A deep learning model for estimation of human body measurements from images

Bartol, Kristijan

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Kristijan Bartol

A DEEP LEARNING MODEL FOR ESTIMATION OF HUMAN BODY MEASUREMENTS FROM IMAGES

DOCTORAL THESIS

Zagreb, 2023

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Kristijan Bartol

A DEEP LEARNING MODEL FOR ESTIMATION OF HUMAN BODY MEASUREMENTS FROM IMAGES

DOCTORAL THESIS

Supervisor: Professor Tomislav Pribanić, PhD

Zagreb, 2023

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Kristijan Bartol

MODEL DUBOKOG UČENJA ZA PROCJENU MJERA LJUDSKOG TIJELA IZ SLIKA

DOKTORSKI RAD

Mentor: prof. dr. sc. Tomislav Pribanić

Zagreb, 2023.

The dissertation was done on the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Information Processing.

Supervisor: Professor Tomislav Pribanić, PhD

Dissertation length: 133 pages

Dissertation num.: ___________

About the Supervisor

Tomislav Pribanić was a Visiting Researcher with the INRIA Grenoble Rhône-Alpes, Grenoble, France, and the Fraunhofer IGD, Darmstadt, Germany. He was a Fulbright Visiting Scholar with the University of Wisconsin, Madison, USA. He is currently a Professor with the Faculty of Electrical Engineering and Computing, University of Zagreb. He teaches several undergraduate and graduate courses in the field of algorithms and data structures, image processing, sensors, and human motion analysis. He has led a number of domestic and international scientific projects, collaborating with researchers from within and outside the EU. His main research interests include computer vision and biomedical signal measurement and analysis. The results of his research have been implemented in technological projects, and he has received recognition for innovations. He is a Senior Member of IEEE and a Collaborating Member of the Croatian Academy of Engineering.

O mentoru

Tomislav Pribanić bio je gostujući istraživač na INRIA-i u Grenoblu, Rhône-Alpes, u Francuskoj, i Fraunhoferu IGD, u Darmstadt u Njemačkoj. Bio je Fulbrightov gostujući stipendist na Sveučilištu Wisconsin, Madison, SAD. Trenutno je profesor na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. Predaje nekoliko kolegija na preddiplomskom i diplomskom studiju iz područja algoritama i podatkovnih struktura, obrade slike, senzora i analize ljudskog pokreta. Vodio je niz domaćih i međunarodnih znanstvenih projekata, surađujući s istraživačima iz i izvan EU. Njegov glavni istraživački interes uključuje računalni vid i mjerenje i analizu biomedicinskih signala. Rezultati svojih istraživanja implementirani su u tehnološke projekte, a dobio je i priznanja za inovacije. Viši je član IEEE i član suradnik Akademije tehničkih znanosti Hrvatske.

Abstract

The understanding of body measurements in and between populations is important and has many applications in medicine surveying, the fashion industry, fitness, and entertainment. Recent advances in human body measurement and shape estimation have been significantly driven by statistical models and deep learning, enabling methods that estimate 3D human meshes from 3D point clouds and 2D images - so called mesh regression methods. This thesis builds upon the state-of-the-art mesh regression approaches from multiple images. The first step is to propose the simplest method and use it as a baseline. The baseline is a linear regression models that takes only person's self-estimated height and weight and estimates the corresponding mesh. The baseline performs surprisingly well compared to the state-of-the-art methods. The second contribution is a 3D human pose estimation model from multiple camera views. The novelty of the model is in the fact that it can take any set of camera views as input, regardless of their relative arrangement and the number of cameras. The third contribution is a model for estimating the parameters of human pose, shape, and clothes from a single image. The estimated parameters are interpretable and, thus, controllable, which is a significant advantage compared to the previous approaches and important for many anthropometric applications. The three proposed models are evaluated in details and compared to the state-of-the-art methods.

Keywords: anthropometry, deep learning, doctoral thesis, human pose estimation, human shape estimation, statistical body models, computer vision

Prošireni sažetak

Analiza mjera i oblika ljudskog tijela važna je za razumijevanje razlika i sličnosti unutar i između populacija te razvoja i promjena tih populacija kroz vrijeme. Znanost koja se bavi mjerenjem ljudskog tijela naziva se antropometrija i ima brojne primjene u medicini, istraživanju, modnoj industriji, vježbanju i zabavnoj industriji. Iako antropometrija postoji već stoljećima, njen razvoj je značajno ubrzan devedesetih godina s pojavom prvih komercijalnih 3D skenera. Komercijalni skeneri omogućuju 3D snimanje čovjeka (skeniranje) i spremanje snimke u obliku oblaka točaka (eng. *point cloud*) ili mreže (eng. *mesh*). Snimanje traje od nekoliko sekundi do nekoliko minuta, što je u usporedbi s ručnim mjerenjem značajno brže. Ipak, ručno mjerenje od strane stručnog mjeritelja i dalje se smatra zlatnim standardom.

Najnoviji napretci u mjerenju ljudskog tijela i procjeni oblika značajno su potaknuti statističkim modelima tijela, stvorenih na temelju skupova podataka 3D skeniranih tijela. Statistički modeli sažimaju informaciju o varijacijama oblika te omogućuju prikaz različitih tjelesnih poza, npr. stajanje s raširenim rukama, sjedenje, ležanje, itd. Sažimajući informaciju o varijacijama u obliku tijela, statistički modeli moćan su alat u interpretaciji tih varijacija i omogućuju stvaranje novih primjeraka tijela na osnovi početne populacije (iz skupa podataka skeniranih tijela) - takav postupak mogao bi se formalno opisati kao *interpolacija* unutar dane distribucije oblika tijela. Konačno, statistički modeli omogućuju procjenu 3D oblika i poze ljudskog tijela iz slika, što je jedna od glavnih tema ovog doktorskog rada. Procjena 3D oblika i poze može se opisati kao regresijski problem u kojem se traže parametri oblika i poze tijela koji najbolje opisuju osobu prikazanu na slici.

Znanstveni doprinosi

Kao prvi doprinos, u doktorskom se radu pokazuje da se statistički modeli mogu koristiti za procjenu mjera ljudskog tijela koristeći samo informaciju o visini i težini osobe. Štoviše, pokazuje se da za procjenu antropometrijskih mjera nije potrebna točna informacija o visini i težini već je dovoljno da ta informcija bude *samoprocijenjena*. S obzirom da u sklopu doktorskog rada nije napravljena studija sa stvarnim osobama koje bi dale procjenu svoje visine i težine, istraživanje je provedeno na sintetičkim podacima. Pogreške samoprocjene modelirane su dodavanjem normalne distribucije na podatke o visinama i težinama.

Kako bi se mogle procijeniti mjere tijela za bilo koju osobu za koju se ne zna njihova samoprocijenja visina i težina, mora se koristiti drugačija strategija. Ostatak doktorskog rada fokusira se na procjenu parametara iz statističkog modela tijela na temelju slika. Drugi doprinos je, stoga, vezan uz procjenu karakteristika ljudskog tijela iz slika, točnije predložen je model učenja procjene položaja na temelju više pogleda iste osobe. Na ulazu se uzimaju 2D poze koje su dobivene prethodno naučenim modelom. Važno je napomenuti da te poze nisu savršeno

dobro procijenjene, odnosno, da su položaji pojedinih zglobova pomaknuti u odnosu na željeno središte zgloba. Ranije metode učenja za 3D procjenu poze izrazito su uspješne u filtriranju pogrešaka iz modela za procjenu 2D poze, ali imale su značajno ograničenje da nisu mogle biti upotrijebljene na skupovima kamera koje nisu korištene za treniranje. Predloženi model je nov na način da može uzeti bilo koji kalibrirani skup kamera i procijeniti 3D pozu, a pritom ne izgubiti značajno na točnosti procjene.

Treći i posljednji doprinos je prijedlog modela procjene parametera poze, oblika tijela i odjeće iz samo jedne slike. Predloženi model koristi prethodno predstavljene modele za opis poze, oblika i deformacije odjeće pomoću parametara, a novost je u tome što je po prvi puta moguće te parametere procijeniti iz slika. Implementacija predloženog modela koristi samo jedan primjer prethodno predstavljenog parametarskog modela, ali opisana računarska strategija (eng. *computational strategy*) nije ograničena tim modelom. Uz računarsku strategiju, predložena su i dva nova skupa podataka - *ClothSurreal* i *ClothAgora*, koji predstavljaju inačice postojećih sintetičkih skupova podataka ljudi u minimalnoj odjeći (donjem rublju), uz dodatak odjeće. Ovakvi skupovi podataka vrlo su važni za učenje predloženog modela, kao i za učenje budućih modela za procjenu obućenih ljudi iz slika.

Opisana tri modela detaljno su evaluirana i uspoređena s najsuvremenijim tehnikama. Konačno, predloženi pristupi su uspoređeni i raspravljeni u smislu njihove iskoristivosti za mjerenje ljudskog tijela iz slika, s ili bez korištenja informacije o visini i težini osobe.

Struktura doktorskog rada

Doktorski rad se sastoji od sedam poglavlja i tri dodatka te je strukturiran na sljedeći način. Prvo poglavlje pruža sveobuhvatni uvod u antropometriju, proučavanje fizičkih mjerenja i dimenzija ljudskog tijela. Poglavlje pokriva različite aspekte antropometrije, uključujući tradicionalne i digitalne metode mjerenja tijela, 3D tehnike skeniranja i statističke modele korištene u stvaranju 3D mreže ljudskog tijela. Poglavlje počinje raspravom o tradicionalnoj antropometriji, koja uključuje mjerenje ljudskog tijela pomoću konvencionalnih alata. Naglašava se važnost točnih mjerenja jer ona čine osnovu statističkih modela ljudskog tijela. Sljedeće, poglavlje ulazi u digitalnu antropometriju, koja koristi digitalnu tehnologiju za dobivanje preciznijih i učinkovitijih mjerenja ljudskog tijela. Ističu se prednosti ovog pristupa u odnosu na tradicionalne metode, posebno u područjima medicinskog istraživanja, ergonomije i virtualnog dizajna. Poglavlje zatim istražuje 3D tehnologiju skeniranja, koja se koristi za stvaranje 3D oblaka točaka i mreže ljudskog tijela. Potom se uvode i opisuju statistički modeli ljudskog tijela, na osnovi kojih se mogu stvoriti 3D mreže i iz njih izračunati pripadne antropometrijske mjere. Konačno, na osnovi opisanih tehnika predlažu se koraci za općeniti postupak mjerenja tijela korištenjem digitalne antropometrije.

U drugom poglavlju ove doktorske disertacije (2. Antropometrija na temelju slika), predstavljen je koncept antropometrije iz slika. Poglavlje raspravlja o različitim metodama koje postoje za procjenu mjera ljudskog tijela iz slika i stavlja ih u kontekst antropometrije u cjelini. Jedna od ključnih tema u poglavlju su modeli dubokog učenja koji se koriste za izvlačenje značajki iz slika kojima se može naučiti procjena 3D mreže ljudskog tijela. Ti su modeli razvijeni kako bi se omogućila točnija procjena 3D položaja, oblika i mjera tijela iz slika. Poglavlje opisuje razvoj i implementaciju modela dubokog učenja u antropometriji na temelju slika, ističući prednosti i izazove ovog pristupa. Modeli su osmišljeni za analizu slika ljudskog tijela i izvlačenje značajki koje su korisne za procjenu tjelesnih mjera, poput tjelesne težine, visine i indeksa tjelesne mase (BMI). Poglavlje pregledava postojeću literaturu o modelima dubokog učenja za antropometriju na temelju slika i pruža uvid u trenutno stanje u tom području.

U trećem poglavlju ove doktorske disertacije (3. Bazni model procjene mjera tijela) predlaže se jednostavan linearni regresijski model za procjenu mjera ljudskog tijela na temelju visine i težine osobe. Model je osmišljen tako da procijeni preostale mjere, poput duljine i opsega tijela, koristeći samo informacije koje osoba procjenjuje za sebe, posebno visinu i težinu tijela. Da bi se simulirale realistične procjene visine i težine, šum se iz normalne distribucije dodaje podacima. Predloženi model postiže točnost koja je usporediva s najboljim postojećim metodama za procjenu mjera ljudskog tijela, a u nekim slučajevima je čak i precizniji od dubokih modela učenja. Poglavlje opisuje razvoj i primjenu predloženog linearnog regresijskog modela, ističući njegove prednosti i ograničenja. Model je osmišljen da bude jednostavan i lako se koristi, što ga čini primjenjivim u različitim postavkama za brzu i preciznu procjenu mjera tijela, bez potrebe za složenom ili skupom opremom. Ukratko, treće poglavlje daje detaljan opis predloženog linearnog regresijskog modela za procjenu mjera ljudskog tijela.

Četvrto poglavlje ove doktorske disertacije (4. Učenje procjene položaja tijela iz slika) predstavlja novu metodu za procjenu 3D položaja tijela koristeći slike dobivene iz više sinkroniziranih kamera. Predloženi model je dizajniran da radi u specifičnim uvjetima, odnosno da postoji samo jedna osoba u sceni i da je osoba vidljiva iz najmanje dva pogleda u svakom trenutku. Poglavlje pruža detaljan opis predloženog modela, ističući njegove prednosti i ograničenja. Predloženi model postiže najbolje rezultate u usporedbi s konkurentnim modelima na slikama koje nisu dobivene istim kamerama kao u skupu za treniranje. Svojstvo predloženog modela da postigne jednako ili približno dobre rezultate koristeći drugačiji raspored kamera pokazuje otpornost modela i sposobnost generalizacije predložene metode izvan uvjeta iz skupa za učenje. Predloženi pristup ima potencijal za upotrebu na skupovima podataka snimljenim različitim brojem kamera i različitom prostornom organizacijom, što s ranijim pristupima nije bilo moguće.

Peto poglavlje ove doktorske disertacije (5. Učenje procjene oblika tijela iz slika) predstavlja novu metodu za procjenu 3D oblika ljudskog tijela iz jedne slike. Predložena metoda ima nekoliko jedinstvenih značajki u odnosu na postojeće metode za procjenu 3D položaja i oblika. Jedna od ključnih prednosti predložene metode je njena sposobnost procjene oblika ljudi iz slika u širokoj odjeći. Još jedna važna značajka predložene metode je njena sposobnost procjene parametara same odjeće. To omogućuje metodi da bolje računa s odjećom povezanim distorzijama u procijenjenom 3D obliku tijela. Predložena metoda predstavlja značajan korak prema pojednostavljenju i ubrzanju procjene ljudskih značajki iz slika koje nemaju značajnih ograničenja.

U šestom poglavlju doktorskog rada pružena je opsežna rasprava koja se bavi cjelokupnim istraživačkim radom predstavljenim u disertaciji. Konkretno, analizira se performansa predloženih modela u smislu procjene mjera tijela. To uključuje bazni model opisan u poglavlju 3 i model za procjenu oblika opisan u odjeljku 5. Poglavlje pruža kritičku evaluaciju predloženih modela, naglašavajući njihove snage i ograničenja. Nadalje, detaljno se raspravlja o pretpostavkama korištenim tijekom evaluacije modela. Konačno, navodi se budući rad, koji sugerira potencijalne smjerove za daljnja istraživanja u području antropometrije i 3D skeniranja. Rasprava pruža vrijedan uvid u doprinose i značaj istraživačkog rada predstavljenog u doktorskom radu, otvarajući put za daljnje razvoje u polju antropometrije upotrebom računalnog vida.

Završno poglavlje doktorskog rada daje zaključna razmatranja, sažimajući ključne doprinose istraživačkog rada. Poglavlje ističe važnost predloženih modela za procjenu 3D položaja, oblika i mjera ljudskog tijela iz slika. Nadalje, poglavlje pruža smjernice za daljnja znanstvena istraživanja na temu, predlažući potencijalna područja za istraživanje i poboljšanje. Konačno, završno poglavlje nudi sveobuhvatan sažetak disertacije i njezinih doprinosa u području antropometrije, 3D skeniranja i posebno mjerenja tijela na temelju slike.

Dodatci doktorskom radu

Tri dodatka opisuju detalje koji su logički izvan dosega rada ili služe kao dodatni izvor informacija koji nisu potrebni za razumijevanje predloženih doprinosa.

Prvi dodatak opisuje fizičku implementaciju stereoskopskih zrcala (engl. *catadioptric stereo*) kao dodatka na pametne telefone. Stereoskopska zrcala omogućuju da prednja i stražnja kamera mobilnog telefona budu korištene kao stereoskopski par u svrhu 3D rekonstrukcije, što originalno nije moguće jer dvije kamere nemaju zajedničko pregledno polje (engl. *field-of-view*). U svrhu provedenog istraživanja, dizajniran je i 3D isprintan prototip adaptera za stereoskopska zrcala. Detaljno su opisane mogućnosti izrađenog adaptera u smislu zajedničkog preglednog polja te veličine preglednog polja s obzirom na stupnjeve slobode koje adapter omogućuje, npr. razmak između zrcala, veličina zrcala te kutevi između zrcala i kamera. Konačno, prototip adaptera za stereoskopska zrcala korišten je za 3D rekonstrukciju poze tijela te su na takvoj 3D procjeni izračunate mjere tijela i uspoređene s ručno dobivenim mjerama. Pokazuje se da je stereoskopska zrcala moguće koristiti za izračun pojedinih mjera ljudskog tijela.

Drugi dodatak sadrži više detalja o vrstama 3D skenera često korištenih za 3D skeniranje

ljudskog tijela. Skeneri su podijeljeni u skupine s obzirom na različita svojstva. Prvo svojstvo je korištena tehnologija skeniranja, koja može biti fotogrametrija, strukturirano svijetlo i *time-of-flight*. Svaka tehnologija ima prednosti i nedostatke, a neke od najvažnijih svojstava su da je fotogrametrija u pravilu najbrža, a strukturirano svijetlo u pravilu najtočnije. Ipak, postoji mnogo komercijalnih skenera koji koriste kombinaciju različitih tehnologija 3D skeniranja istovremeno. Drugo svojstvo je veličina skenera. Postoje skeneri u koje osoba ulazi cijela (stacionarni skeneri), skeneri koji se mogu držati u ruci dok traje skeniranje tijela, te skeneri koji koriste samo tehnologiju pametnog telefona ili tableta. Većina je stacionarnih skenera u pravilu točnija od ručnih i mobilnih skenera, ali takvi su skeneri manje praktični i skuplji za prosječnog potrošača. Posebno su zanimljivi skeneri koji koriste samo tehnologiju pametnog telefona jer bi takva rješenja, uz pretpostavku približno jednake točnosti, potencijalno zamijenila veće skenere (barem za potrebe antropometrije). Ipak, mobilni skeneri još uvijek nisu dovoljno točni i koriste oslanjaju se na statističke modele tijela. Takvi će modeli raditi bolje za mjerenje ljudi koji su bliže prosječnoj osobi iz originalne populacije i obrnuto. Ipak, noviji uređaji u sebi imaju i *time-of-flight* senzore pa u tom smislu mogu ipak biti potencijalno zanimljivi i bez isključivog oslanjanja na statističke modele. Ostala svojstva 3D skenera opisanih u drugom dodatku su njihova cijena, brzina, točnost, rezolucija i prilagodba za potrebe antropometrije (ako postoji).

Treći dodatak pruža praktične preporuke za određene primjene mjerenja ljudskog tijela, poput medicine, fitnessa i zabavne industrije, te povezuje dane preporuke s prethodno opisanim tipovima 3D skenera. Dane su preporuke značajne za svaku osobu i organizaciju koja direktno ili indirektno koristi 3D skenere i mjerenje tijela u svom poslu, a posebno je važno pri eventualnoj odluci o kupnji skenera i prikladnom tipu s obzirom na danu djelatnost. Primjer praktične preporuke je postupak koji se sastoji od koraka pripreme (označavanje tijela, upute o položaju tijela te zauzimanje položaja), 3D skeniranja, i izvlačenja mjera tijela iz ručno obrađene 3D snimke. Ovakav je postupak dovoljno točan za većinu primjena, ali uključuje korake koji zahtijevaju stručno osoblje i poznavanje programa za obradu snimaka. Preporuka praktičnog postupka koji zahtijeva manju razinu stručnosti, ali ne garantira visoku razinu točnosti se sastoji od koraka 3D ili 2D skeniranja (fotografiranja), procjene parametara oblika tijela, te izvlačenje mjera iz mreže koja odgovara procijenjenim parametrima oblika.

Ključne riječi: antropometrija, duboko učenje, doktorski rad, procjena položaja tijela, procjena oblika tijela, statistički modeli tijela, računalni vid

Contents

Chapter 1

Introduction

Anthropometry is a subfield of applied metrology that studies how to measure the physical properties of humans. In general, anthropometry includes the complete process of data collection, documentation, summarization, and analysis [1]. In a more narrow sense, anthropometry studies *body measurement*, where lengths, breadths, heights, and circumferences are used to numerically describe body segments and the overall body shape [2]. Body measurement is crucial in quantifying the correlations in and between populations of different countries, ethnicities, cultures, and ages [3, 4], and it strongly impacts medicine [5, 6], fitness [7], fashion industry [1], surveying [1, 8], and entertainment [9].

Human body dimensions may be obtained in various ways, e.g., they can be measured manually using traditional tools such as tape measures and calipers [1] or semi-automatically using 3D scanners. To ensure both repeatability and comparability, body measurements are standardized by the means of postures and body landmarks [10, 11]. Although manual measurement is the benchmark for anthropometry, several reports suggested that expert measurement assessors (measurers) and 3D scanners achieve comparable accuracy and that the repeatability is usually higher when 3D scanners are used [12, 13, 14]. An additional advantage of using 3D scanners compared to expert assessors is the efficiency in terms of time [15]: the duration of an automatic scan is usually under several seconds, although it goes up to 30 seconds for high-quality scans[*]. Therefore, even though the first commercial 3D body scanners appeared in the 1990s [16] and were expensive, requiring trained personnel and extensive manual postprocessing [17], the scanning technology is currently mature enough and is comparable in performance to human assessors [18, 19, 20].

[*] See Appendix A for more details

1.1 Manual (Traditional) Anthropometry

Traditional anthropometry involves the use of tools such as tape measures and calipers [1], as illustrated in Fig. 1.1. Measurements obtained by an expert are considered the benchmark and are used as ground truth [21]. The public body measurement databases such as ANSUR are collected by expert measurement assessors. On the other hand, even the measurements taken by the experts are not absolutely accurate [14]. For that reason, the allowable error values (or the "expert" errors) [22] for each measurement are defined in the standard ISO:7250 [10, 11].

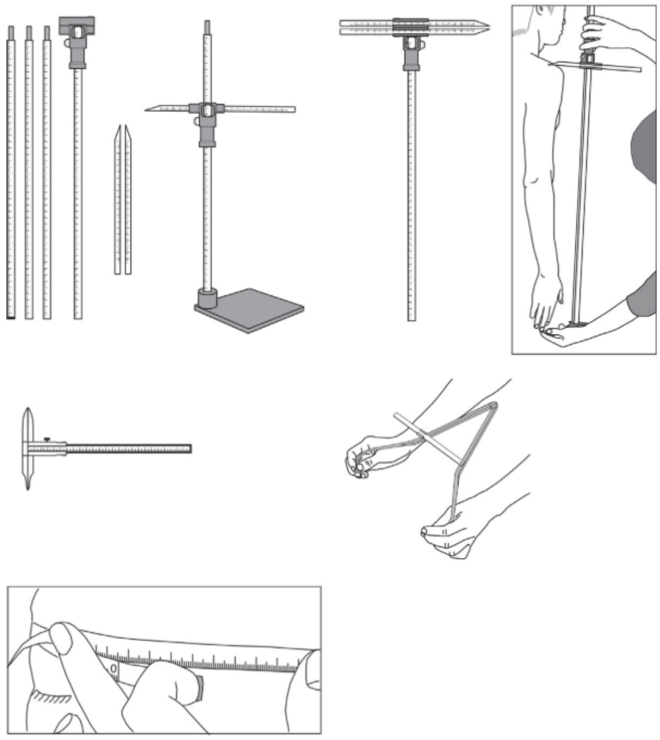


Figure 1.1: The examples of traditional body measurement tools - calipers and tape measures. The image is adapted from [1].

1.2 Measurement Standards and Evaluation

Although manual expert measurements are considered to be the benchmark, they will never be perfect, first of all, because the human body is not rigid, for example, some body measurements change throughout the day [23]. There are many proposed methods for the analysis of body measurement errors [24, 25, 26, 27], but the conclusion is that none of the methods are ideal for manual anthropometry in general. [28]. The most common, and simple to calculate, easily interpretable approach to the body measurement error analysis is proposed by Gordon et al. [22]. They came up with the threshold error values, called the allowable errors (AEs). The aim of any accurate body measurement technique should be to be as close as possible to the AEs. The allowable errors are the mean absolute errors (MAEs) made by the expert measurers. The authors of the AE values also took into account many factors that potentially affect measurement accuracy, such as posture, time of the day (morning, evening), measurement technique and instrument, etc. We refer to the allowable errors throughout the dissertation.

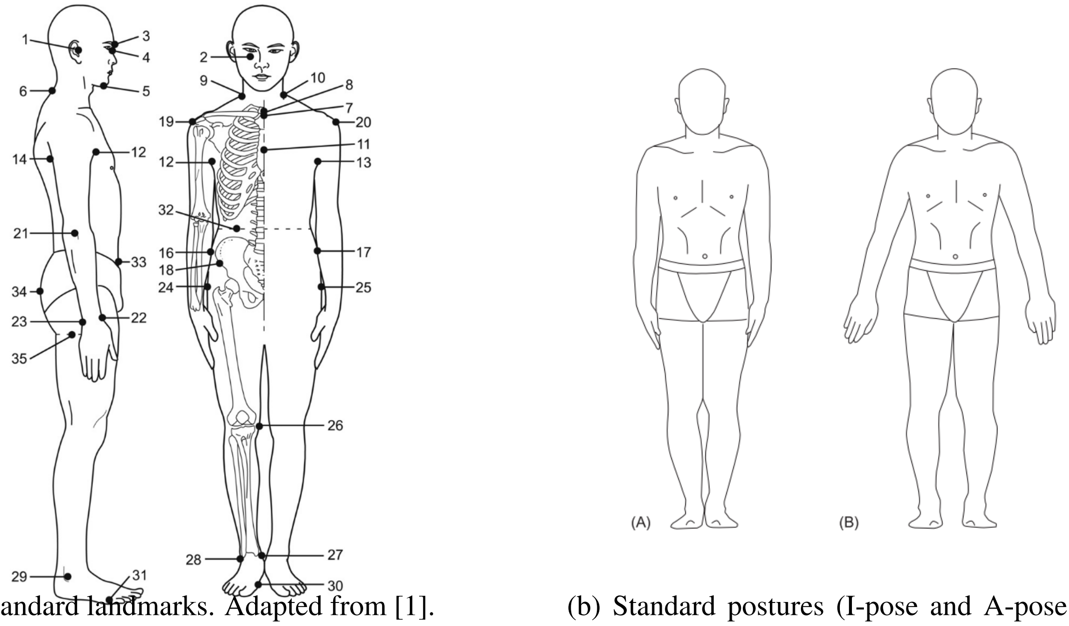
Table 1.1: The list of human body landmarks according to ISO 7250-1:2017 standard [10]. The numbers correspond to the numbers in Fig. 1.2a. The letters *R* and *L* abbreviate right and left.

Human body landmarks (ISO 7250)		
1 tragion	12 axilla pnt. ant. R	23 stylion ulnare R
2 orbitale	13 axilla pnt. ant. L	24 trochanterion R
3 glabella	14 axilla pnt. post. R	25 trochanterion L
4 sellion	15 axilla pnt. post. L	26 tibiale R
5 gnathion	16 iliocristale R	27 sphyrion R
6 cervicale	17 iliocristale L	28 sphyrion fib. R
7 suprasternale	18 iliospinale ant. R	29 supratarsale fib. R
8 front neck	19 acromiale R	30 metatarsale tib. R
9 side neck R	20 acromiale L	31 metatarsale fib. R
10 side neck L	21 radiale R	32 waist level
11 mesosternale	22 stylion R	33 abdom. ext. level

1.2.1 Landmarking, Body Postures, and Standards

To ensure the comparability of measurements between the anthropometric surveys [1] and to be able to compare the results quantitatively, standardization of body landmarks, measurements, and postures is required.

The list of body landmarks is shown in Table 1.1 and the list corresponds to the ISO standard 7250-1:2017 [10]. The landmarks are located on the skin to reduce the ambiguity in their locations between the subjects, i.e., in such a way that they have the same semantics for every measured body. The landmark locations corresponding to Table 1.1 are shown in Fig. 1.2a. Before the measurement, markers that represent body landmarks are placed on the surface of the human skin. Fig. 1.3 shows the procedure of landmarking and marker placing on human skin. The particular landmark location is first determined and then marked. The dome or a similar convenient round object is then placed in the marked location. The markers are useful for feature extraction [29], however, marked placing is a tedious and fallible process. In that sense, successful markerless systems have been proposed [30, 31].



(a) Standard landmarks. Adapted from [1].

(b) Standard postures (I-pose and A-pose, respectively). Adapted from [1].

Figure 1.2: Standard landmarks and postures.

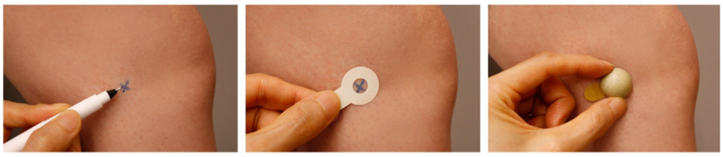


Figure 1.3: The landmarking process on the human body. The image is taken from [1].

There are several standard standing poses recommended by the ISO 20685-1:2018 [11] (Fig. 1.2b). The person takes one of the standard poses, holds his or her breath during the scanning, and tries to keep as static as possible [30]. In the I-pose, the subject stands tall with the shoulders completely relaxed and arms hanging down naturally, holding the feet together. In the A-pose, the feet are about 20 cm (0.7 feet) apart, the elbows are straight and the palms face backward [1]. The arms form an angle of 20 degrees with the torso. Using the standard postures is not always required when measuring the body, but usually, it is when recording the datasets that capture shape variations [18, 19, 20, 32, 33, 34, 35].

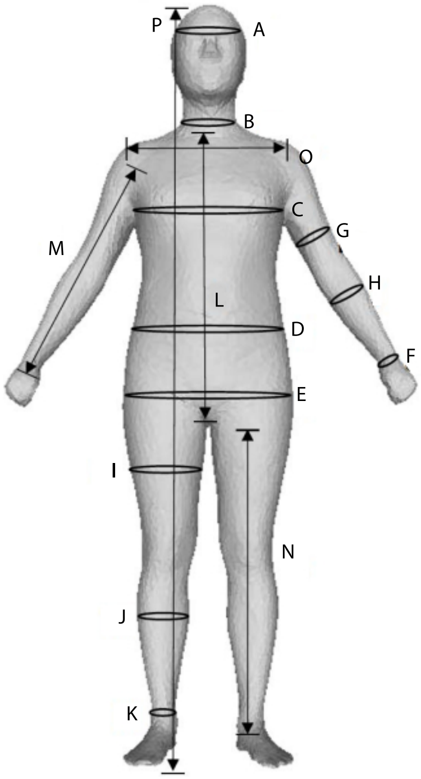
Finally, a list of standard body measurements [10] is shown in Table 1.2. The measurements consist of distances (lengths, breadths, depths, and heights), circumferences, and soft biometrics (weight, height, body mass index (BMI)). In the next subsection, we specify which of these measurements are used for the comparison between the state-of-the-art anthropometric methods.

1.2.2 Body Measurements for Evaluation

Even though different anthropometric methods can be compared against standardized body measurements defined in the ISO 7250-1:2017 [10] standard, authors still tend to report their

Table 1.2: The list of 44 standardized human body measurements [10].

Human body measurements			
1 eye	12 forearm circum. L	23 weight	34 bicep circum. R
2 cervicale	13 forearm circum. R	24 height	35 shoulder breadth
3 shoulder-elbow L	14 neckbase breadth	25 BMI	36 elbow circum. L
4 shoulder-elbow R	15 thigh clearance	26 neck circum.	37 elbow circum. R
5 crotch height	16 wall-acromion distance	27 chest circum.	38 knee circum. L
6 tibial height	17 grip and forward reach	28 waist circum.	39 knee circum. R
7 chest depth	18 elbow-wrist L	29 thigh circum. L	40 neck base circum.
8 body depth	19 elbow-wrist R	30 thigh circum. R	41 neck circum.
9 thorax depth	20 hip circum.	31 calf circum. R	42 head circum.
10 chest breadth	21 buttock-popliteal	32 calf circum. R	43 trouser waist circum.
11 hip breadth	22 buttock-knee	33 bicep circum. L	44 iliac spine breadth



	Measurement	Allowable Errors
A	Head circum.	± 5 mm
B	Neck base circum.	± 11 mm
C	Chest circum.	± 15 mm
D	Waist circum.	± 12 mm
E	Hip circum.	± 12 mm
F	Wrist circum.	-
G	Bicep circum.	-
H	Forearm circum.	-
I	Thigh circum.	± 6 mm
J	Calf circum.	-
K	Ankle circum.	± 4 mm
L	Shoulder-crotch length	-
M	Shoulder-wrist length	-
N	Inside leg length	-
O	Shoulder breadth	± 8 mm
p	Height	± 10 mm

Figure 1.4: A set of body measurements used for evaluation and comparison between the state-of-the-art along with their corresponding allowable errors. The measurements are abbreviated: C stands for circumference, L for length and B for breadth. The image is adapted from: [36].

evaluations for different sets of measurements, which are not always equal [29]. This makes it difficult to objectively and comprehensively compare between the proposed anthropometric methods. To somehow bypass these difficulties, a review paper on body measurements [29] evaluates anthropometric methods on 16 common body measurements. These 16 body measurements shown specified in Figure 1.4 and specified in the table next to the Figure. In particular, the table specifies body measurement labels (A-P), names, and their corresponding allowable errors. The first 15 body measurements are also used for the evaluation of methods in this dissertation (excluding overall height[†]).

1.2.3 Evaluation Metric

There are several evaluation metrics concerning reliability, precision, and accuracy [37] that are commonly reported [38, 39, 40]. The lack of standardized evaluation measures complicates straightforward comparisons of various anthropometric methods since different error metrics cannot be converted from one to another. To still be able to compare ourselves against the state-of-the-art with respect to body measurement estimation, in this dissertation we use only a single metric, the mean absolute error (MAE) since it is almost always reported in the published anthropometric approaches. The MAE is a measure of accuracy, and is calculated between the body measurement method estimation E_{est} and the ground truth E_{gt} as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \mid E_{est} - E_{gt} \mid \tag{1.1}$$

for every subject i from the dataset, where N is the total number of subjects. The ground truth is usually obtained using manual measurement as described in the ISO 20685-1:2018 [11] standard.

1.3 Digital Anthropometry - 3D Scanning

Over the years, numerous 3D scanning systems have been proposed. Fig. 1.5 highlights three phases of the development of 3D scanners. The first phase consists of large and fixed 3D scanners. These scanners still exist nowadays and are usually the most accurate ones. The downside is their price and size. In many cases, they still require manual postprocessing to get the best results. The second phase consists of smaller 3D scanning platforms and handheld scanners, which are more convenient to use and are generally less expensive. Finally, the third phase consists of smartphone scanners, where the body measurements are extracted either from images or depth sensors embedded on the back side of the device. A more detailed overview

[†] The overall height is excluded because estimating height often includes scale estimation which is a particularly difficult problem, especially for the monocular methods. We additionally discuss these difficulties in relation to our proposed baseline in Chapter 3

and comparison between 3D scanners are given in Appendix B. A comprehensive overview is given in the survey paper [29].

With respect to 3D scanning technologies, three popular techniques can be highlighted, following the survey paper [29]: passive stereo, structured light, and time-of-flight imaging, which we now go into greater depth about.

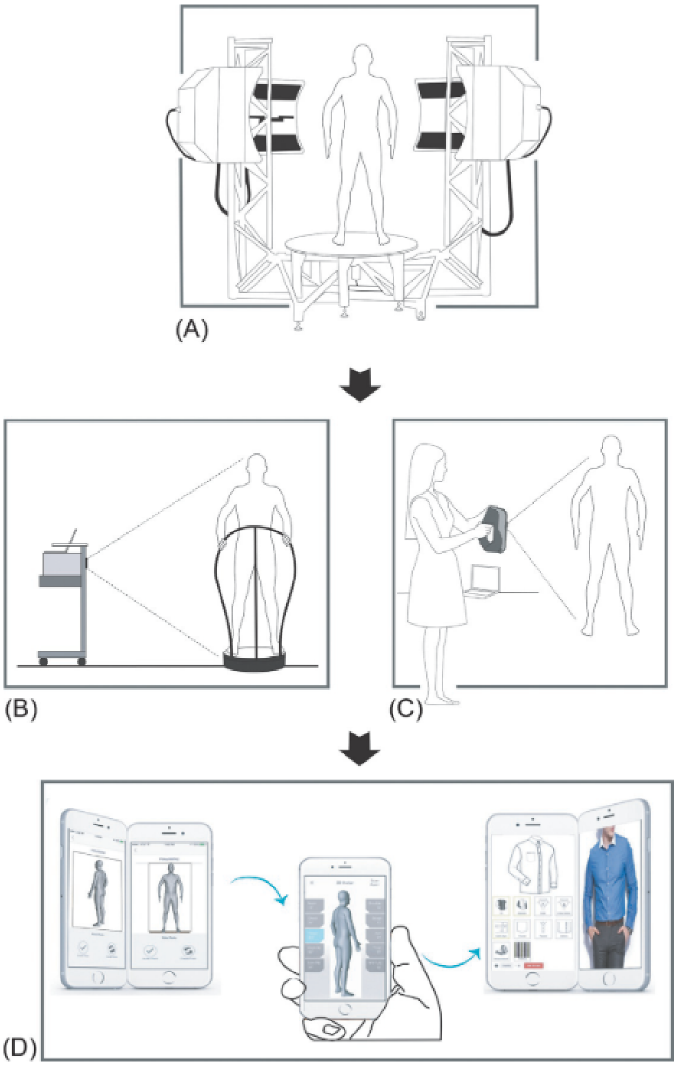


Figure 1.5: The development of 3D body scanning systems over the years. Adapted from [1].

1.3.1 Photogrammetry

Passive stereo is a measuring technique for 3D reconstruction from multiple camera views. Photogrammetry is the science of measuring objects from photographs. Passive stereo and photogrammetry are sometimes used interchangeably in the context of 3D scanning [41, 42, 43]. For clarity, we use the passive stereo in the remainder of the dissertation. PS-based 3D scanners use RGB cameras to obtain color images. The PS assumes that multiple cameras are pointing to a person. Under passive stereo, in this Section, we describe the principles of stereo and monocular reconstruction, as well as motion capture systems.

Stereo. The simplest PS setup is a binocular stereo, which consists of two RGB cameras that are either horizontally or vertically oriented (see Fig. 1.6). The triangulation and correspondences discovered on the photos provide the foundation for the reconstruction [44]. According to Fig. 1.6, the point P in the 3D scene projects to pixels p_1 in the first image and p_2 in the second image. The corresponding pixel location p_2, on the other hand, is not known a priori given a fixed pixel location p_1. An image block at p_1 is matched with the most similar block along the epipolar line l to find the location p_2. The discrepancy between the associated pixel coordinates [‡], $|p_1 - p_2|$ (called the disparity) is used to calculate the depth of a point P using triangulation [46]. By coupling particular camera pairs [47] or by utilizing multi-view-stereo approaches [48], the stereo approach can be expanded to include more than two cameras.

Monocular. Each viewpoint (frame) is treated as a separate camera in a monocular moving-camera-based 3D reconstruction, which is a specific case of stereo reconstruction. Together, the general monocular approaches [49, 50, 51] reconstruct a 3D scene and estimate camera locations for each frame. To locate the correspondences, the keypoints are first found using the standard algorithms [52, 53, 54, 55] and then they are quickly and accurately matched between the images. Following bundle adjustment (BA) refinement [56], the correspondences are then used for the initial 3D reconstruction and camera parameter estimation. In general, human 3D scanning is easier since camera viewpoints can be known before the reconstruction. This is implemented in a way that either the person is standing on a revolving platform or the camera is spinning around them to resemble camera rotation. Keep in mind that the subject must remain still throughout the quasi-static scanning. Based on timestamps, the relative camera locations with regard to the subject are retrieved. The above-mentioned stereo reconstruction principles can be used to acquire a dense 3D reconstruction.

Motion. Motion capture or MoCap is a (semi-) passive stereo technique that uses body markers visible under standard or near-infra-red light. The MoCap markers are usually small, round objects with reflective surfaces. MoCap produces sparse 3D reconstructions and is usually used for motion tracking. The number of body markers is between 30 [35] and 300 [57]. Multiple markers are often used to estimate the location of a single keypoint (joint), as markers can only be placed on the surface of the body.

For an overview of 3D scanning devices based on photogrammetry, refer to the Appendix B.

[‡] Note that the images are usually rectified [45].

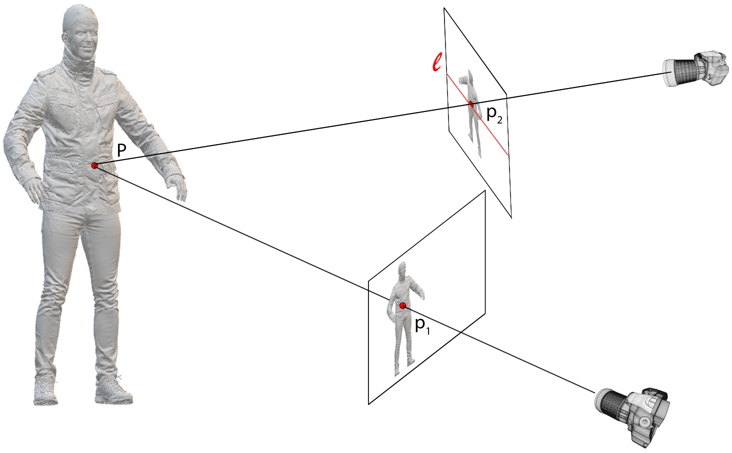


Figure 1.6: Passive stereo approach. Point p_2 is the most similar image pixel to point p_1 along the epipolar line l. The image is adapted from [29].

1.3.2 Structured Light

The standard solution is to project a textured pattern over the scene in order to improve PS's poor 3D reconstruction quality in cases when the texture is low or repeats. Active stereo (AS) [58, 59, 60, 61] upgrades PS by casting a light pattern onto the body to enhance the search for the corresponding image pixels between viewpoints. Structured light (SL) approaches [62, 63, 64, 65], conversely, look for the correspondences between the camera and the light pattern. The methodology and technologies used in SL are the main topics of the next paragraphs.

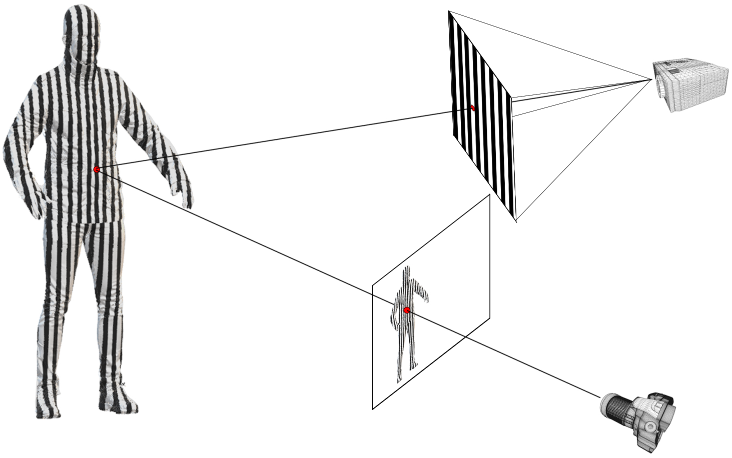


Figure 1.7: Structured light (projector-based) approach. The image is adapted from [29].

Based on SL technology, we separate laser scanners from projector scanners. A laser scanner [58, 66, 67], seen in Fig. 1.8, projects dot or striped patterns across the scene with a laser. Laser scanners have the accuracy of less than one-millimeter [1, 68, 69] and less complicated decoding than projector-based scanners [16]. However, because the laser line must traverse the entire body, laser scanners often have a long scanning time [70]. Generally speaking, projector-based scanners are quicker than laser scanners [64], since the entire body may be scanned

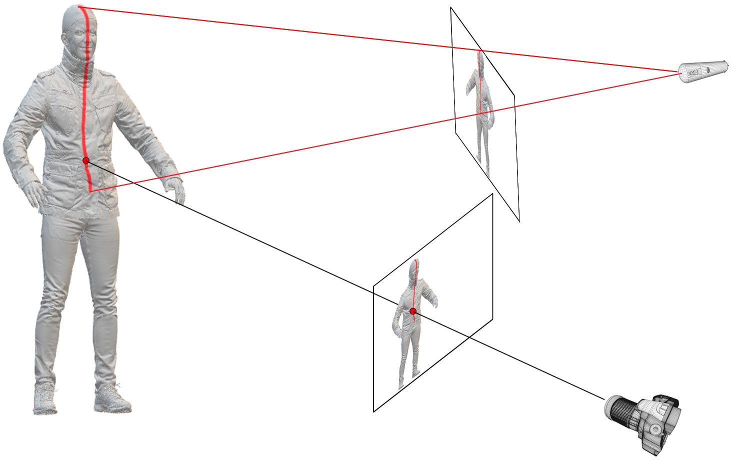


Figure 1.8: Structured light (laser-based) approach. The image is adapted from [29].

at once from a single view and more sophisticated 2D patterns can be presented. Additionally, compared to laser scanners, projector-based scanners have less safety limitations [71]. Projector-based scanners' precision range (μm-mm) is adequate for high-quality body measuring even if they are not as accurate as laser scanners (see Sec. 3.4).

The projected light patterns have generally been categorized in a variety of ways, such by the number of projected patterns (single- or multi-shot), color (achromatic or colored), transitions (discrete or continuous), or organized form (stripes, grids, dot arrays, gradients, etc.) [64, 72, 73, 74, 75], as seen in Fig. 1.9a-1.9d. Short-duration achromatic multi-shot patterns are typically utilized for (quasi-)static human 3D scanning, providing a trade-off between acquisition speed and reconstruction accuracy [64]. Single-shot patterns are more appropriate for dynamic scenarios where quick capture is required [76].

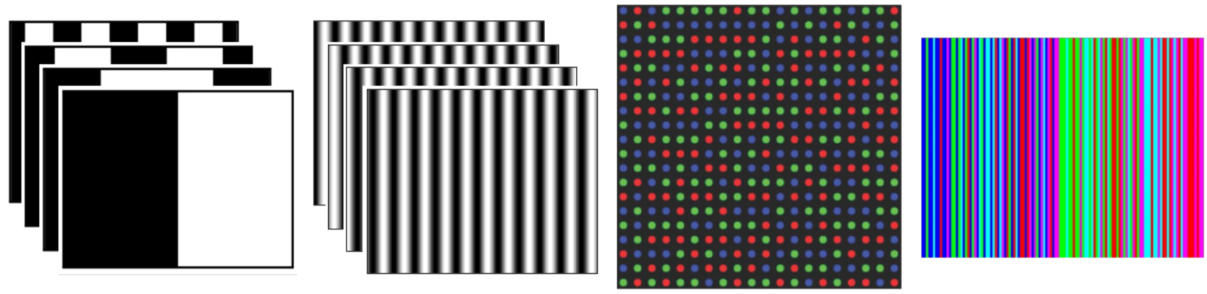
Depending on the projected pattern and light, the correspondences between the camera and the light source are discovered. To locate the (monochromatic) light projections in the image, laser-based techniques mostly use pattern detection algorithms [77, 78]. Visible-light scanners, on the other hand, have more complex pattern decoding mechanisms [79, 80], especially in the case of multiple projectors and light interference [80, 81]. We direct readers to the pertinent survey publications [73] for more information. After the correspondences have been found, the 3D human body can be rebuilt using ray-to-ray or ray-to-plane triangulation [8, 58, 63, 72].

Refer to Appendix B for a summary of structured light-based 3D scanning equipment.

1.3.3 Time-of-Flight

The time it takes for an emitting light signal to travel from the illumination source to the three-dimensional scene and back to the sensor is what ToF scanners measure, as seen in 1.10. The time of flight of the light signal directly correlates to the distance information [8, 86, 87, 88].

The light emitter and the photodetector are a ToF scanner's primary parts [8]. The light



(a) Discrete achromatic multi-shot stripe pattern - Gray Code [82].

(b) Continuous achromatic multi-shot stripe pattern - Phase Shift [83].

(c) Discrete colored single-shot grid pattern - M-array [84].

(d) Discrete colored single-shot stripe pattern - De Brujin code [85].

Figure 1.9: The examples of different types of structured light patterns. The images are adapted from [29].

emitter sends a modulated beam of light, often in the NIR range [86], using a laser or an LED. The light from the emitter is usually dispersed throughout the entire scene using the lens. A matrix of point-wise sensors is typically used by the photodetector [87]. CCD/CMOS matrix sensors are typically utilized for 3D scanning of humans.

Pulsed-light (direct) and continuous-wave (indirect) reconstruction techniques can be separated [8, 87]. The round-trip time of an emitting light pulse is indirectly measured using continuous-wave (CW) techniques, which also gather data on the signal's time-dependent intensity [88, 89]. The phase shift of the sent and received light signals is then used to demodulate (retrieve) the distance of a location [87, 90]. Typically, sine or square waves are used to modulate the emitted light signal's amplitude [91]. As a result of the waves' periodicity, there is a limit to their scanning range at half of the modulation wavelength, after which an ambiguity issue develops [92]. While reducing the maximum range, increasing the modulation frequency improves measurement accuracy [87]. Multiple modulation frequencies can be used to increase the measurement range [93, 94]. Fortunately, since human bodies are scanned from close range in anthropometric applications, this rarely poses an issue. Time-to-digital (TDC) or duration-to-amplitude (TAC) circuitry is used in pulsed-light (PL) technologies to directly detect the round-trip time of an emitted light pulse [87, 89]. Since light travels at a very high speed, PL approaches need timing data that is extremely precise—on the order of picoseconds—to determine a millimeter-to-millimeter distance range [8, 86, 89]. PL is therefore rarely utilized for 3D body scanning.

ToF cameras offer affordable, small, precise, dependable, and low-power-consumption sensors [88, 92, 95]. ToF is texture-independent, with a minimal post-processing time and lower-light capabilities [96]. The major issue with single ToF camera scanners is a low scanning resolution, even though fast frame rates that are appropriate for dynamic scanning can be accomplished [88, 96]. Utilizing numerous ToF cameras can increase resolution [97], but this raises complex light interference difficulties that must be resolved [90]. ToF is therefore still

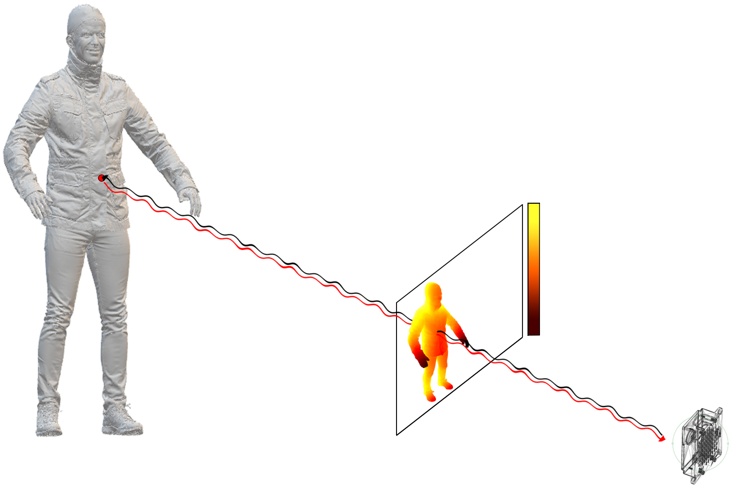


Figure 1.10: Time-of-flight approach. The black arrow indicates the emanated light signal path. The red arrow indicates the received light signal path. The image is adapted from [29].

less useful for body measuring and quasi-static scanning.

Refer to Appendix B for a summary of time-of-flight-based 3D scanning equipment.

1.3.4 Comparison Between the Scanning Technologies

Table 1.4 provides a comparison of the three scanning technologies. To triangulate 3D locations in space, PS and SL rely on establishing correspondences between the views, whereas ToF uses time-to-distance conversion and hence does not have to rely on the correspondences. Potential (self-) occlusions between the images pose a typical problem for triangulation techniques and could lead to gaps in the 3D point cloud [48]. Utilizing several cameras or views (which can be accomplished by rotating the subject or the scanner) and the T-pose, which reduces self-occlusions, are two ways to deal with the occlusions.

SL and ToF use light sources. It benefits SL in sections of the body with low texture, but it also restricts its uses to particular interior lighting scenarios. Light is the cause of the interference issues for multi-ToF scanners. The light source has a limit on the SL and ToF scanning ranges. Although the optics theoretically limit the PS scanning range, for 3D human body scanning the practical limit is usually up to several meters. The entire scanning spectrum is appropriate for scanning the human body.

PS is the best option for scanning dynamically changing objects due to its quick acquisition time, strong overall reconstruction performance, and lack of light interference concerns [98, 99]. ToF has a high reconstruction frame rate, making it suitable for dynamic applications. Dynamic scanning with single-shot patterns is also possible with SL. Note that, compared to multi-shot patterns, single-shot SL patterns give lesser reconstruction accuracy.

Last but not least, SL is the preferred technique for quasi-static scanning and body measuring because it provides the best resolution and accuracy. The quantity of commercial SL

Table 1.4: Main properties of the three 3D scanning technologies with respect to human body scanning.

	Passive stereo	**Structured light**	**Time-of-flight**
Method	triangulation	triangulation	time-to-distance conversion
Illumination	passive (ambient)	active (visible, IR)	
Scanning range	several meters	$< 5\,\mathrm{m}$ (illumination source limited)	
Dynamic scanning	yes	yes (slower movement only)	yes
Accuracy range	mm - cm	μm - cm	mm - cm
Resolution range	mm	μm - mm	mm
Main issues	textureless body parts	light interference	lower resolution, multi-camera interference

scanners is another indicator of this[§]. The accuracy and resolution range of PS and ToF are comparable (see Table 1.4), while ToF often has a lower resolution.

1.3.5 Human Body Scanning

When employing 3D scanners, the human body may be measured either in a stationary position [32, 33, 100] or while it is moving [101, 102, 103]. In static scanning, the subject is instructed to strike a fixed stance and maintain stillness for the duration of the scan. Subjects may unintentionally move during acquisition for 3D scanners with longer acquisition times, such as handheld scanners or scanners with moving heads, which increases mistakes; we can distinguish such situations as quasi-static scanning. The most accurate body measurements can be obtained using static scanning, which is frequently used to create reasonably big and varied public 3D human body datasets [18, 19, 20, 32, 33]. Typically, scanning in motion is done with PS or ToF technologies. The most popular motion capture (MoCap) systems are PS-based and use markers affixed to the body to measure movement [34, 101, 102, 104, 105] . Other dynamic 3D scanning technologies [103] record a subject as they are moving and examine the changes in soft tissue over time [106].

Typically, scanning yields a set of RGB pictures, a 3D point cloud, or one or more depth maps. In the case of dynamic scanning, so-called 4D scans are obtained [103] (where the fourth dimension is time). Some or all of these data are used in the processing stage to extract the measurements.

[§] See Appendix A for more details.

1.4 Statistical Human Body Models

Statistical models represent the population of human bodies with regard to pose and shape variations. The shape variations are usually represented by the principal components (PCs) of the mesh vertices' offsets from the 3D mesh that represent an average human body, while the pose variations represent rotations for the selected subset of human joints. Each scan in a dataset must undergo the mesh fitting technique in order to generate a statistical model. The mesh being fitted or deformed to each 3D scan from the scanning dataset is called a *template mesh*.

1.4.1 Template Mesh Fitting

Deforming a 3D template mesh that most accurately depicts an input is done using a collection of techniques called mesh or model fitting. A 3D scan, 2D or 3D keypoints, or a silhouette can be used as the input(s) to the mesh deforming procedure. When estimating body measurements, template meshes have the advantage of having a fixed number of vertices and matching vertices with the same semantics across all registered meshes in the dataset. When body measures have been determined for one mesh, they can be determined similarly for all the meshes. We differentiate between mesh regression and mesh fitting (registration or deformation) using statistical models.

More precisely, mesh fitting is an optimization process of deforming an initial, template mesh to the 3D scan[¶]. Pose and shape fitting make up mesh fitting [34, 108, 109, 110, 111]. A 3D scan is often subsampled before optimization so that the number of points is equal to or greater than the number of vertices in the template mesh [34, 40]. First, the 3D scan and mesh are roughly aligned using the landmarks [112] (pose fitting step). The body skeleton components of the template mesh are then rigged [113] before surface points are skinned using linear blend (LBS) [34, 109] or dual quaternion skinning (DQS) [111]. Once the pose satisfies the convergence criterion, shape fitting is done using a non-rigid registration, minimizing a loss function that usually consists of three components: a landmark term, a smoothness term, and a data term. The landmark term minimizes the distance between the corresponding landmarks of the template mesh and the 3D scan. The smoothness term minimizes the difference between the spatial transformations of the neighboring vertices. Finally, the data term minimizes the distances between the corresponding vertices. Note that the correspondence is determined at the beginning of a shape-fitting phase. Pose and shape fitting are typically performed back-to-back until final convergence [112]. Some works [19] additionally consider texture, which enhances fitting. When employing high-quality scanners, the described fitting is the go-to method for 3D

[¶] For simplicity, we only describe mesh fitting on 3D scans, but similar techniques can be applied to features or images [107].

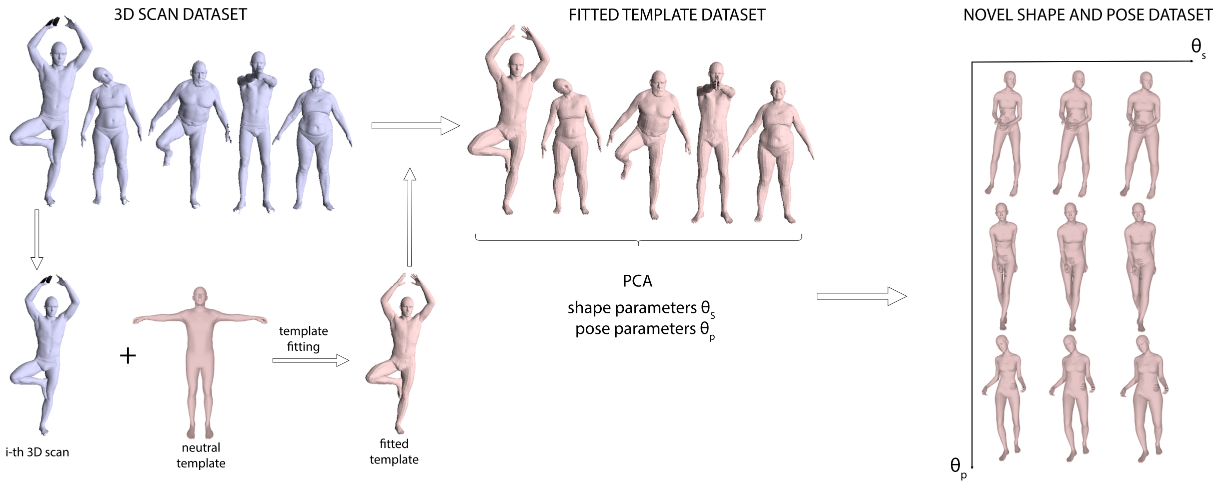


Figure 1.11: A visual summary of the mesh fitting procedure for building statistical models. A neutral template mesh is registered to each 3D scan in the scanning dataset, creating a dataset of registered template meshes. PCA can be used to develop the statistical model based on the shape and pose variations of the registered templates. The pose-shape space can be utilized to create new 3D meshes using the principle components. The FAUST dataset [19] contains the 3D scans and template models. SMPL-X is used to create novel 3D meshes [110]. The image is adapted from the review paper [29].

scans that are practically entirely complete. A clean mesh that fills in the gaps in the first, noisy 3D scan is the end result of fitting.

The principal component analysis (PCA) is used to explain shape changes in the collection of fitted template meshes of the statistical model. Finding the shape principle components that account for the majority of the dataset's variance allows PCA to condense the dataset of registered meshes. The ability of PCs to create innovative template meshes [36, 105, 114, 115] from a shape parameter space is a significant benefit of PCA. By also specifying the joint rotations of the mesh via pose parameters, a full 3D body mesh can be produced. The CAESAR [32], Size-UK [33], ScanDB [18], and potentially other datasets containing 3D scans [19, 20, 34, 35, 116] are the datasets that are frequently utilized for creating statistical body models (SMs).

1.4.2 The Examples of Statistical Body Models

The first SM for pose and shape deformations, as well as pose-dependent shape alterations, is SCAPE [34] (for example, muscle contractions in different poses). They construct about 150 additional markers using a set of initial physical markers and the correlated correspondence algorithm [117]. When they want an articulated human model, they use non-rigid registration. Each body part is rotated independently in SCAPE, which causes artifacts near joints and is one of its main drawbacks. The artifacts issue is resolved by BlendSCAPE's [108], which smooths SCAPE body part segmentations across part boundaries. Triangle deformations are used for PCA in both BlendSCAPE and SCAPE, which is a drawback. SMPL [109], one of the most well-known statistical models, has shown that utilizing vertex transformations rather than triangle transformations enhances the final SM. In order to create models that are more aesthetically acceptable for animation, SMPL also requires body symmetry. However, maintaining the symmetry comes at the expense of realism in some stances. A STAR [111] model that imposes spatially local and sparse pose corrective blend shapes and is unrelated to the symmetry optimization component is an improvement over SMPL. The fact that STAR was created utilizing the largest database, a mix of the SizeUSA dataset (5000 scans) and the CAESAR dataset (4000 scans), makes it the most expressive SM (9000 scans).

1.4.3 Mesh Regression from 3D scans

A statistical model can be utilized for mesh regression after it has been constructed. The goal of mesh regression is to identify the SM's pose and shape characteristics that best match an input. Some of the significant prior works on mesh regression from 3D scans are given in the continuation.

Volumetric-template-fitting, used by Kwok et al. [118] as an illustration of such a method, entails iteratively choosing the mesh from the statistical pose-shape space and tailoring the clothing to match the input 3D image. Prokudin et al. [119] provide a deep learning model for template fitting that is supervised by SMPL templates that were fitted to the dataset prior to learning (efficient-learning-with-basis-point-sets). The basis point set, a set of 3D scan features, and the ground truth template mesh are the two points on which the learning is based. The (slow) rendering step that is required to verify the parameters is omitted when employing the characteristics to identify optimal parameters.

1.5 Human Body Measurement

1.5.1 Body Measurement from a 3D Scan

The measurements can also be taken directly from a 3D scan after the body has been 3D scanned. The use of landmarks can aid in measuring circumferences and some distances [120]. The circumferences are estimated from a point cloud using a convex hull polygonal approximation method in Lu and Wang's work [30]. Using a perpendicular plane to slice the point cloud, the circumferential points are obtained. The highest X coordinate point is when the algorithm begins (Fig. 1.12b). The counterclockwise direction's next point is chosen as the one with the smallest angle between the Y-axis and the line connecting the current point X and the following point (Fig. 1.12b). Up till the polygon is closed, the operation is repeated. The sum of the line lengths between the chosen locations serves as an approximation of the circumference.

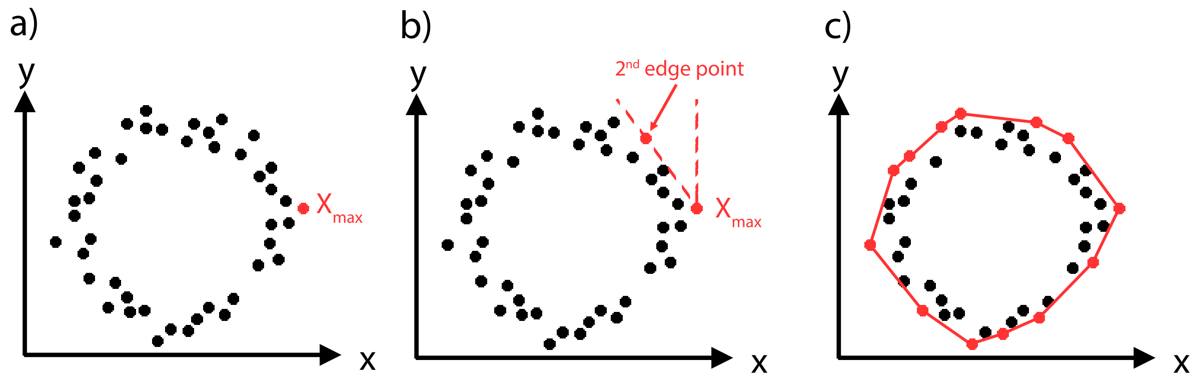


Figure 1.12: The convex hull polygonal approximation method. Adapted from [29].

1.5.2 Body Measurement from Template Mesh

Once the template mesh is estimated from either 3D or 2D data, body measurements can be extracted as illustrated in Fig. 1.13. A detailed description of how to extract specific body measurements is given in Chapter 3, where the linear regression model for body measurement estimation is also proposed and described.

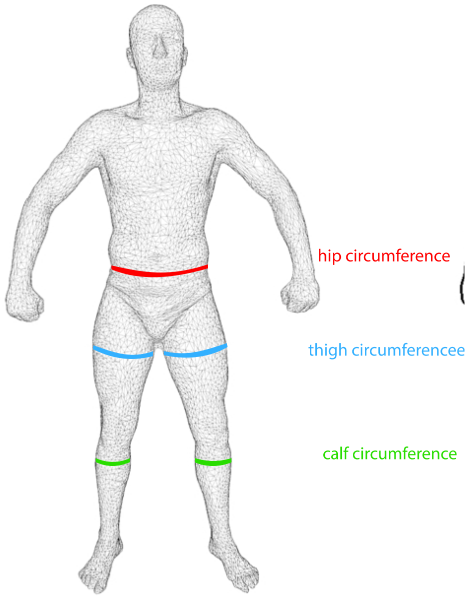


Figure 1.13: A visualization of body measurement extraction from the template mesh. Adapted from [29].

1.6 Five Steps of Digital Anthropometry

According to the review paper on digital 3D human body measurement, the body measurement processing pipeline can be divided into five steps: (1) preparation, (2) scanning, (3) feature extraction, (4) model fitting, and (5) measurement extraction (Fig. 1.14). Markers that indicate

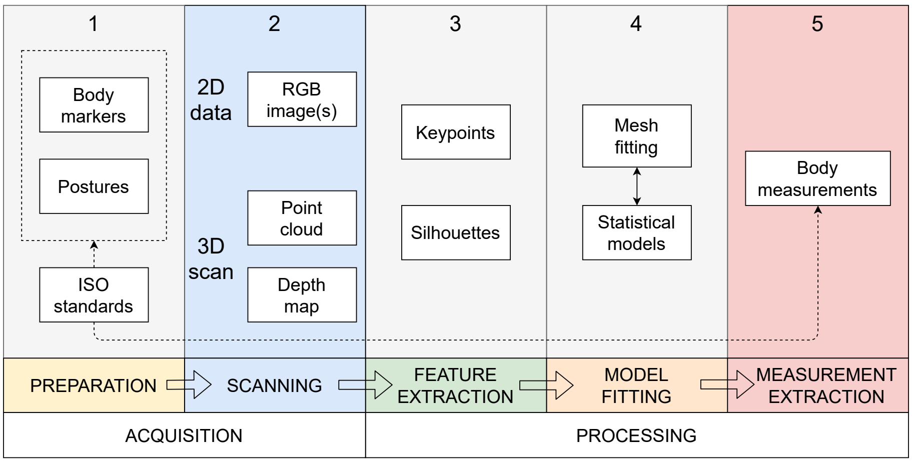


Figure 1.14: The five steps for digital anthropometry proposed by [29]: preparation, 3D scanning, feature extraction, model fitting, and measurement extraction. The image is adapted from [29].

common body landmarks may be applied to the body during preparation (step 1) [19, 32, 120]. The subject is instructed to strike a specific position [10] and remain motionless for the duration of the scan. If RGB cameras are employed, scanning (step 2) generates a 3D point cloud or depth map(s) together with the collection of images. Step 3 involves taking 3D scan and picture data and extracting features like keypoints and silhouettes. In step 4, the ideal human 3D template mesh is estimated using the characteristics or raw picture data [107]. The main benefit of using a model as a template and adapting it to a 3D scan is that any measurement can be conveniently and easily inferred from the semantics of the model. Statistical body models can be created using mesh fitting techniques, as described in Section 1.4. Direct template mesh regression from photos and image characteristics is made possible by the statistical models. In step 5, the data that has been processed (the 3D scan, pictures, features, and template mesh) is used to extract body measurements.

Chapter 2

Image-Based Anthropometry

A collection of methods for estimating body measurements using images and image attributes is known as image-based anthropometry, which is the main focus of the dissertation. Body measurements can be calculated either directly from photos or indirectly from extracted image features like keypoints and silhouettes [29] and then regressing the pose and shape parameters of the statistical body models (Section 1.4).

Note that the keypoint detection algorithms and models described in this chapter are specific to human bodies, i.e., they represent the selected subset of human joints [35, 121]. In contrast, the standard keypoint detection algorithms [52, 53] find the keypoints which contain distinctive information compared to the rest of the image, such as edges. The keypoints found using the latter algorithms struggle more to find the corresponding keypoints across images of the common scene [122]. The former algorithms know the correspondences in advance, but struggle with estimating the exact joint location, which degrades the final 3D pose estimation (Chapter 4).

We initially provide a broad overview of the methods and information utilized for image-based body measurement (Section 2.1). Then, we go over backbone deep learning architectures used for human body measurement, pose, and shape estimation (Section 2.2). In Section 2.3, we cover the feature extraction techniques. Finally, in Section 2.4, a summary of the previous synthetic data used for training and evaluating the suggested models is provided. Note that, in Section 5.3.2, we propose and describe a novel dataset designed for body measurement and shape estimation of clothed people.

2.1 The Overview of Image-Based Anthropometry

We categorize image-based anthropometry:

- in the context of other (semi-automatic) body assessment approaches/techniques;
- the surroundings and recording (imaging) conditions, such as the number of views, syn-

thetic, i.e., real data, the kind of clothing, etc.;

- input features and the pipelines used to estimate each body measurement.

We link certain subjects to the dissertation's corresponding parts. Note that the three research contributions should be highlighted:

- self-estimation of body measurements using linear regression model;
- pose estimation from multi-view images;
- shape estimation from a single image.

2.1.1 Body Measurement Approaches

Fig. 2.1 shows an overview of anthropometric approaches w.r.t. the input data and the technique used to extract the body measurements. The upper part of the diagram highlights the techniques which directly obtain body measurements, while lower part shows techniques that first estimate the human body model, and then extract the body measurements from it, as will be described in Section 3.2.1. Body measurements can be obtained either manually (traditional anthropometry described in Section 1.1), extracted from 3D scans (Section 1.5.1), self-estimated (Chapter 3), or estimated from images (Chapter 5). We propose two novel approaches, one for self-estimation of body measurements (Contribution 1 in the Figure) and one for body model (shape) estimation from images (Contribution 3 in the Figure).

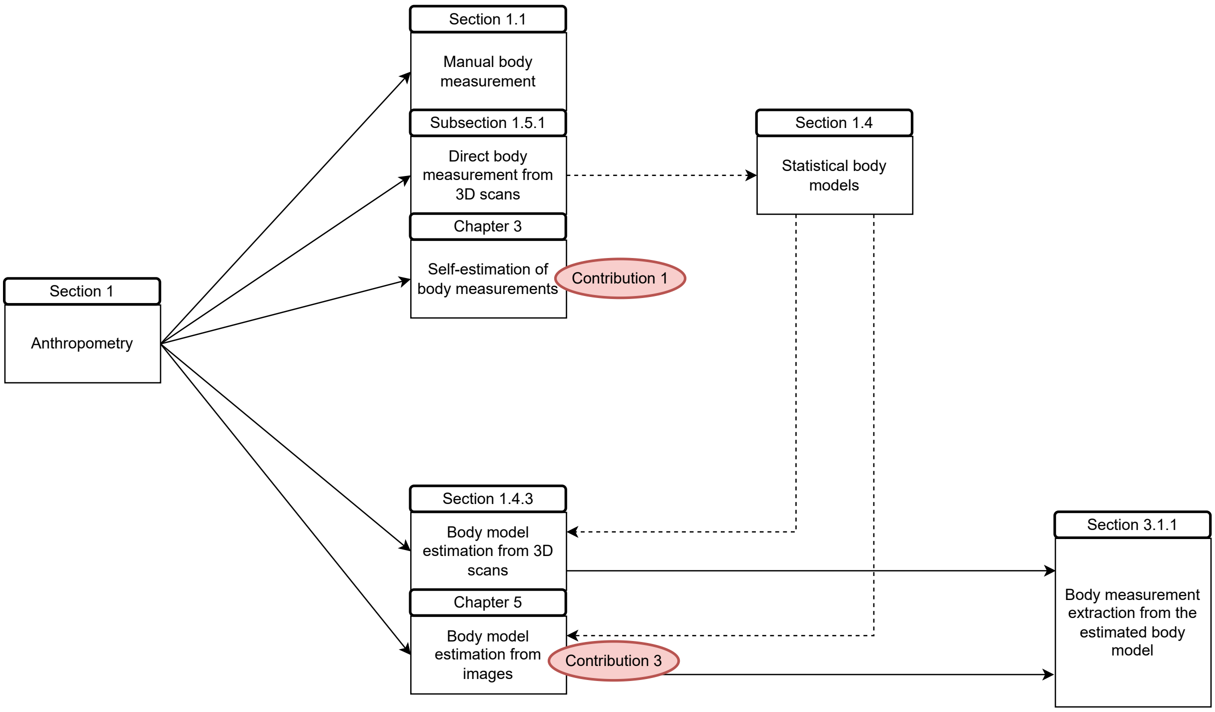

Section 1.1
Manual body measurement
Subsection 1.5.1
Direct body measurement from 3D scans
Chapter 3
Self-estimation of body measurements
Contribution 1
Section 1.4
Statistical body models
Section 1
Anthropometry
Section 1.4.3
Body model estimation from 3D scans
Chapter 5
Body model estimation from images
Contribution 3
Section 3.1.1
Body measurement extraction from the estimated body model

Figure 2.1: An overview of anthropometry w.r.t. input data and the technique used for body measurement. In particular, body can be measured manually (traditional anthropometry), and can be measured from 3D scans and 2D images. Based on the given data, different general techniques can be applied, as shown highlighted in the diagram.

2.1.2 Imaging Conditions

Image-based anthropometry can be analyzed with regard to imaging conditions, or the recording conditions in a more general sense (digital anthropometry and 3D scanning). Body pose, type of data (actual or synthetic), garments, number of views, and occlusions are five categories of body recording conditions that are highlighted in Fig. 2.2*.

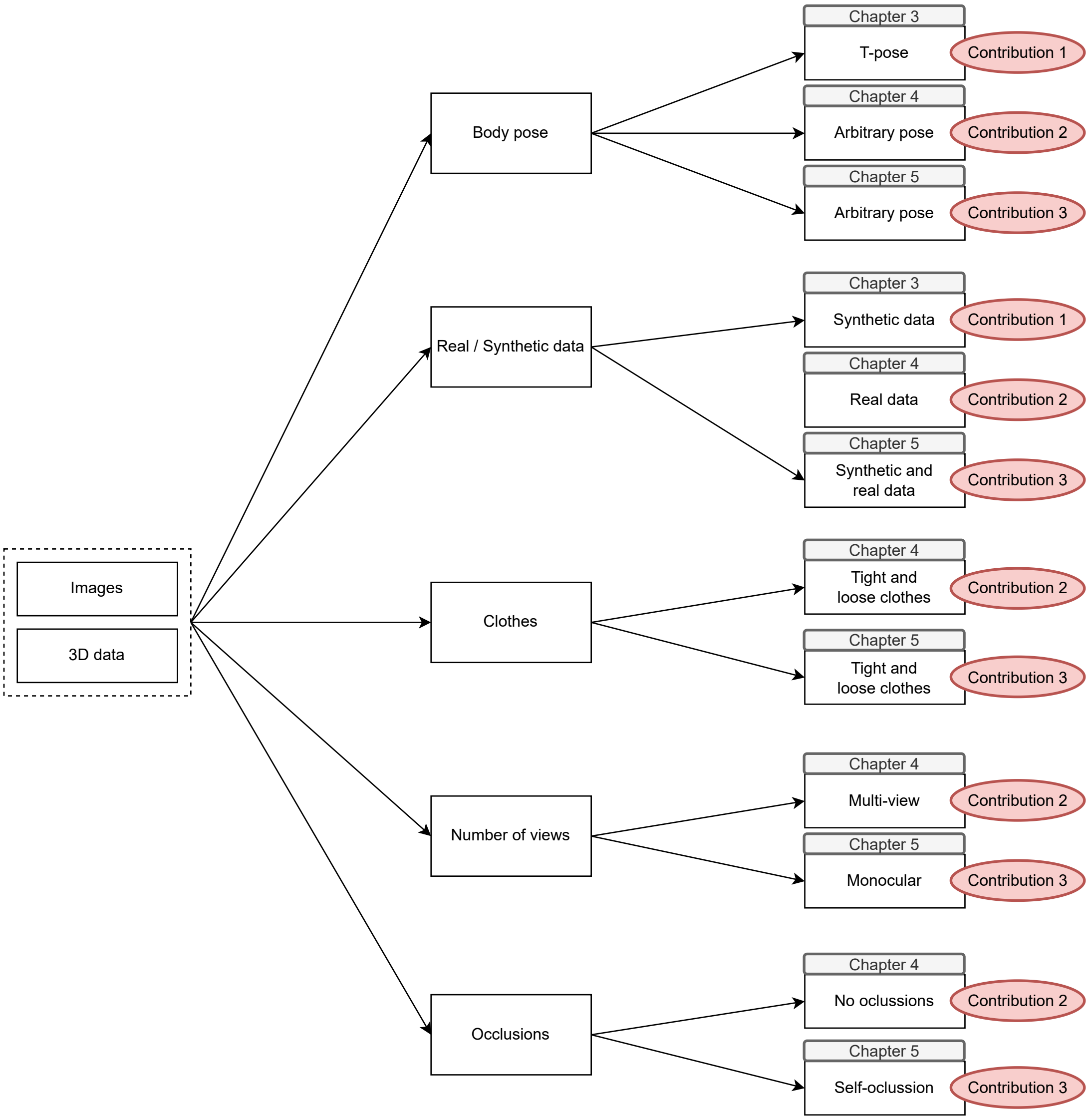


Figure 2.2: An overview of anthropometry w.r.t. various types of recording conditions, namely body pose, real or synthetic data, types of clothes (tight or loose), number of views, and occlusions.

Body pose. In the case of 3D scanning, the body pose is typically fixed and is one of the standard body poses [29]. On the other hand, the majority of cutting-edge image-based mesh regression techniques predict shape and posture parameters simultaneously, allowing one to extract body measurements in any pose using these models (see Chapter 5). Of course, there

*In general, one can also take into account other factors like illumination and image quality, but we emphasize these five because they are the dissertation's main topic are still several problems with precise joint assessment. Shape-in-a-fixed-position, or T-pose shape estimation, is the focus of a sizable portion of the dissertation (Chapter 5).

Real / Synthetic data. The identity obfuscation laws (GDPR), the difficulty of obtaining ground-truth body measurements for the people in the scenes (one would, for example, need to 3D-scan people and extract their accurate body measurements), and the difficulty of recording people of various shapes and in various poses make realistic datasets with people and their corresponding ground-truth body measurements particularly difficult compared to general 3D scene datasets [123, 124]. As a result, synthetic datasets [125, 126, 127] are a crucial tool for training and assessing models for estimating human pose, shape, and body measurement. We also dedicate a significant effort in using synthetic datasets for the purposes of this dissertation. In particular, we describe previous synthetic datasets in Section 2.4, extend previous body measurement dataset for the purpose of fitting our linear regression model in Chapter 3, and propose a novel synthetic dataset of clothed people for the estimation of shape and clothes in Chapter 5.

Clothes. The quantity and type of clothing a person wears is an important factor to take into account. Assuming that the clothes do not significantly alter the 3D body shape, it is nevertheless expected that the subject be either naked (swimsuit) or in tight clothing for the purpose of estimating shape and body dimensions. Many previous methods are limited to tight clothes [128, 129, 130, 131]. We go beyond and propose pose (Chapter 4) and shape (Chapter 5) estimation approaches from images of people in either tight or loose clothes.

Number of views. The majority of state-of-the-art mesh regression techniques are monocular [107, 131, 132, 133]. By utilizing different viewpoints, 3D geometry and geometric techniques like triangulation of the matching keypoints [46, 134, 135] are made possible. This methodology is particularly useful for 3D human posture estimation [134, 135, 136, 137, 138]. Based on any number of available views, a different collection of probabilistic algorithms estimates the most likely posture and shape parameter hypothesis jointly [129, 130]. We utilize a probabilistic model for monocular estimation of human shape and several perspectives to enable the triangulation of the 3D human position (Chapter 4). We contend that by integrating these two methods, 3D posture and shape estimation would be enhanced in the future (Chapter 6).

Occlusions. The (self-) occlusions are disregarded in a controlled environment, which is typically a 3D scanning environment or an image recording in a scene without objects because they are not a significant factor in body measurement estimate. Contrarily, occlusions are a significant component of in-the-wild image-based anthropometry, particularly in indoor scenarios with a lot of 3D objects and in cases with multiple views (the likelihood that some part of the body will be obstructed in some of the views increases with the number of views). A common method for dealing with occlusions is to give the model the ability to communicate how confident it is in its predictions, either at the level of 2D keypoint detection [139, 140] or at the level of body shape and pose parameters [128, 129, 130], where the occluded parts are likely to be less confident. We don't specifically focus on training the shape estimation model to handle occlusions (Chapter 4 and 5), leaving that to future study. We also assume no substantial occlusions in our 3D pose estimation model (Chapter 4).

2.1.3 Feature Types

Keypoints and silhouettes, two popular types of features [36, 131], can be retrieved from images and 3D data. A very similar effect is produced by edges [128], which we explore and experiment with in Chapter 5.

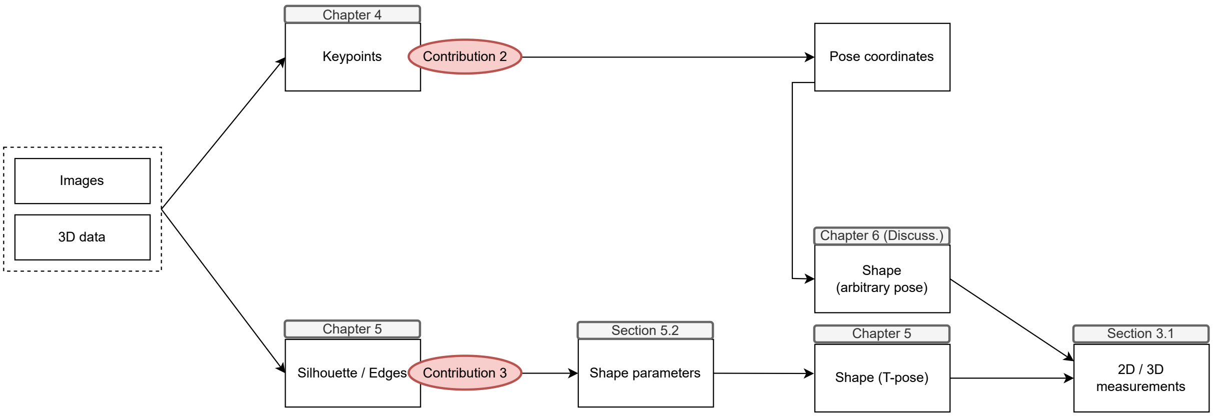


Figure 2.3: The overview of anthropometry w.r.t. to feature types used to estimate the body measurements.

There are several popular pose parameter estimation approaches based on image keypoints [141, 142, 143, 144]. Once the pose parameters are specified or estimated, the 3D pose coordinates can be easily derived from the corresponding 3D human mesh [†] (see the upper portion of Fig. 2.3). In Chapter 4, we propose a straightforward 3D pose estimation technique. Pose estimation, on the other hand, enables in-the-wild body measurement (arbitrary pose), which we briefly discuss regarding future approaches in Chapter 6.

In Chapter 5, we propose a deep learning approach for estimating shape parameter from edges[‡]. Based on the estimated shape parameters, we derive body measures from the calculated shape parameters in accordance with the requirements explained in Section 3.2.1.

2.2 Backbone Architectures for Feature Extraction

Convolutional neural networks (CNNs) are the essential and fundamental building blocks for keypoint and semantic estimation tasks in computer vision. Note that whereas the human silhouette is a binary mask (single-class), the semantic segmentation mask is typically a set of

[†] The non-estimated shape parameters can be set to mean (zero) shape, which corresponds to $\beta = \mathbf{0}$.

[‡] Note that similar approach could be proposed using silhouettes, as suggested and demonstrated in [131].

binary masks (multi-class).

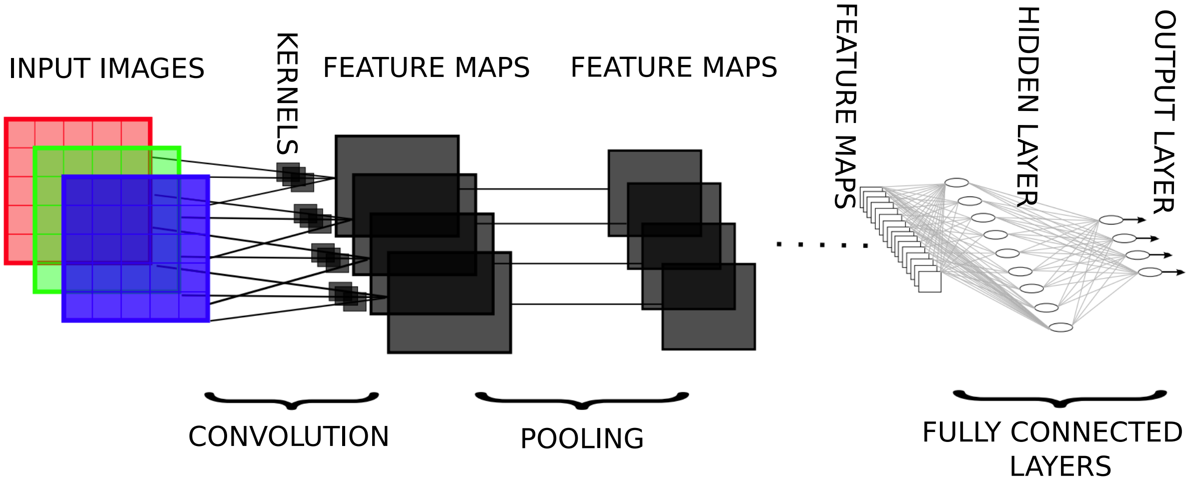


Figure 2.4: The typical, generic convolutional network architecture for RGB image processing. Adapted from [145].

The typical convolutional neural network architecture consists of three types of layers:

- convolutional layer helps to abstract the input image as a feature map via the use of kernels.
- pooling layer helps to downsample feature maps by summarizing the presence of features in patches of the feature map.
- fully connected layer connects every neuron in one layer to every neuron in another layer.

2.2.1 Hourglass Models

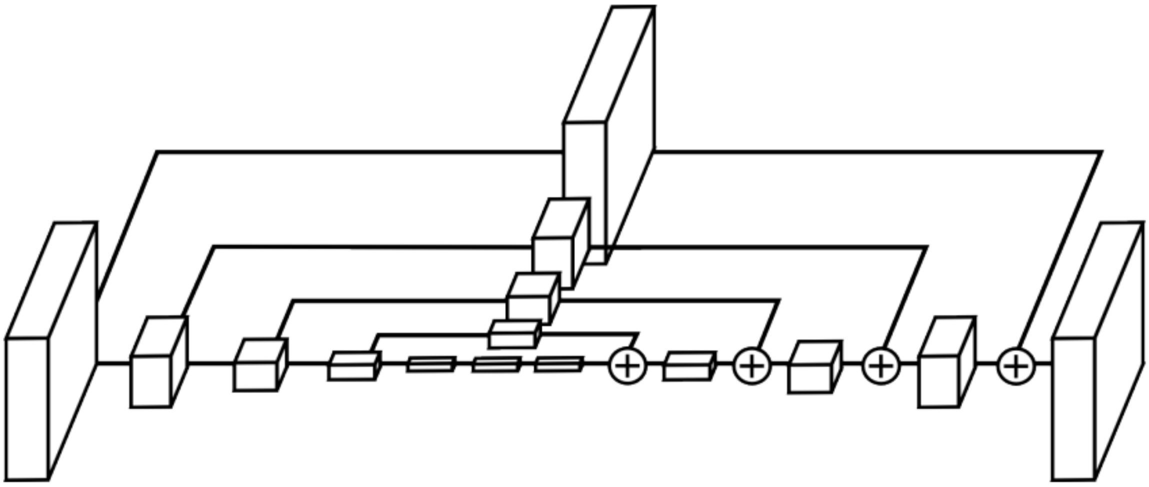


Figure 2.5: A single block of an hourglass architecture, which consists of several downsampling and upsampling layers. The downsampling and upsampling layers are concatenated as shown in the image (the plus (+) signs). Note that the image is adapted from [146].

The stacked hourglass architecture [147] and the concept of "hourglass" (see Fig. 2.5) is the base idea for the dominant approaches [146, 148, 149] on MPII 2D keypoint estimation benchmark [150]. It consists of a multi-stage architecture with repeated bottom-up, top-down processing and a skip (residual) layer feature concatenations (see Fig. 2.5 as an example of a single block).

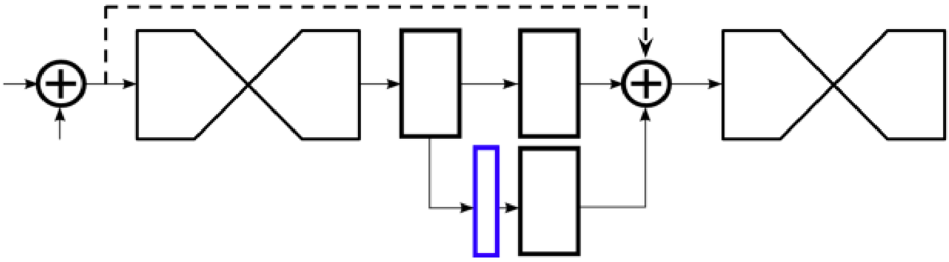


Figure 2.6: The immediate supervision is used after each hourglass block. The plus (+) signs represent residual (skip) connections between the layers. The contracting-expanding blocks represent the blocks of the hourglass architecture, shown in Fig. 2.5. The remaining black rectangles represent other network layers whose internal structure is not relevant to the overall description. The narrow rectangle shows the immediate supervision layer, which is one of the key characteristics of the hourglass architecture. The image is adapted from [147].

Another key feature of the stacked hourglass architecture is the use of immediate supervision after each hourglass block, as shown in Fig. 2.6. Note that there are usually several hourglass blocks in the hourglass model. The idea of the immediate supervision strategy is that loss is applied both on coarse and more refined feature maps, which proved beneficial for the model's performance.

2.2.2 Cascaded Pyramid Networks

Cascaded pyramid network (CPN) [151] was the leading method on COCO 2017 keypoint challenge [152]. Similar to hourglass architecture, CPN also involves skip (residual) layers and 2D keypoint detection step, where the model is expected to detect visible and, in general, "easy" keypoints. This is called the GlobalNet part of the architecture (see Fig. 2.7). The keypoints that are occluded and more difficult to detect based on the image appearance features are left to the RefineNet part of the architecture (Fig. 2.7).

The role of RefineNet is visualize even better in Fig. 2.8. RefineNet takes the heatmaps for all the keypoints and exploits all these information in order to reason about the "hard" keypoints that are not distinguishable based on image appearance only. This more abstract reasoning about the human body and pose is usually called the *human pose prior*. We specifically analyze our proposed 3D pose estimation model from Chapter 4 w.r.t. human pose prior.

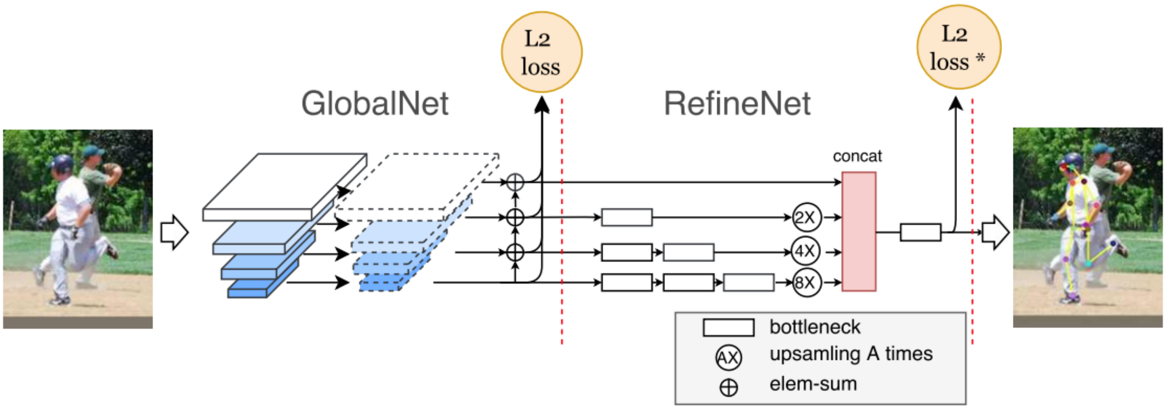


Figure 2.7: The Cascaded Pyramid Network (CPN) consists of GlobalNet and RefineNet. Note that the crossed-out blocks in the GlobalNet part simply point out that there are multiple convolutional layers in this part of the network, and the plus (+) symbols point out that these layers also contain the residual (skip) connections. The GlobalNet output is immediately supervised via L2 loss and propagated to the RefineNet part. The RefineNet is particularly useful to localize "hard" keypoints, which is further explained in the Fig. 2.8. The image is adapted from [151].

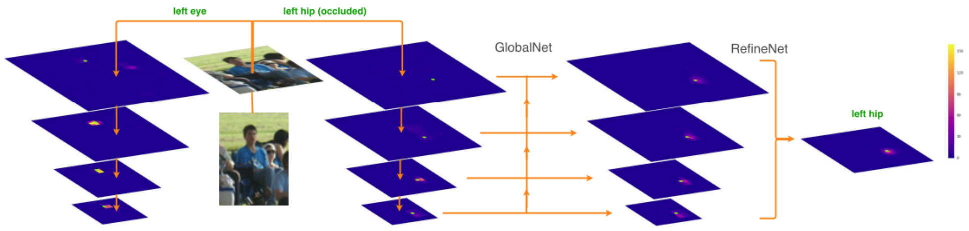


Figure 2.8: The RefineNet part of the CPN architecture is particularly useful for localizing the "hard" keypoints. In the case of CPN, these keypoints are the ones that require a larger perceptual context from the image, such as hips. For example, in the image above, the left hip is occluded and not well localized in the GlobalNet part. In order to improve the localization, the features of lower resolution are upsampled and then concatenated with the ones of higher resolution in order to finally produce the refined location of the left hip. Adapted from [151].

2.2.3 Using Deconvolutional Layers

Compared to the stacked hourglass and cascaded pyramid networks, Xiao et al. [139] proposed a much simpler architecture for 2D keypoint detection. Instead of stacking several downsampling and upsampling hourglass blocks, i.e., cascading (concatenating) multi-resolution heatmaps, they propose to use deconvolutional layers (Fig. 2.9) to replace the upsampling parts. The architecture of their model is similar to the previously described architectures, consisting of several encoder-decoder blocks with residual layers. Their proposed model demonstrated state-of-the-art performance for 2D keypoint detection on COCO 2017 challenge [152]. Their architecture is used as a backbone for both the proposed model in Chapter 4 and 5.

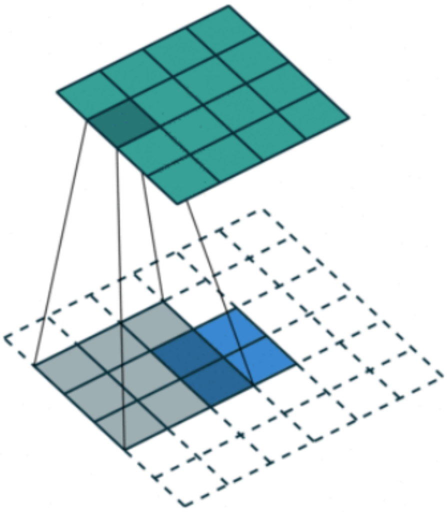


Figure 2.9: The illustration of deconvolutional layer, which was the main ingredient in the architecture proposed by Xiao et al. [139]. The image is adapted from the same paper.

2.3 Feature Extraction

Two types of features that are usually extracted from 3D scans and images are keypoints and silhouettes. The location of keypoints[§] can be determined based on markers or can be estimated automatically from a 3D scan [30, 31]. Silhouettes either represent the points or pixels for the whole human body, or the body segments.

2.3.1 Keypoint extraction

The majority of keypoint estimation algorithms can identify human joints from images of a person. The keypoints might be represented as 3D points in the scene or as 2D pixel coordinates in an image. The accuracy of the estimation can be increased if there is a moving subject by using temporal smoothness and taking advantage of the time component [153]. Therefore, keypoint estimation methods can be divided into: single-image [140, 154, 155, 156, 157], multi-frame [153], and multi-view methods [134, 136, 137] for 2D [140, 154, 155, 156, 158] or 3D [134, 136, 137, 153, 157] keypoint estimation. As seen in Fig. 2.10, the keypoint estimation algorithms typically find between 14 and 21 keypoints. Due to the availability of large annotated datasets, the majority of cutting-edge keypoint estimate techniques are deep learning-based [35, 116, 125, 159, 160]. Typically, landmark extraction from 3D scans is not integrated with the extracted 2D and 3D keypoints for mesh fitting [111, 142, 144] (see the following subsection).

Motion capture [101] is a movement tracking method that allows for the direct gathering of ground truth coordinates for 2D and, in particular, 3D keypoints[¶]. The majority of the keypoint

[§] Note that keypoints are called landmarks if they refer to standardized body locations [10].

[¶] Another way to obtain 2D pose estimation data is to manually label human joints on a large number of images.

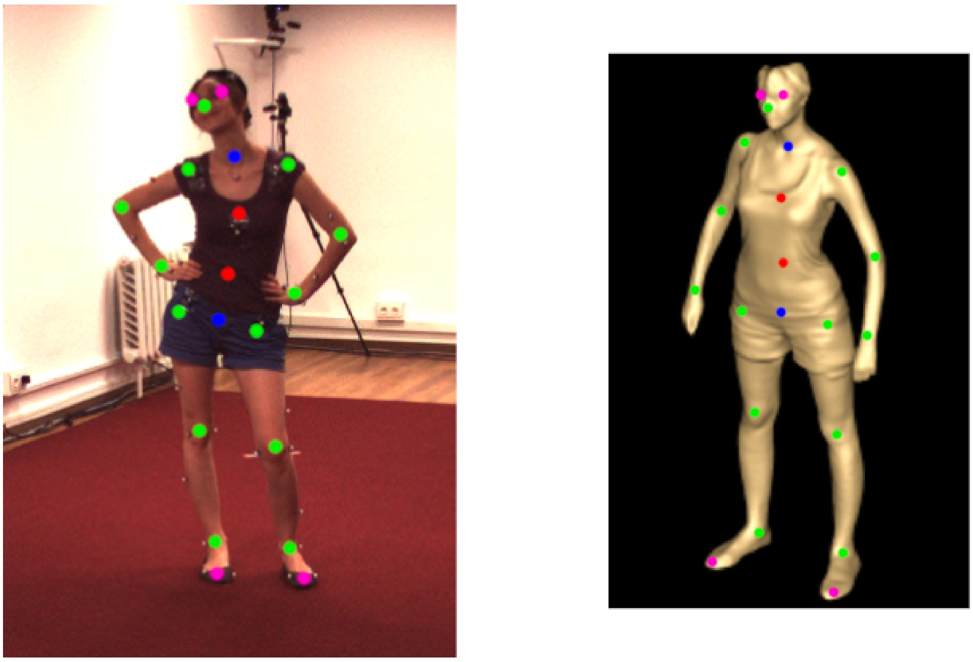


Figure 2.10: The example of keypoint detection in 2D image (left) and on the point cloud scan (right) corresponding to the person in the image on Human3.6M dataset [35].

estimation algorithms mentioned earlier benefit from the ground truth information obtained from motion capture. Human3.6M dataset [35], HumanEva [116], and TotalCapture [101] are a few examples of MoCap datasets. Motion capture systems' drawback is that they can't be used in real-world situations.

2.3.2 Silhouette extraction

Methods for extracting silhouettes separate pixels that depict pixels of interest (a person) from other pixels in the image [161]. Modern semantic segmentation techniques that are related to human pose estimation, such as cross-domain multi-person part segmentation [162], self-correction for human parsing method [163], dense image prediction [164], body part parsing posture-guided method [165], and joint pose and part segmentation [166], are likewise deep learning-based. There are body-part-segmented datasets [167, 168] in addition to whole-body segmentation [121, 169]. Both whole body and body part segmentation methods achieve relatively high accuracy[‖], even on difficult examples. Therefore, semantic segmentation models are convenient to be used as part of the body measurement estimation from images pipeline [131]. Note that we do not use semantic segmentation in this dissertation, because edge detection has proven to be even more beneficial for human pose and shape estimation from images, i.e., edges [128].

However, this is impractical and unreliable in case of 3D data.

[‖] The accuracy is measured as a mean IoU (intersection over union).

2.3.3 Edge Detection in PyTorch

In this subsection, we are going to describe canny edge detector [170], the most popular edge detector algorithm. We use edge detection in Chapter 5 and use the extracted edge maps as input features to learn human body shape from images. The canny edge detection algorithm consists of five steps: applying the Gaussian filter, finding the intensity of the gradients of the image, applying gradient magnitude thresholding, applying a double threshold to determine potential edges, finalize the detection.

Gaussian filtering. In order to avoid false detection caused by noise, which can easily impair any edge detections, the noise in the image must be filtered out. A Gaussian filter kernel is convolved with the image to smooth it out. In order to lessen the impact of significant noise on the edge detector, this step will slightly smooth the image. The formula for a size-dependent Gaussian filter kernel is explicitly given by:

$$G_{ij} = \frac{1}{2\pi\sigma^2} \exp\big(-\frac{(i-(k+1))^2 + (j-(k+1))^2}{2\sigma^2}\big), \tag{2.1}$$

where the Gaussian filter is of size $(2k+1) \times (2k+1)$, $1 \leq i$, and $j \leq (2k+1)$. The performance of the canny edge detector depends on the kernel size. In general, the detector's sensitivity to noise decreases with increasing the kernel size. However, the edge localization error will slightly increase with larger kernel sizes.

Intensity of the image gradients. The Canny algorithm employs four filters in order to identify horizontal, vertical, and diagonal edges in the blurred image because an edge in an image can point in a number of different directions. The edge detection operator [171] returns a value for the first derivative in the horizontal direction (G_x) and the vertical direction (G_y). From these values the edge gradient and direction can be determined:

$$G = \sqrt{G_x^2 + G_y^2}$$

$$\Theta = \arctan(G_y, G_x)$$

The edge direction angle is rounded to one of four angles representing vertical, horizontal, and two diagonals.

Gradient magnitude thresholding. To locate the areas with the sharpest change in intensity value, the algorithm needs to assess how strong the edges of the current pixel are in comparison to how strong they are in the positive and negative gradient directions. The value will be retained if the edge strength of the current pixel is greater than that of the other pixels in the mask pointing in the same direction (for example, a pixel pointing in the y-direction will be compared to the pixels above and below it in the vertical axis). The value will be suppressed in any other case.

Double thresholding. Remaining edge pixels after the use of non-maximum suppression offer a more realistic representation of actual edges in an image. However, there are still some edge pixels with noise and color variance. Edge pixels with a low gradient value must be removed in order to account for these erroneous answers, while edge pixels with a high gradient value must be kept. By choosing high and low threshold values, this is achieved. A pixel is designated as a strong edge pixel if its gradient value exceeds the high threshold value. A weak edge pixel is one whose gradient value is greater than the low threshold value and lower than the high threshold value. An edge pixel will be suppressed if its gradient value is less than the low threshold value. The specification of the two threshold values, which are derived empirically, is based on the input image's content.

Finalizing the detection. As they have been retrieved from the true edges in the image, the strong edge pixels should undoubtedly be included in the final edge image. The weak edge pixels, however, can either be retrieved from the true edge or the noise/color variations. The latter causes' weak points should be eliminated to produce an accurate outcome. As opposed to noise responses, weak edge pixels resulting from real edges are typically coupled to strong edge pixels. Blob analysis [172] is used to track the edge connection by examining a weak edge pixel and its eight connected neighboring pixels. That weak edge point can be recognized as one that needs to be kept as long as the blob contains at least one strong edge pixel.

To implement the above edge detection algorithm in PyTorch [173], the main requirement is that all of the above steps need to be differentiable so that the gradients can backpropagate. For efficiency, the pipeline should be written for parallel execution, i.e., for-loops and similar sequential strategies should be avoided at all costs.

2.3.4 Mesh Regression

The main methodology used in this thesis for image-based body measurements is mesh regression (see Fig. 2.11). The mesh regression is a learning task that optimizes human pose and shape parameter estimation, i.e., the estimation of parameters from the statistical human body population (Chapter 1). The mesh regression can be done using features described in this section, such as keypoints, as shown in Fig. 2.11. The Figure highlights one of the first mesh regression approaches, called SMPL-X [110]. In Chapter 5 we propose and describe in detail a novel mesh regression approach for pose, shape, and clothes estimation from the features of a single image.

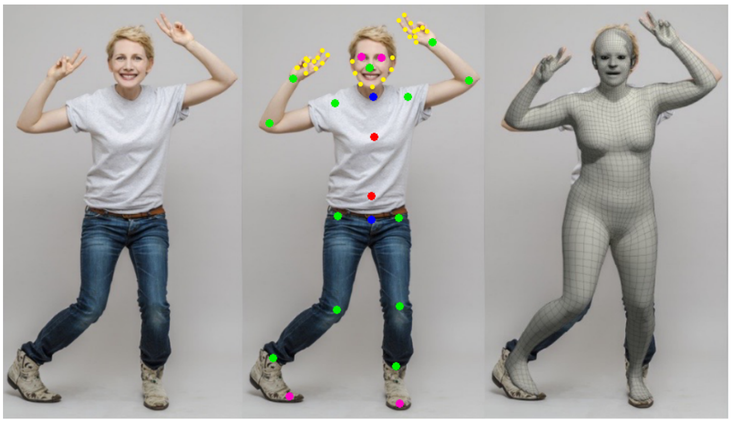


Figure 2.11: An illustration of a mesh regression method using SMPL-X that is based on 2D keypoint estimation [143]. Notably, the hands and face are represented by yellow keypoints, although typically being modelled independently. The image is adapted from [110].

2.4 Synthetic Data

The main advantage of using synthetic data is that perfectly accurate ground truth is available along with the images, such as depth, body parts, optical flow, 2D/3D pose, surface normals, keypoints, etc. For our image-based methods, we only use keypoints obtained from the given 3D template meshes (SMPL). The keypoint locations are transformed into heatmaps, as described in Chapter 5, which has proven useful for the performance of deep learning model.

In particular, the advantage of using 3D template meshes in combination with synthetic data is that the body measurements can be extracted in a standardized way, which guarantees a standardized measurement (Chapter 3). Another advantage is that synthetic data avoid privacy issues and approvals. Therefore, the promising future direction is the creation of large, more realistic, and more diverse synthetic anthropometric benchmarks. The statistical models such as SMPL are currently the best approximation of the overall population and are likely to be used as a tool for generating the body measurement benchmarks in the coming years.

2.4.1 SURREAL

SURREAL [125] is the first large-scale synthetic human dataset. It consists of 6M frames. The images are photo-realistic renderings of people under large variations in shape, texture, viewpoint and pose. To ensure realism, the synthetic bodies are created using the SMPL body model, whose parameters are fit by the MoSh [104] method given raw 3D MoCap marker data. The backgrounds are 2D images from LSUN dataset [174]. The advantage of SURREAL is that the renderings are diverse and controllable and that the random backgrounds further improve the diversity of the final images. Some of the disadvantages are that the body poses w.r.t. to the background scene are often unrealistic and that the lower quality of the textures further sacrifice realism. Lack of realism in the synthetic data is particularly limiting when training deep learning models for anthropometric application of real people. Finally, the textures are only applied on top of the SMPL mesh, i.e., no clothes geometry is available. Even though, SURREAL is still a very popular synthetic human dataset [128, 129, 130, 131]. The example images from SURREAL are shown in Fig. 2.12.

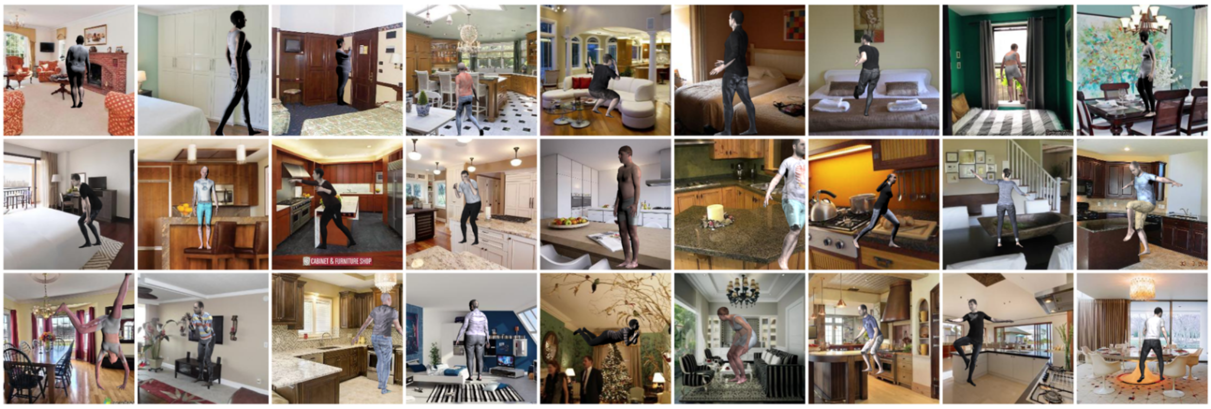


Figure 2.12: The examples from the SURREAL synthetic dataset.

2.4.2 3DPeople

3DPeople [160] is the first dataset of dressed humans with specific geometry representation for the clothes. It contains around 2M images with 40 male and 40 female performing 70 actions. The significance of this dataset, compared to SURREAL, is that the actual clothes geometry is added on top of the body, as shown in Fig. 2.13. This allows direct training of clothing-geometry-aware models, while still having all the advantages of SURREAL such as diverse human characters and backgrounds. However, 3DPeople does not use SMPL model, which prevents straightforward body measurement extraction and makes controllable character generation more complicated. We tackle some of these limitations in Chapter 5, where we propose novel synthetic dataset of clothed humans.

2.4.3 AGORA

AGORA is a highly-realistic synthetic human dataset. It consists of 4240 commercially available, high-quality, textured human scans in diverse poses and natural clothing; this includes 257 scans of children. These scans are rendered in diverse 3D scenes using Unreal Engine [175], as shown in Fig. 2.14. The dataset contains ground-truth 3D poses and body shapes by fitting the SMPL-X body model [110] (with expressive face and hands) to the 3D scans, taking into account clothing. AGORA has around 14K training and 3K test images by rendering between 5 and 15 people per image using either image-based lighting or rendered 3D environments. In total, AGORA consists of 173K individual person crops. The disadvantage of AGORA is that

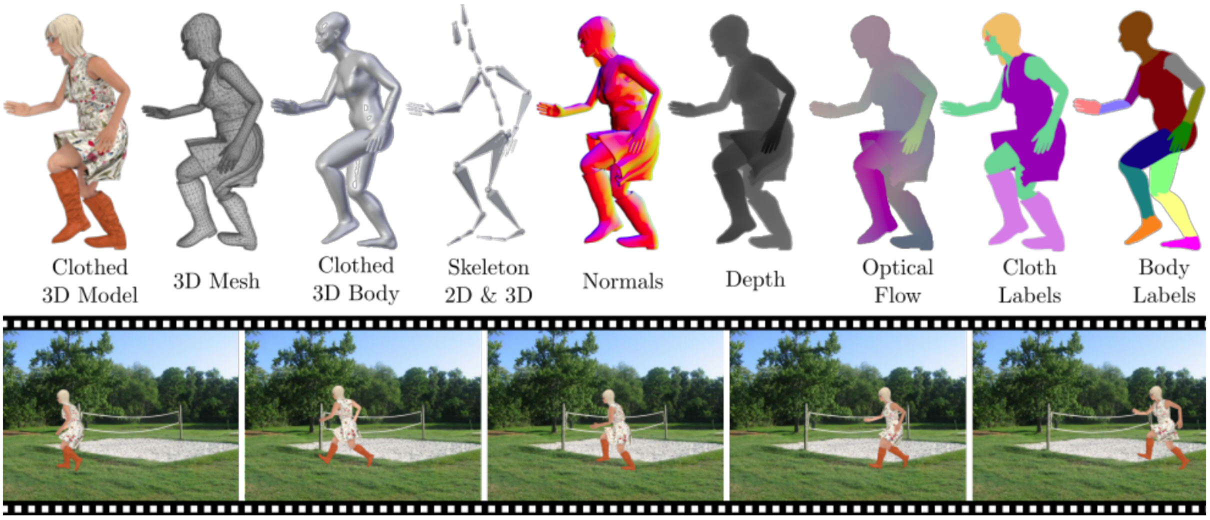


Figure 2.13: The example sequence and ground truth information given with the 3DPeople dataset. The image is adapted from [160].

the scans are not controllable which puts limits to potentially generate any number of different people in various poses and with different body measurements. We also tackle this limitation in Chapter 5.

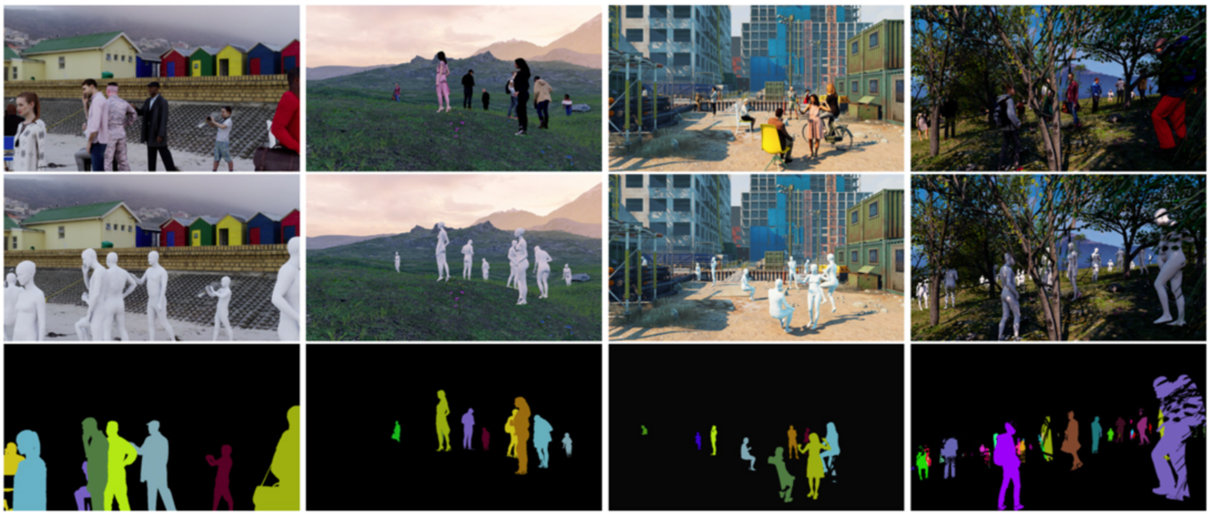


Figure 2.14: AGORA dataset populates highly realistic 3D scenes with high-quality 3D scans of clothed people.

Chapter 3

Body Measurement Estimation Baseline

In this section, we propose and describe a linear regression model that requires only the information that any person can self-estimate, such as height and weight [176]. By comparing the proposed model's performance to the state-of-the-art for body measurement estimation on two publicly available datasets, BODY-fit [177] and ANSUR [22], we show that it performs well. The linear model offers a clear and easy technique to estimate body measures, which makes it useful for augmented reality and virtual try-on.

3.1 Related Work

In this section, we briefly cover prior works relevant to the proposed body measurement estimation baseline. With respect to the baseline, we divide prior works into three groups: the ones that also propose linear regression for anthropometry, the nonparametric approaches, and the parametric approaches. We compare the baseline with each group of methods in the remainder of the chapter.

Linear Regression for Anthropometry. Based on the public databases, several works propose linear regression models for the estimation of measurements [178, 179, 180] and other body characteristics such as skeletal muscle mass [181]. We extend these analyses by focusing specifically on using self-estimated height and weight as input, as well as on using the statistical body models.

Nonparametric Approaches. With the advances in 3D scanning technology, more automatic approaches to body measurement have been proposed [182, 183, 184]. Most 3D-based body measurement methods use landmarks to determine distances and calculate measurements such as arm and leg length, shoulder-to-crotch, etc. Circumferences can be obtained by slicing the point cloud with a vertical plane and summing up distances between the nearest points [182]. The 3D-based methods are generally the most accurate among the (semi-)automatic body measurement methods. However, their main drawback is that they require 3D scanning, which is cumbersome and relatively expensive.

There are a number of image-based (2D) nonparametric models [107, 185, 186, 187] that freely deform meshes to best fit the input features. To improve convergence, they start with the template human mesh. However, the final deformed mesh does not necessarily retain the original vertex semantics. This property makes current image-based non-parametric models less suitable for body measurements. On the other hand, compared to their parametric counterparts, nonparametric models might have the advantage of better fitting out-of-distribution samples.

Parametric Approaches. The first ten principal components (shape parameters) of the SMPL parametric model are usually used for mesh regression. Tsoli et al. [188] first registers template meshes to the 3D scans and additionally learns features for body measurement prediction. BUFF [20] addresses the problem of 3D body shape estimation under clothing by allowing the mesh to deviate from the template but regularizes the optimization to satisfy the anthropometric constraints. Similar to non-parametric approaches, 3D parametric approaches are generally more accurate than image-based approaches, but also require 3D scans as input.

Image-based approaches (2D) can be divided into shape-aware pose estimation methods, which typically regress pose and shape parameters in-the-wild either from 2D keypoints or directly from images [107, 132, 133, 142, 143, 144, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205], and shape estimation methods which regress shape from silhouettes, usually in fixed pose and minimal clothing [21, 177, 206, 207, 208, 209, 210, 211, 212, 213]. We compare the proposed baseline against the state-of-the-art 3D- and 2D- based approaches for human body measurement estimation and achieve comparable performance to the best methods, while outperforming several deep learning models (see Sec. 3.3).

3.2 Linear Regression Model

In this section, we go over how to extract body measurements from a template SMPL mesh and the proposed linear regression model. For males and females, separate regression models were used. The proposed approach aims to show that body measures may be calculated without taking any further measurements using only the information that each individual normally knows about themselves, i.e., height and weight, making it appropriate as a baseline.

The method is shown in Figure 3.1. Each human body sample i is defined by 10 shape parameters, θ_{Si}, and 63 pose parameters, θ_{Pi}. The height and the 15 body measurements of the model are extracted from the template mesh. The input to the model consists of height (h) and weight (w). Weight should be available in the dataset; otherwise, it can be estimated from the calculated mesh volume. The output consists of the 15 other body measurements (A–O). The measurement names are listed in Table 3.1, and their extraction is described in

Section 3.2.1. Then, the linear model for the j-th measurement is described as:

$$y_j = x^T a_j + b_j, \quad j \in \{1, 2, \ldots, 15\} \tag{3.1}$$

where $x^T \in \mathbb{R}^2$ is a row vector consisting of the height and weight of the samples (an independent variable), $a_j \in \mathbb{R}^2$ is a column vector of the linear coefficients (the slope), $b_j \in \mathbb{R}$ is an intercept of the regression model, and $y_j \in \mathbb{R}$ is an output measurement (a dependent variable). These equations can be written more compactly in matrix form as:

$$Y_j = XA_j, \quad j \in \{1, 2, \ldots, 15\} \tag{3.2}$$

where $X \in \mathbb{R}^{N\times3}$ are the heights and weights from the N samples from the training dataset and $Y_j \in \mathbb{R}^{N\times1}$ are their output measurements. Additionally, a column of ones is added to X to account for b_j, which is now included in $A_j \in \mathbb{R}^{3\times1}$, representing all the model parameters.

Therefore, the least-squares closed-form solution is:

$$A_j = \left(X^T X\right)^{-1} X^T Y_j, \quad j \in \{1, 2, \ldots, 15\} \tag{3.3}$$

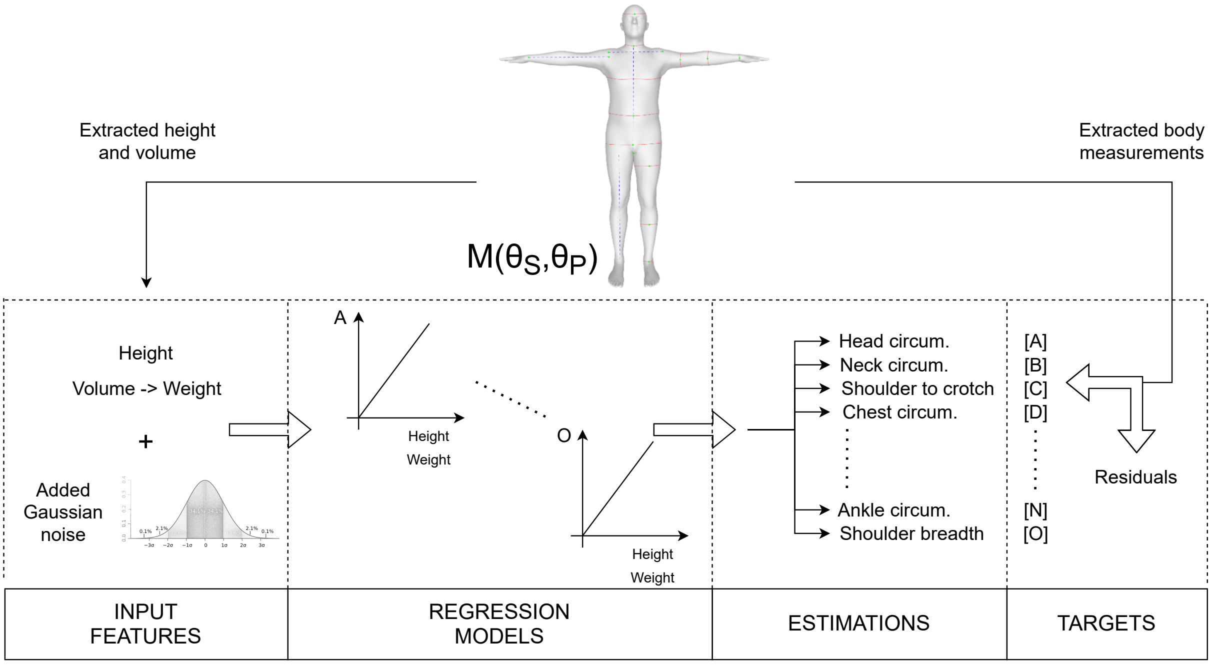


Figure 3.1: An overview of the linear regression model for a single sample from the statistical model. The sample mesh is defined using the shape and pose parameters, $M(\theta_S, \theta_P)_i$. Both input and output are extracted from the sample template mesh, M_i. The input consists of height and weight. Weight can be either available in the dataset or estimated using the calculated mesh volume. The output consists of the 15 body measurements (A–O), which are listed in Table 3.1. To account for the errors in self-reporting height and weight, we model height and weight as stochastic variables by adding Gaussian noise to the input. A linear regression model is fitted to each of the 15 body measurements. The image is adapted from [176].

The linear regression has four assumptions: linearity, homoscedasticity, independence, and normality. The linearity assumes that the relationship between X and the mean Y is linear. The homoscedasticity assumes that the variance of residual is the same for any value of X. Independence assumes that the observations are independent of each other. Finally, the normality assumes that for any fixed value of X, Y is normally distributed. The latter three assumptions are related to residuals, and we verify those in Section 3.4.1. For further details on linear regression, we refer readers to the relevant literature [214].

Note that we also experiment with the interaction terms (h is height, w is weight): $\frac{w}{h^2}$ (BMI), wh, w^2, h^2, and add them to the input vector x^T. The model with no interaction terms is called the *baseline*, and the models with added interaction terms are called the *augmented baselines*. The augmented baselines are marked in the remainder of the chapter as Baseline (I = N), where N is the number of interaction terms. The interaction terms I = 2 correspond to $\frac{w}{h^2}$ and wh, and I = 4 corresponds to $\frac{w}{h^2}$, wh, w^2, and h^2.

In case that the weight measurement is not available in the dataset, we estimate it using the extracted volume. The volume is extracted using a standard algorithm [215]. Then, the body weight (w) is estimated based on the human body density, which is approximately $\rho = 1 \pm 0.005$ kg/L [21, 216], and the extracted volume (V). To account for the variation in body density w.r.t. weight, we model the volume as a normal stochastic variable:

$$V = V_{extracted} + \mathcal{N}(\mu_V = 0, \sigma_V = 5)\ [L]. \tag{3.4}$$

Note that the standard deviation of 5 L applied to the extracted volume propagates to the standard deviation of 5 kg applied to the estimated weight. Additionally, to account for the variation in self-estimation of height and weight, we model self-estimation using another two stochastic variables, $h = h_{extracted} + \mathcal{N}(\mu_h\ =\ 0,\, \sigma_h = 1)$ cm, and $w = V \cdot \rho + \mathcal{N}(\mu_w = 0, \sigma_w = 1.5)$ kg.

3.2.1 Extraction of Body Measurements

We use a total of 18 body measurements, 15 of which are a set of measurements that have been used consistently in previous studies [21, 177, 206, 207, 208, 209, 210, 211, 212, 213], and 3 of which are used specifically to compare with Virtual Caliper [21] (see Table 3.1). The measurements, which are calculated using their associated landmarks, are either lengths or circumferences. We slice the mesh with a horizontal or vertical plane at the designated landmark location to extract the circumferences, such as the waist or thigh circumference, and then add the resulting line segments [177, 217]. Table 3.2 displays the whole list of landmarks together with their appropriate SMPL vertex index. We determine the Euclidean distances between the two relevant landmarks in order to extract the lengths, such as the arm length and shoulder

Table 3.1: The list of 15 (+3) body measurements. The 15 measurements (A–O) are used to compare to the state-of-the-art. The arm length (J) and the three additional measurements are specifically used to compare with [21].

Measurement Set		Measurement	Landmark Index
Standard	A	Head circumference	14
	B	Neck circumference	10
	C	Shoulder to crotch	1, 10
	D	Chest circumference	4
	E	Waist circumference	13
	F	Hip circumference	19
	G	Wrist circumference	9
	H	Bicep circumference	20
	I	Forearm circumference	15
	J	Arm length	2, 9
	K	Inside leg length	11, 12
	L	Thigh circumference	16
	M	Calf circumference	17
	N	Ankle circumference	18
	O	Shoulder breadth	2, 3
Additional [21]	-	Arm span	7, 8
	-	Inseam height	2, 19
	-	Hip width	5, 6

breadth.

3.3 Evaluation

On the BODY-fit and the ANSUR datasets, we assess the linear baseline. More particular, weight estimates derived from the collected mesh volumes are added to the BODY-fit dataset. The extended dataset is known as BODY-fit+W. We compare the baseline on BODY-fit+W with the aforementioned cutting-edge image-based techniques: SMPLify [142], ExPose [191], and Yan et al. [177]. Note that because we do not have the original images of CAESAR, we use the reported results of other works instead [217] datasets. This work references a total of six datasets, which are mentioned in Table 3.3.

Table 3.2: The list of 20 landmarks and their corresponding SMPL vertex indices.

Landmark Index	Landmark Name	Vertex Index
1	Inseam point	3149
2	Left shoulder	3011
3	Right shoulder	6470
4	Left chest	1423
5	Left hip	1229
6	Right hip	4949
7	Left mid finger	2445
8	Right mid finger	5906
9	Left wrist	2241
10	Shoulder top	3068
11	Low left hip	3134
12	Left ankle	3334
13	Lower belly point	1769
14	Forehead point	335
15	Right forearm point	5084
16	Right thigh point	4971
17	Right calf point	4589
18	Right ankle point	6723
19	Mid hip point	3145
20	Right bicep point	6281

Table 3.3: The list of datasets referenced in this work. Note that the baseline is evaluated on the first three datasets.

Dataset	Samples	Data Type	Availability	Approach	Reported by
BODY-fit	4149	SMPL mesh	Public	2D-based	[177], [176]
BODY-fit+W	4149	SMPL mesh	Public	2D-based	[142, 177, 191], [176]
ANSUR	6068	Tabular	Public	Regression	ISO [22], [176]
CAESAR	3800	Point cloud	Proprietary	3D-based	[206, 208, 210, 211, 212] [18, 188, 205, 207, 218]
NOMO3D	375	Point cloud	Public	3D-based	[217]
Virtual Caliper	20	Point cloud	Private	Regression	[21]

3.3.1 Datasets

There are 1474 male and 2675 female SMPL meshes in the BODY-fit collection. The original 3D scans of people, which are not made available to the public, are used to fit the template meshes. Additionally, weights that are computed from related mesh volumes are included in the BODY-fit+W dataset. Figure 3.2 displays the distributions of male body measurements on BODY-fit+W that were derived by measuring the template meshes in accordance with Section 3.2.1. Male and female ranges in height from 145 to 196 cm and 135 to 190 cm, respectively. Male ranges in weight from 40 to 130 kg, while female ranges are between 30 and 130 kg. They also fluctuate proportionally to other metrics' absolute values. Ankle circumferences, for instance, are often less than waist circumferences, etc. According to Figures 3.2 and 3.3, the body measures on BODY-fit+W are just as diverse as the body measurements from the ANSUR dataset. The BODY-fit+W dataset represents the population of SMPL meshes fitted to 3D scans, whereas the ANSUR dataset represents the true human body population. It should be noted that ANSUR dataset does not contain all of the body measurements used in the BODY-fit+W dataset. The ANSUR attributes and expressions used to obtain the corresponding measurements from BODY-fit+W are listed in Table 3.4.

Table 3.4: The specification of ANSUR attributes and expressions corresponding to the body measurements extracted from the SMPL meshes.

	SMPL Mesh (BODY-fit+W)	**ANSUR Attribute/Expression**
A	Head circumference	headcircumference
B	Neck circumference	neckcircumference
C	Shoulder to crotch	sittingheight - (stature - acromialheight)
D	Chest circumference	chestcircumference
E	Waist circumference	waistcircumference
F	Hip circumference	buttockcircumference
G	Wrist circumference	wristcircumference
H	Bicep circumference	bicepcircumferenceflexed
I	Forearm circumference	forearmcircumferenceflexed
J	Arm length	acromialheight - wristheight
K	Inside leg length	crotchheight - lateralmalleolusheight
L	Thigh circumference	thighcircumference
M	Calf circumference	calfcircumference
N	Ankle circumference	anklecircumference
O	Shoulder breadth	biacromialbreadth

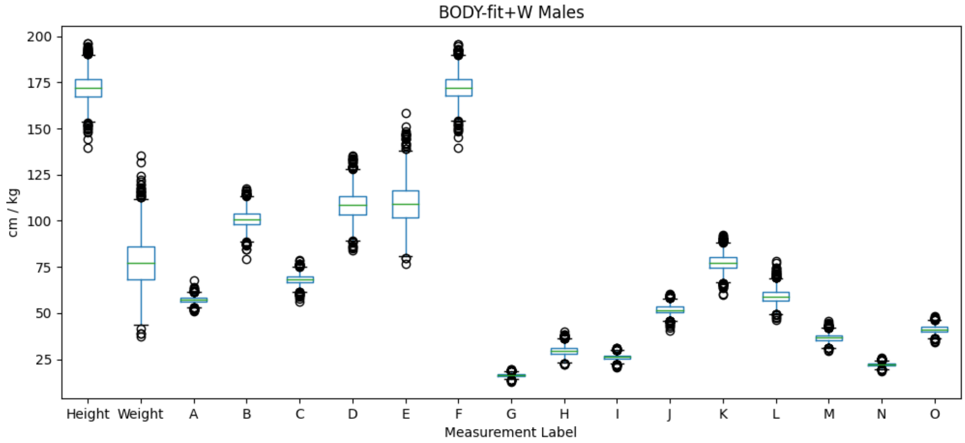


Figure 3.2: The distribution of body measurements for male subjects in the BODY-fit+W dataset. Adapted from [176].

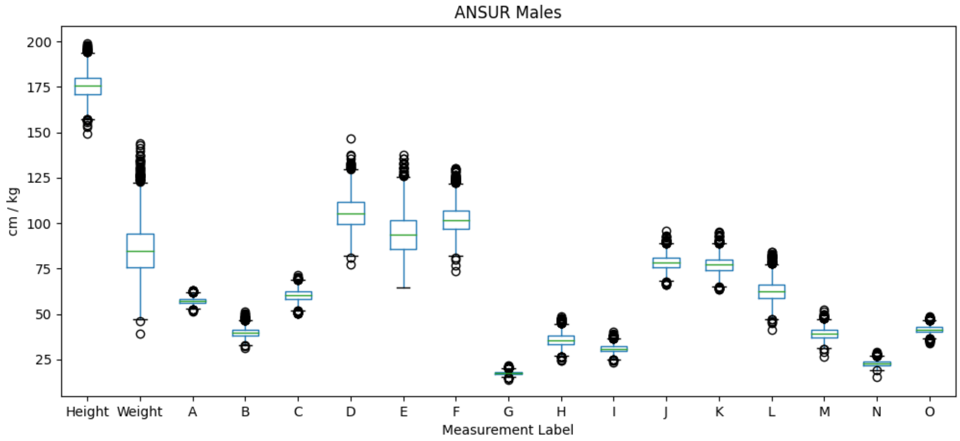


Figure 3.3: The distribution of body measurements for male subjects in the ANSUR dataset. Adapted from [176].

3.3.2 Quantitative Evaluation

The methods are compared quantitatively against the 15 standard body measurements and the three additional measurements to compare the baseline with the Virtual Caliper [21]. The metrics used for comparison are as follows:

- Mean absolute error (MAE), $E_{j,MAE} = \frac{1}{N}\sum_{i}^{N} y_{est,j}(i) - y_{gt,j}(i)$, where i is the sample index, j represents the measurement, and N is the number of samples;
- Mean relative error (MRE), $E_{j,MRE} = \frac{1}{N}\sum_{i}^{N} \frac{y_{est,j}(i) - y_{gt,j}(i)}{y_{gt,j}(i)}$, where i is the sample index, j represents measurement, and N is the number of samples;
- Expert ratio (%<Expert), $\%{<}\text{Expert}_j = \frac{\#{<}Expert_j}{N}$, where j is the measurement and N is the number of samples. This metric shows the ratio of samples that are within the expert errors [10, 11, 22]. The expert errors are shown in Tables 3.5 and 3.6 (*Expert error* rows).

We contrast our linear models with the competing techniques in a number of settings:

- Against the methods that use ground-truth features as input, such as ground truth silhouettes [206, 207, 208, 210, 211, 212]. In this case, we evaluate the baseline using the ground truth volume from the original BODY-fit data, i.e., the volume, height, and weight are modeled as deterministic variables.

- Against the state-of-the-art methods that use estimated or extracted features, including both 3D-based [188, 217, 218] and 2D-based [142, 143, 177, 191, 205, 206, 207, 208, 210, 211, 212] methods. The volume, height, and weight are modeled as stochastic variables (see Section 5.2).
- Against other methods such as the Virtual Caliper [21] that estimates body lengths using a VR headset.
- More detailed comparison with the representative 2D-based [142] and 3D-based methods [217]. On top of MAE, we also report the mean relative error (MRE) and the percentage of the samples within the expert errors (%<Expert).

Ground Truth Methods. We quantitatively compare the baseline with approaches that incorporate data from the real world. The ground truth silhouettes were used to publish the findings of a number of earlier silhouette-based body measurement estimate methods [206, 207, 208, 210, 211, 212]. We enter the volume and the BODY-fit information to the linear model to assess the baseline. By doing this, in addition to using the actual data, we can estimate the weight. Table 3.5 presents the results. Notably, a number of approaches [206, 207, 208] operate inside the bounds of professional mistake. All of these techniques, including the baselines, operate similarly and achieve body measuring errors that are much less than 1 cm for all body measurements. For practical anthropometric applications, however, adopting body volume or silhouettes based on ground truth is unfeasible. Following the earlier efforts, we offer these assessments for the sake of thoroughness.

Table 3.5: Ground truth silhouettes are used in quantitative comparison to image-based body measurement techniques. We also demonstrate the effectiveness of the linear baseline in this instance utilizing the deterministic variables volume, height, and weight (unrealistic). Because ANSUR lacks volume measurements, we simply show the performance of the BODY-fit model in our demonstration. Four more interaction terms are used in the baseline (I=4), as explained in Section 5.2. The results of evaluating methods indicated with a *dagger* on various, non-public data are presented in [206] (MAEs, in mm).

Measurement	Dataset	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Mean
†Xi '07 [212]	CAESAR	50.0	59.0	119	36.0	55.0	23.0	56.0	146	182	109	19.0	35.0	33.0	61.0	24.0	67.1
†Chen '10 [211]	CAESAR	23.0	27.0	52.0	18.0	37.0	15.0	24.0	59.0	76.0	53.0	9.0	19.0	16.0	28.0	12.0	31.2
†Boisvert '13 [210]	CAESAR	10.0	11.0	4.0	10.0	22.0	11.0	9.0	17.0	16.0	15.0	6.0	9.0	6.0	14.0	6.0	11.1
Expert error [22]	ANSUR	5.0	6.0	15.0	12.0	12.0	-	-	-	6.0	-	4.0	-	-	-	8.0	8.5
†Dibra '17 [208]	CAESAR	3.2	1.9	4.2	5.6	7.1	6.9	1.6	2.6	2.2	2.3	4.3	5.1	2.7	**1.4**	2.1	3.6
†Dibra '16 [207]	CAESAR	**2.0**	2.0	3.0	**2.0**	7.0	4.0	2.0	2.0	**1.0**	3.0	9.0	6.0	3.0	2.0	2.0	3.3
†Smith '19 [206]	CAESAR	5.1	3.0	**1.5**	4.7	**4.8**	**3.0**	2.5	2.7	1.9	**1.7**	**1.5**	**2.4**	**2.3**	2.1	**1.9**	**2.7**
Baseline (I = 4)	BODY-fit	7.9	**1.2**	5.6	10.1	9.2	3.6	**0.6**	**1.4**	1.3	5.3	8.0	8.4	2.6	**1.4**	6.6	4.9

State-of-the-Art Methods. The performance of cutting-edge body measurement estimation techniques is shown in Table 3.6, in comparison to our baselines fitted on the BODY-fit+W and ANSUR datasets. The volume, height, and weight are modeled as stochastic variables in the baseline fitted on the BODY-fit+W dataset. We scale their meshes to match ground truth height for comparison with SMPLify, and ExPose since otherwise, their mesh estimations would be considerably compromised by height estimation errors. Our baseline outperforms numerous well-known deep learning algorithms, including HMR [205], SMPLify, and ExPose, in comparison to the competing methods. Note that the baseline obtains MAEs for the neck circumference (B), shoulder-to-crotch (C), and forearm circumference that are within the expert errors (I). The baseline evaluated on the ANSUR dataset performs competitively while being on average less accurate. In Section 3.4.3, a more thorough comparison between BODY-fit+W and ANSUR is provided.

Table 3.6: Comparative analysis of body measurement and shape estimate methods based on images (MAEs in mm). Self-reported height and weight are used as stochastic variables in this instance to demonstrate the effectiveness of the linear baseline (more realistic). The baseline performance is presented using the ANSUR dataset and the BODY-fit+W dataset. The baseline used I = 2 interaction terms, as described in Section 5.2. The best results are shown in bold.

Method	**Dataset**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**Mean**
HMR [205]	CAESAR	16.7	35.7	33.8	92.8	118	68.7	12.2	29.3	20.6	29.9	44.3	38.5	25.8	14.0	26.5	39.8
ExPose [191]	BODY-fit	17.4	13.1	31.4	96.0	116.7	54.8	7.7	33.3	15.3	12.3	29.5	37.3	18.2	8.9	23.0	34.3
SMPLify [142]	BODY-fit	15.3	7.7	8.7	57.5	74.7	39.7	5.1	21.0	9.5	5.7	11.4	27.2	12.3	6.5	10.4	21.6
Hasler '09 [18]	CAESAR	7.5	17.0	7.5	13.0	19.0	16.2	-	-	-	10.4	-	-	-	6.6	-	12.2
Anthroscan [218]	CAESAR	7.4	21.1	7.5	12.4	17.0	7.5	-	-	-	11.7	-	-	-	7.6	-	11.5
Tsoli '14 [188]	CAESAR	5.9	15.8	5.5	12.7	18.6	12.4	-	-	-	10.1	-	-	-	6.2	-	10.9
Yan '20 [177]	BODY-fit	12.0	13.6	8.9	22.2	16.9	14.2	4.8	10.0	8.0	6.8	7.5	13.8	9.1	5.9	8.2	10.8
Dibra '16 [207]	CAESAR	9.3	10.0	6.6	22.8	24.0	20.0	9.9	12.0	7.9	6.4	8.9	15.5	13.2	7.6	6.0	10.7
Expert Error [22]	ANSUR	**5.0**	6.0	15.0	**12.0**	**12.0**	-	-	-	6.0	-	4.0	-	-	-	8.0	8.5
Yan '20 [217]	NOMO3D	-	**3.7**	-	13.2	12.4	**8.9**	4.5	**5.5**	**3.0**	13.2	-	**7.9**	**3.0**	10.6	12.4	8.2
Smith '19 [206]	CAESAR	6.7	8.0	**5.1**	12.5	15.8	9.3	9.3	8.1	5.7	**5.1**	**6.8**	8.8	7.2	5.0	**4.5**	**7.9**
Baseline (I = 2)	BODY-fit+W	9.1	4.2	6.6	30.3	39.5	28.0	**2.7**	10.0	4.9	5.7	9.5	16.0	7.3	**3.3**	9.0	12.4
Baseline (I = 2)	ANSUR	11.9	10.7	17.4	29.1	37.9	21.6	4.4	13.2	9.3	17.6	19.6	17.0	12.8	8.7	11.4	16.2

Other Methods. We also make comparisons to the Virtual Caliper [21], which suggests measuring bodies with a VR headset. The Virtual Caliper performs a more realistic evaluation of its performance when compared to ours by comparing measurements taken by their expert to those taken on real subjects. They estimate height but also utilize self-reported weight as input. As stated in the supplementary section of Table 3.1, we estimate three additional body measurements, including arm span, inseam height, and hip width, for fair comparison. Table 3.7 contains the results. The measurements made with the Virtual Caliper fall short of the baseline. The proposed baseline has the important benefit of not requiring a VR headset.

More Detailed Comparison. We calculate MAE, MRE, and %Expert metrics to compare the representative 2D- and 3D-based methods to the baseline. The representative 2D-based method is SMPLify [142], even though it does not achieve the best performance among the 2D-based methods (see Table 3.6). However, the best performing method, by Smith et al. [206], does not provide the source code to evaluate on BODY-fit+W. As shown in Table 3.8, the 3D-based method by Yan et al. [217], evaluated on the NOMO3D dataset, achieves the best overall performance, with the exception of wrist (G) and ankle circumference (N). For all body dimensions, the baseline matched to ANSUR achieves MREs < 5%, which is appropriate and practical for anthropometric applications. It's interesting to note that the ratio of samples with expert mistakes exceeding 50% is found in the shoulder-to-crotch distance (C) and wrist circumference measurements (G). In general, most ratios for the majority of body dimensions are higher than 25%.

Table 3.7: Comparison to the Virtual Caliper [21] (MAEs in mm) w.r.t. four body measurements—arm length (J) and the three additional measurements (arm span, inseam height, and hip width). We present the baseline evaluated on the same data as in Table 3.6 (BODY-fit+W). Better results are shown in bold.

Measurement	**Dataset**	**Arm Span**	**Arm Length**	**Inseam Height**	**Hip Width**	**Mean**
Virtual Caliper [21]	Virtual Caliper	17.2	7.6	24.6	**6.5**	14.0
Baseline (I = 2)	BODY-fit+W	**13.1**	**5.7**	**8.8**	6.7	**8.6**

Table 3.8: Detailed comparison between 2D-based methods (SMPLify [142]), 3D-based methods (Yan et al. [217]), and the two linear baselines (with I = 2 interaction terms), one fitted to the BODY-fit+W dataset, and one fitted to the ANSUR dataset [22]. Note that to fairly compare with Yan et al., the expert error values are extended according to [217]. The best results in each row for MAEs and %<Experts are shown in bold.

	2D-Based			**3D-Based**			**Baseline (I = 2)**			**Baseline (I = 2)**		
	SMPLify [142] (BODY-fit+W)			**Yan et al. [217] (NOMO3D)**			**BODY-fit+W**			**ANSUR**		
	MAE [mm] ↓	**MRE (%) ↓**	**%<Expert ↑**	**MAE**	**MRE**	**%<Expert**	**MAE**	**MRE**	**%<Expert**	**MAE**	**MRE**	**%<Expert**
A	15.3	2.3	25.1	-	-	-	**9.1**	1.5	**43.1**	11.9	2.1	27.9
B	7.7	4.4	50.7	**3.7**	-	87.6	4.2	1.1	74.9	10.7	2.9	34.9
C	8.7	1.4	85.2	-	-	-	**6.6**	0.9	**94.2**	17.4	3.0	50.1
D	57.5	5.5	13.6	**13.2**	-	**67.6**	30.3	1.4	23.8	29.1	2.9	27.3
E	74.7	7.0	9.3	**12.4**	-	**58.7**	39.5	1.6	19.8	37.9	4.2	20.6
F	39.7	5.9	12.3	**8.9**	-	**72.4**	28.0	1.1	23.4	21.6	2.1	35.8
G	5.1	3.3	59.5	4.5	-	66.5	**2.7**	0.7	**87.0**	4.4	2.7	63.2
H	21.0	7.5	17.2	**5.5**	-	**65.8**	10.0	1.4	33.8	13.2	4.0	28.5
I	9.5	3.8	40.0	**3.0**	-	**74.2**	4.9	0.9	63.9	9.3	3.2	40.1
J	**5.7**	1.6	-	13.2	-	-	**5.7**	1.2	-	17.6	2.3	-
K	11.4	1.7	21.0	-	-	-	**9.5**	1.4	**26.8**	19.6	2.6	13.9
L	27.2	4.5	14.5	**7.9**	-	**47.5**	16.0	1.7	23.4	17.0	2.7	25.1
M	12.3	3.4	27.5	**3.0**	-	**82.5**	7.3	1.0	40.7	12.8	3.4	25.5
N	6.5	2.9	41.4	10.6	-	26.7	**3.3**	0.8	**60.5**	8.7	8.7	28.0
O	10.4	3.2	49.2	12.4	-	-	**9.0**	1.8	**56.2**	11.4	2.9	43.1

3.4 Discussion

Strong performance on the public datasets is shown by the baseline that is presented. This section includes an analysis of the residual hypotheses, emphasizing p-values, and R2 scores (Section 3.4.1); a detailed discussion of using height and weight for body measurement estimation (Section 3.4.2); a detailed comparison of the BODY-fit+W and ANSUR datasets (Section 3.4.3); and a discussion of previous image-based mesh regression techniques (Section 3.5).

3.4.1 Residuals, p-Values, and R^2 Scores

In this section, we assess the homoscedasticity, independence, and normality assumptions of linear regression with reference to residuals, as well as the p-values and $R2$ scores for the regression models of each body measurement. The residuals for the BODY-fit+W and ANSUR models on train and test splits are shown in Figure 3.4. The homoscedasticity requirement is satisfied since the variance of the residuals is typically constant for all values of both models. Since the values are dispersed fairly randomly, the independence assumption is met. Finally, the residuals' means are zero and they are regularly distributed, satisfying the requirement for normalcy, as shown in the right sides of the two figures.

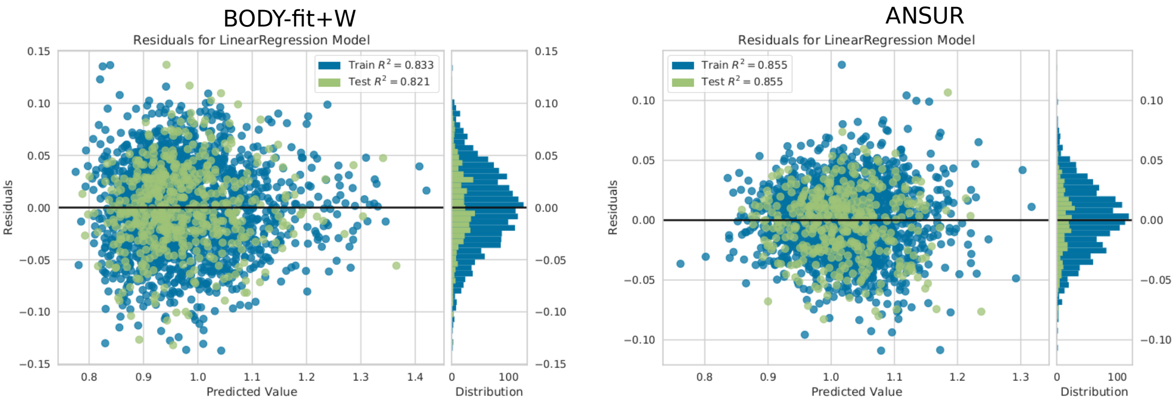


Figure 3.4: An analysis of the residuals for hip circumference (F), for BODY-fit+W and ANSUR, on train and test splits. The image is adapted from [176].

Tables 3.9 and 3.10 show the p-values, R^2 scores, MAEs, and RMSEs for male and female models, on BODY-fit+W and ANSUR, respectively, with two interaction terms (I = 2). We can observe that the vast majority of p-values are within the <0.05 threshold. For the simplicity of the analyses, we keep all the input variables that might lead to increased variance in the predictions and hence larger RMSEs [219]. Ideally, the R^2 scores should be as high as possible. Most of the scores for BODY-fit+W datasets are above or close to 0.8, except for head circumference (A) and shoulder breadth (O). It is reasonable that the head circumference is more difficult to estimate based only on height and weight and their derivative terms. Note that based on p_{BMI} of the shoulder breadth, it would make sense to fit the model without the BMI input term, which may improve the R^2 score. The ANSUR model has somewhat lower R^2 scores, particularly for shoulder-to-crotch (C), wrist circumference (G), and ankle circumference (N). In Table 3.4, it is noted that the shoulder-to-crotch measure was calculated from three manual measurements, which may have contributed to the lower score. It is more challenging to estimate wrist and ankle circumferences because, intuitively, they only slightly match a person's height and weight. The most significant finding of these studies is that, despite the linear models' potential for improvement, their current performance is competitive with that of the most advanced 2D- and 3D-based approaches.

Table 3.9: The linear regression statistics for the BODY-fit+W dataset, for males and females (I = 2). The p-values correspond to the intercept (b), height (h), weight (w), BMI ($\frac{w}{h^2}$), and wh, respectively. In addition, we report adjusted R^2 scores, MAEs (mm), and RMSEs (mm).

	MALE								**FEMALE**							
	p_b	p_h	p_w	p_{BMI}	p_{wh}	**Adj. R^2**	**MAE**	**RMSE**	p_b	p_h	p_w	p_{BMI}	p_{wh}	**Adj. R^2**	**MAE**	**RMSE**
A	0.004	0.000	0.000	0.000	0.000	0.508	8.90	12.36	0.000	0.000	0.000	0.000	0.000	0.442	8.31	12.67
B	0.000	0.255	0.000	0.033	0.000	0.796	4.00	5.04	0.000	0.000	0.000	0.000	0.000	0.795	4.12	5.22
C	0.351	0.000	0.000	0.000	0.000	0.892	6.95	8.77	0.001	0.000	0.000	0.000	0.000	0.903	6.16	7.73
D	0.402	0.000	0.000	0.067	0.000	0.805	26.68	33.81	0.053	0.000	0.000	0.000	0.000	0.808	31.81	40.29
E	0.094	0.000	0.000	0.000	0.000	0.811	38.55	49.18	0.000	0.000	0.000	0.000	0.000	0.819	38.27	48.95
F	0.904	0.000	0.000	0.001	0.000	0.829	22.25	28.50	0.000	0.000	0.000	0.000	0.000	0.833	30.77	39.45
G	0.264	0.000	0.000	0.001	0.000	0.845	2.74	3.46	0.52	0.000	0.000	0.000	0.000	0.85	2.44	3.14
H	0.718	0.005	0.000	0.055	0.000	0.811	8.46	10.76	0.000	0.000	0.000	0.000	0.000	0.825	10.47	13.40
I	0.019	0.000	0.000	0.354	0.000	0.841	4.44	5.71	0.000	0.000	0.000	0.000	0.000	0.848	4.99	6.44
J	0.489	0.000	0.421	0.281	0.359	0.930	5.81	7.81	0.000	0.000	0.055	0.648	0.055	0.923	6.07	8.19
K	0.600	0.000	0.000	0.128	0.001	0.903	10.10	13.26	0.000	0.000	0.007	0.009	0.111	0.920	8.91	11.51
L	0.580	0.000	0.000	0.021	0.002	0.742	14.01	18.70	0.846	0.000	0.000	0.000	0.000	0.790	8.91	21.57
M	0.011	0.000	0.000	0.113	0.000	0.810	7.10	9.29	0.012	0.000	0.000	0.000	0.000	0.835	6.69	8.70
N	0.257	0.000	0.000	0.000	0.000	0.856	2.76	3.49	0.925	0.000	0.000	0.000	0.000	0.848	3.17	4.11
O	0.000	0.034	0.000	0.980	0.015	0.679	8.54	10.78	0.000	0.000	0.000	0.000	0.000	0.689	8.45	10.81

Table 3.10: The linear regression statistics for the ANSUR dataset, for males and females (I = 2). The p-values correspond to the intercept (b), height (h), weight (w), BMI ($\frac{w}{h^2}$), and wh, respectively. In addition, we report adjusted R^2 scores, MAEs (mm), and RMSEs (mm).

	MALE								**FEMALE**							
	p_b	p_h	p_w	p_{BMI}	p_{wh}	**Adj. R^2**	**MAE**	**RMSE**	p_b	p_h	p_w	p_{BMI}	p_{wh}	**Adj. R^2**	**MAE**	**RMSE**
A	0.000	0.011	0.116	0.014	0.349	0.266	10.50	13.25	0.000	0.370	0.654	0.002	0.642	0.152	13.19	17.02
B	0.000	0.863	0.000	0.030	0.001	0.671	12.01	15.09	0.002	0.374	0.000	0.902	0.028	0.606	9.45	12.14
C	0.262	0.000	0.185	0.312	0.402	0.484	16.90	21.70	0.635	0.000	0.234	0.779	0.263	0.405	17.90	21.96
D	0.000	0.137	0.000	0.000	0.000	0.866	24.46	31.34	0.049	0.632	0.000	0.400	0.001	0.740	33.96	43.56
E	0.000	0.123	0.000	0.080	0.000	0.848	36.50	45.31	0.344	0.804	0.000	0.440	0.000	0.777	39.37	49.78
F	0.000	0.574	0.000	0.000	0.001	0.887	20.79	26.68	0.000	0.892	0.000	0.000	0.026	0.855	22.34	28.38
G	0.140	0.000	0.000	0.548	0.000	0.547	4.90	6.15	0.001	0.020	0.300	0.015	0.826	0.543	3.89	4.90
H	0.000	0.000	0.122	0.000	0.370	0.712	15.26	19.72	0.019	0.625	0.000	0.249	0.001	0.800	11.09	14.11
I	0.000	0.435	0.014	0.000	0.696	0.662	10.65	13.44	0.000	0.779	0.004	0.005	0.182	0.678	8.03	10.15
J	0.592	0.000	0.054	0.209	0.091	0.649	15.59	19.39	0.323	0.000	0.875	0.862	0.830	0.646	14.29	18.05
K	0.017	0.000	0.265	0.440	0.478	0.714	19.60	25.00	0.121	0.000	0.605	0.341	0.644	0.692	19.63	24.11
L	0.000	0.000	0.076	0.000	0.026	0.859	17.12	22.15	0.000	0.000	0.109	0.000	0.402	0.840	16.71	21.59
M	0.000	0.272	0.000	0.000	0.228	0.711	12.28	15.62	0.000	0.063	0.587	0.000	0.381	0.648	13.35	16.62
N	0.000	0.029	0.004	0.000	0.154	0.533	8.07	10.28	0.000	0.494	0.470	0.000	0.210	0.385	9.31	11.61
O	0.072	0.000	0.000	0.808	0.004	0.420	11.26	14.09	0.054	0.000	0.845	0.016	0.883	0.373	11.49	14.50

3.4.2 Height and Weight for Body Measurement Estimation

The substantial correlation between height and weight and the population's body measures in the statistical model is not entirely unexpected. The principal components of the SMPL and SCAPE statistical body models, respectively, are analyzed in previous publications like [109, 112]. They add several standard deviations separately to each principal component of the mean shape. The resulting explained variance for the first 10 components is shown in Figure 3.5, and the variation in shapes are shown in Figure 3.6. It can be observed that the first two principal components explain most of the variance in body shapes (Figure 3.5), which in turn define the extreme shape variations visible in Figure 3.6, particularly in terms of height and weight. For the third and fourth components, there are only slight differences. Body measurements and shape are not considerably influenced by the other components. Whether or not these linear relationships hold for the general population, is currently not easy to verify due to a lack of public data. Still, the statistical models are expected to be made from a diverse set of human bodies; therefore, we consider this linear relationship relevant.

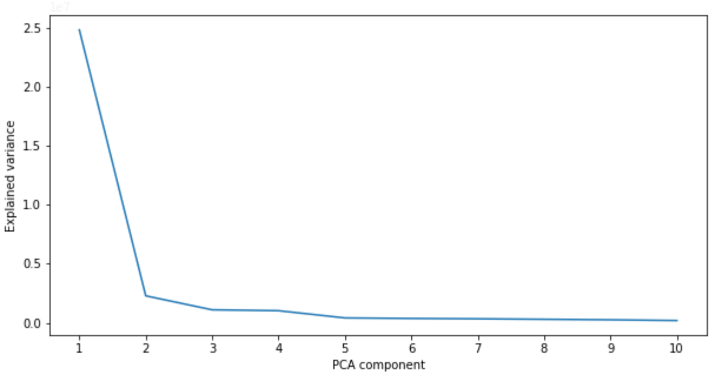


Figure 3.5: Explained variance for the first 10 principal components of the dataset. The graph is generated on CAESAR-fits data [112] by applying PCA to the given vertices. As expected, the first two components, which most significantly correlate to height and weight, explain most of the variance in the data. Adapted from [176].

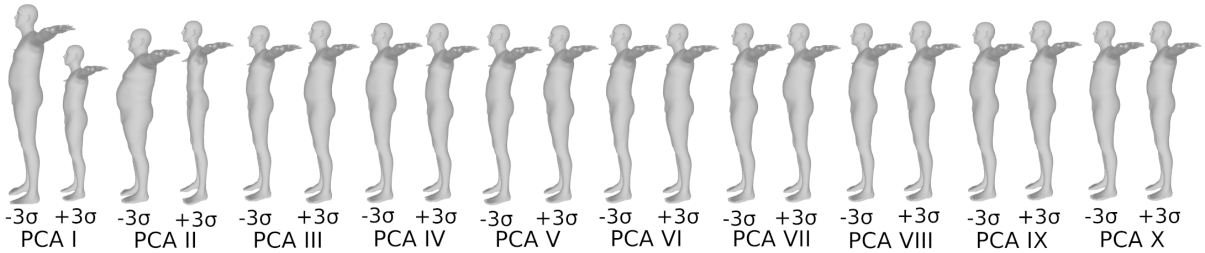


Figure 3.6: Explained variance for the first 10 principal components of the dataset. As expected, the first two components, which highly correlate to height and weight, explain most of the variance in the data. The image is inspired by [112], originally made using SCAPE model [34]. We have drawn the above image using SMPL model [109]. Adapted from [176].

3.4.3 Comparing BODY-Fit and ANSUR Models

The synthetic, statistical population (BODY-fit+W) and the realistic population (ANSUR) are represented by the two datasets that were used to assess the proposed baseline. For some body dimensions, such as neck circumference (B), shoulder-to-crotch (C), arm length (J), inside-leg length (K), and ankle circumference (N), the linear models fitted on the BODY-fit+W and ANSUR datasets perform differently (see Table 3.6).

Even if the volume, height, and weight are represented as stochastic variables, the differences can be explained by the fact that these body measurements are less scattered in the height-weight space, as shown in Figure 3.7. This may indicate that, despite the addition of Gaussian noise to the input, some body measurements of the bodies from the synthetic population still exhibit a linear relationship with height and weight. Another explanation could be that the populations of human bodies in the two datasets varies significantly with respect to these measures. In contrast to the initial group of BODY-fit subjects, which is likely different, ANSUR reflects the population of military members.

At the very least, genuine weight data from BODY-fit participants is needed to confirm these claims. Unfortunately, there is no weight information. Finally, take notice that because there is some inherent error in hand measurements that is difficult to quantify, we simplified the study of ANSUR by ignoring the variation in self-estimation of height and weight.

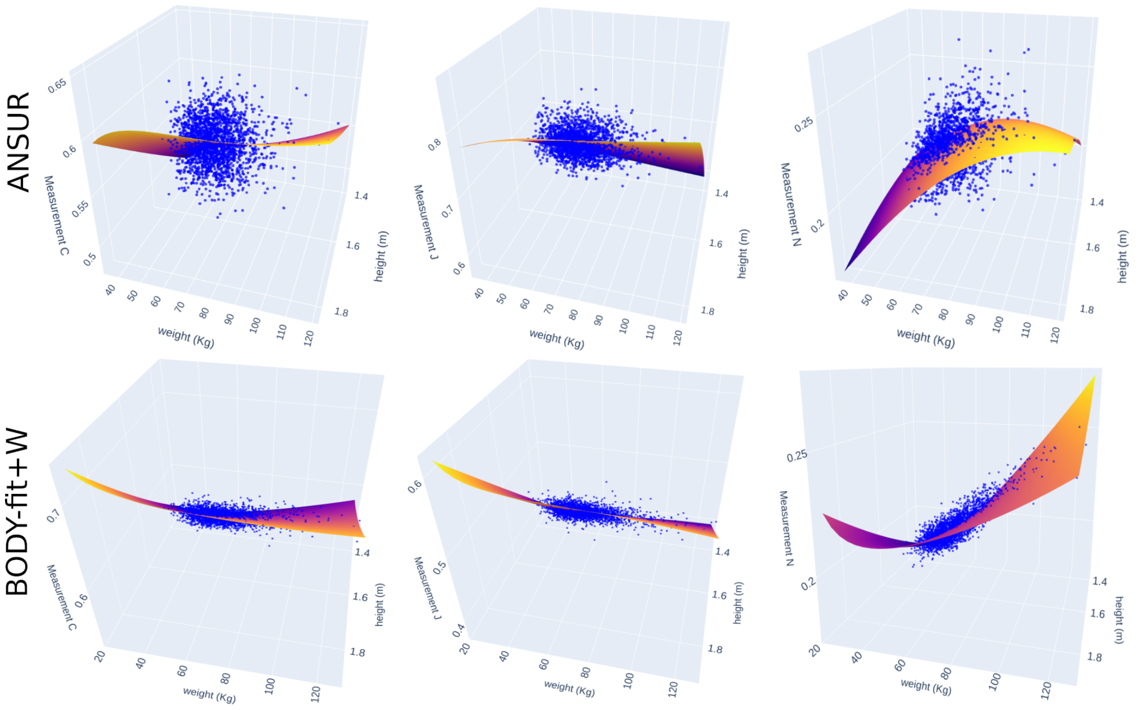


Figure 3.7: The plots above display the three most severe situations of body measurements with various dispersions on the ANSUR (**first row**) and BODY-fit+W (**second row**) datasets: shoulder-to-crotch (C), arm length (J), and ankle circumference (N). The fitted planes correspond to the linear model with two interaction terms (I = 2). Adapted from [176].

3.5 Final Remarks and Future Work

Assumptions. Note that the normal distributions that are added are uncorrelated with other variables to describe all stochastic variables, which is not realistic for all subgroups of the population. As an illustration, underweight persons frequently overestimate their weight, whereas overweight people frequently report inaccurate weights [220].

Future work. Future techniques should be compared against the linear regression based on self-reported height and weight; in other words, no body measurement estimation method should ever do worse that the proposed baseline. The results presented in this section primarily illustrate the approximative relative performances of the baseline and the competing methods on the publicly accessible benchmarks currently in use, but for future reference and datasets, we advise fitting the model to more realistic and diverse data, if applicable. The following stage is to develop more accurate, publicly available benchmarks based on statistical models in order to further analyze the correlation between height and weight as well as other anthropometric parameters.

Chapter 4

Learning Body Pose Estimation from Images

Human pose estimation is a vision task of detecting the keypoints that represent a standard set of human joints. In this chapter, we focus on 3D human pose estimation from multiple views in a single time frame.

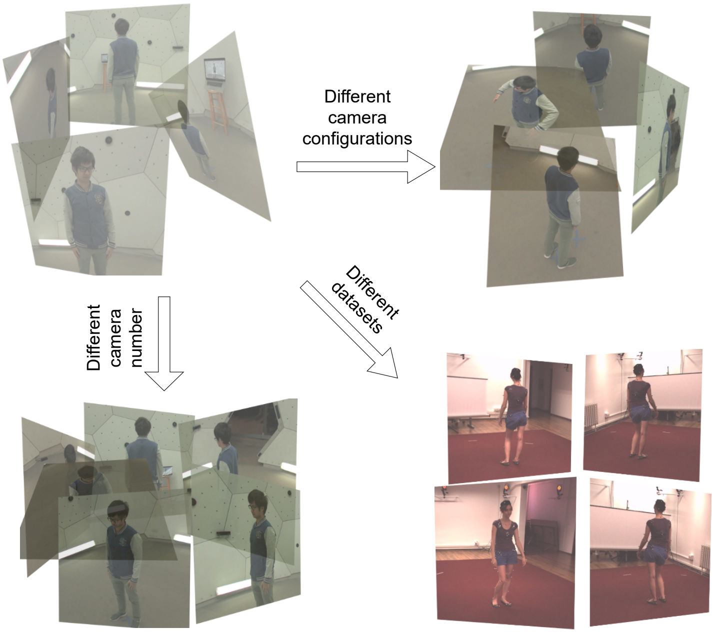


Figure 4.1: We propose a stochastic framework for human pose triangulation from multiple views and demonstrate its successful generalization across different camera arrangements, their number, and different public datasets. The upper two and the lower left image shows different camera arrangements and their number on CMU Panoptic Studio dataset [221]. The lower right part shows the Human3.6M's 4-camera arrangement [35]. The image is adapted from [135].

The typical method for multi-view pose estimation is to: (1) use a pretrained pose detector to find the correspondent 2D keypoints in each view [139, 140, 222], and then (2) triangulate [134, 136, 137, 223, 224, 225]. A crude method applies triangulation from all accessible perspectives to 2D detections as-is. Some views contain incorrect detections as a result of the variety of positions and self-occlusions; these views should be disregarded or their impact should be mitigated throughout the triangulation process. Applying RANSAC [226] and marking the keypoints with reprojection errors that are higher than a predetermined threshold as outliers [136, 227] is one technique to ignore the incorrect detections. End-to-end learning is not possible with vanilla RANSAC since the gradients cannot be back-propagated because it is non-differentiable. Modern techniques for 3D posture estimation collect 2D picture features, such as heatmaps, from several viewpoints and combine them for 3D elevation in an end-to-end process [134, 137, 223]. These techniques are referred to as *learnable triangulation approaches* [134, 137, 223].

The learnable triangulation systems are frequently restricted to a single camera layout and their number due to a generally fixed set of cameras throughout training. The demonstrated performance on novel views is noticeably worse than using the original (base) views, despite several works' attempts to generalize outside the training data [134, 136, 138, 223, 224, 228, 229].

We propose a stochastic framework for human pose triangulation from multiple views and demonstrate its successful generalization across different camera arrangements, their number, and different public datasets [135]. In order to obtain 3D poses from multiple views, we assume the following:

- the images taken from multiple views are synchronized;
- only a single person is visible in the scene;
- the person is visible from at least two views at each given point in time.

4.1 Related Work

We distinguish two types of related work. First, we focus on triangulation-based 3D pose estimation methods and methods that attempt to generalize between the different camera arrangements and datasets. Second, we relate to keypoint correspondence methods and point out how our problem differs from the standard correspondence problem.

Triangulation. Most of the single-person image-based approaches either use robust triangulation (RANSAC) or apply learnable triangulation. Several methods [227, 230, 231] based on robust triangulation use RANSAC on many (more than four) views to apply triangulation only on inlier detection candidates to produce pseudo ground truth data. He et al. [136] exploit epipolar constraints to find the keypoint matches between multiple images and then apply robust triangulation.

The standard approach for learnable triangulation using deep learning models [134, 137, 225, 232, 233] is to first extract 2D pose heatmaps, where each heatmap represents the probability of a keypoint location. Cross-view fusion [137] builds upon the pictorial structures model [234] to combine 2D keypoint features from multiple views to estimate a 3D pose. An algebraic triangulation [134] estimates the confidence for each keypoint detection and applies weighted triangulation. Their volumetric approach combines the multi-view features and builds the volumetric grid, obtaining the current state-of-the-art for single-frame 3D pose. Finally, [223] fuses the features into a unified latent representation that is less memory intensive than the volumetric grids. Similar to us, they also attempt to disentangle from the specific spatial camera arrangement.

Keypoint correspondence. The standard keypoint-based computer vision approaches, such as structure-from-motion [50], rely on sparse keypoint detections to establish initial 3D geometry. The core problem is to determine the correspondences between the extracted keypoint detections across images, under various illumination changes, texture-less surfaces, and repetitive structures [48, 122]. The usual approach is to apply keypoint descriptor such as SIFT [52] and find inlier correspondences using RANSAC [226]. Even though this paradigm is successful in practice, it is not differentiable and, therefore, cannot be used in an end-to-end learning fashion.

Several works have proposed soft and differentiable versions of RANSAC (DSAC) [235, 236, 237, 238]. The successful soft RANSAC alternative [238] learns to extract both local features of each data point, as well as retain the global information of the 3D scene. Similar to us, they also demonstrate convincing generalization capabilities to unseen 3D scenes. On the other hand, DSAC and its variants [235, 236, 237] propose a probabilistic learning scheme, i.e. minimizing the error expectation. We follow their approach but also discover that different strategies work better for our problem (see Sec. 4.3).

In contrast to the standard keypoint matching approaches, we extract keypoints with already known human joint correspondences between the views. However, our correspondent keypoints are noisy, oscillating around the centers of the joints, which potentially leads to erroneous triangulation. Our model demonstrates robustness to erroneous keypoint detections.

4.2 3D Pose Estimation Method

We propose a generalizable triangulation of human posture, which is motivated by stochastic learning [239] and its applications in computer vision [235, 236, 237]. We start by creating a pool of randomly chosen hypotheses. A 3D pose known as a hypothesis is one in which the points are created by triangulating a random subset of views for each joint individually. Each hypothesis that is created is evaluated by a neural network. The loss function is an expectation of the triangulation error, i.e. $\mathbb{E}(h_i) = \sum_i e_i s_i$, where e_i is the error of the hypothesis h_i and s_i is the hypothesis score. In order to learn the distribution of hypotheses, the model minimizes the

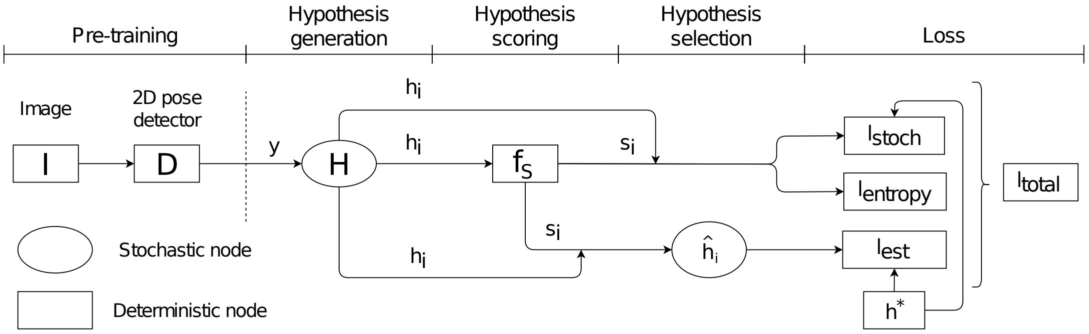


Figure 4.2: An overview of our method. Before stochastic learning, 2D keypoints, y, are extracted. In each frame, the hypothesis pool, $h_i \in \mathbf{H}$, is generated, and the poses are passed through the scoring network, f_S. The hypothesis $\hat{h}_i$ is selected based on the estimated scores s_i. Finally, the total loss, l_{total}, consists of three components (l_{stoch}, $l_{entropy}$, l_{est}), and is calculated with respect to the ground truth, h^*. The image is adapted from [135].

error expectation. The essential concept is to develop the ability to assess 3D pose hypotheses without taking into account the spatial camera configuration utilized for triangulation.

We initially provide a broad overview of the stochastic framework before focusing on how it applies to generalizable pose triangulation. The framework consists of multiple phases, which are depicted in Fig. 4.2:

1. **Pre-training.** The 2D poses (keypoints) for each image in the collection are extracted before stochastic learning. We employ the keypoints that were extracted using a baseline model [139] that was pretrained on the Human3.6M dataset in all of our tests. Therefore, the only input for the stochastic model is keypoint detections, or $\mathbf{y}$. Keypoints of the form JxK, where J is the number of joints and K is the number of views, are found in each frame.
2. **Hypothesis generation, H**. Only a portion of randomly generated hypotheses are produced because it is feasible to produce an incredibly vast number of them. We model the step of hypothesis creation as a stochastic node, following [239] and [235].
3. **Hypothesis scoring, $\mathbf{f_S}$**. Each generated hypothesis $h_i \in \mathbf{H}$ is scored using a scoring function, $f_S(h_i|\mathbf{y}) = s_i$. A multi-layer perceptron (MLP, neural network) serves as the scoring mechanism. The MLP is the only learnable part of our model. The estimated scores s_i, passed through the Gumbel-Softmax, $\sigma_{GS}(s_i)$ (Eq. 4.3), represent the estimated probability distribution of the hypotheses $\mathbf{H}$, $\theta_{\mathbf{H}}$.
4. **Hypothesis selection, $\hat{h}_i$**. We experiment with several hypothesis selection strategies. The one that works the best for us is the weighted average of all hypotheses:

$$\hat{h}_{weight} = \sum_i s_i h_i, \quad \sum_i s_i = 1, \quad h_i \in \mathbf{H}, \tag{4.1}$$

where the scores s_i are used as weights. We also try other strategies, such as the stochastic selection:

$$\hat{h}_{stoch} = h_i, \quad \text{with} \quad i \sim \theta_{\mathbf{H}}, \tag{4.2}$$

where hypothesis h_i is selected based on the estimated distribution $\theta_{\mathbf{H}}$. As shown in Sec. 4.3, the stochastic selection performs worse than the weighted, in contrast to [235].

5. **Loss calculation,** l_{total}. The loss function consists of several components:
   (a) Stochastic loss. We calculate our stochastic loss as an expectation of error for all hypotheses in accordance with [235], $l_{stoch} = \mathbb{E}(e_{\mathbf{H}}) = \sum_i e(h_i, h^*) s_i$, where e_i is the error of the estimated hypothesis with respect to the ground truth, h^*, and s_i represent the probability that the error is minimal.
   (b) Entropy loss. Score estimations s_i tend to quickly converge to zero. To stabilize the estimation values, we follow [236] and minimize an entropy function, $l_{entropy} = -\sum_i s_i \log(s_i)$.
   (c) Estimation loss. We define it as the error of the selected hypothesis with respect to the ground 3D pose, $l_{est} = e_i(\hat{h}_i, h^*)$. The estimation loss, in the case of generalizable pose triangulation, is most similar to the standard 3D pose estimation loss, used by the competing approaches [134, 137, 223, 224, 225].

   Finally, the total loss is a sum of the three components, $l_{total} = \alpha\, l_{stoch} + \beta\, l_{entropy} + \gamma l_{est}$, where α, β, and γ are fixed hyperparameters that regulate relative values between the components.

The predicted scores s_i must have their values normalized into the $[0,1]$ range in order for them to accurately represent the probabilities. Applying the softmax function is the conventional method for normalizing the output values, $\sigma(s_i) = \frac{\exp s_i}{\sum_j \exp s_j}$. To avoid early convergence, we use the Gumbel-Softmax function [240, 241]:

$$\sigma_{GS}(s_i) = \frac{\exp((\log s_i + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log s_j + g_j)/\tau)}, \tag{4.3}$$

where τ is a temperature parameter, and g_i represent samples drawn from *Gumbel*(0, 1) [242] distribution. The range of the distribution is controlled by the temperature τ. Lower-score hypotheses have less of an impact than higher-score hypotheses at lower temperatures ($\tau < 1$), and vice versa. *Gumbel*(0, 1) is used to introduce noise to each sample while keeping the original distribution or distributions intact. This enables the model to be more adaptable in terms of the choice of hypothesis.

Following is the description of the stochastic framework used specifically for learning human pose triangulation.

Pose generation. The 3D human pose hypothesis, $h_i \in \mathbf{H}$, is produced by the subsequent technique. For each joint k, a subset of views, $\mathbf{v}_k$, is selected at random. The detections from the selected views are triangulated to produce a 3D joint.

Pose normalization. The input to the pose scoring network, $f_{S,pose}$, are 3D pose coordinates, $\mathbf{p}$, normalized in the following way — we select three points: left and right shoulder and the pelvis (between the hips) calculate the rotation between the normal of the plane given by the three points, and the normal of the xy-plane, and apply that rotation to all coordinates. In addition to the 16 body part lengths that are provided by all nearby joints, such as the left lower arm, left upper arm, left shoulder, etc., we additionally extract the 3D posture coordinates. We next combine the body part lengths and normalized 3D pose coordinates into a 1D vector and feed it into the network. The result is a scalar, s_i, which represents the hypothesis's score, h_i.

Pose estimation error. The pose estimation error, $e_i(\hat{h}_i, h^*)$, is a mean per-joint precision error (MPJPE) [35] between the estimated 3D pose, $\hat{\mathbf{p}}_i$, and the ground truth, $\mathbf{p}^*$:

$$e_i(\hat{h}_i, h^*) = e_i(\hat{\mathbf{p}}_i, \mathbf{p}^*) = \frac{1}{J}\sum_{k}^{J} ||\hat{p}_{ik} - p_k^*||_2, \tag{4.4}$$

where p_{ik} is the k-th keypoint of the i-th pose.

4.3 Evaluation and Discussion

The Panoptic Studio [221] and Human3.6M [35] datasets are used to assess the proposed stochastic framework. We use Human3.6M for the quantitative comparison to state-of-the-art since the majority of the prior 3D pose estimation methods displayed their results on this benchmark. A significant number of cameras (31) with helpful data annotations are present in Panoptic Studio (camera parameters, 3D and 2D poses). To assess the effectiveness of generalization across various camera configurations and their number, we make use of the Panoptic Studio dataset. Additionally, we assess the generalization between the datasets from Panoptic Studio and Human3.6M. We employ Panoptic Studio sequences with a single person in the scene, as described in [243], because studies are based on a one-person position estimation.

4.3.1 Generalization Performance

The proposed model's ability to generalize effectively to diverse spatial arrangements, number of views, and datasets is one of its most significant features. This overcomes a key drawback of the previous models. We used five various camera configurations to assess the generalization performance across data sets:

1. Cameras $1,2,3,4,6,7,10$ (CMU1),
2. Cameras $12,16,18,19,22,23,30$ (CMU2),
3. Cameras $10,12,16,18$ (CMU3),
4. Cameras $6,7,10,12,16,18,19,22,23,30$ (CMU4), and
5. Cameras $0,1,2,3$ (H36M).

Table 4.1: Five data sets with varying geographical camera placements, number of views, and datasets were used to demonstrate the generalization performance (MPJPE in mm) (CMU Panoptic Studio and Human3.6M). The performance on five test sets for the given train set is displayed in each row. The final column displays the largest difference between the scores for specific test sets. While previous rows display intra-dataset performance, the final row demonstrates inter-dataset generalization performance.

Train	CMU1		CMU2		CMU3		CMU4		H36M		Max diff. ↓
Test	CMU1	25.8	CMU1	25.8	CMU1	25.6	CMU1	25.2	CMU1	25.6	2.3%
	CMU2	25.4	CMU2	26.0	CMU2	25.5	CMU2	25.6	CMU2	25.9	2.4%
	CMU3	24.9	CMU3	26.0	CMU3	25.0	CMU3	25.0	CMU3	25.7	4.4%
	CMU4	25.1	CMU4	25.6	CMU4	25.3	CMU4	25.1	CMU4	25.5	**2.0%**
	H36M	**33.5**	H36M	33.4	**H36M**	**31.0**	H36M	32.5	**H36M**	**29.1**	15.1%

Table 4.2: The evaluation of generalization performance from CMU Panoptic Studio [221] to Human3.6M dataset [35], compared to the volumetric approach of Iskakov et al. [134]. The proposed approach achieves 8.8% better performance on H3.6M compared to [134], when trained on a 4-camera CMU3 dataset (see Table 4.1).

CMU → H3.6M		
Ours [135]	Iskakov et al. [134]	Improvement
31.0 mm	34.0 mm	**8.8%**

Following training on each of the five camera configurations, the effectiveness of generalization is evaluated using the other four configurations. The spatial camera layout and the number of the five selected sets vary. Additionally, the transfer learning skills between the datasets are tested using the fifth camera set (H36M).

Our Generalization Performance. Regardless of the chosen training dataset, Table 4.1 shows consistent performance on each of the five test datasets. The Panoptic Studio dataset in particular shows performance variation between test sets to be within 5%, demonstrating tolerance to alternative camera setups and their number (intra-dataset). Additionally successful is the inter-dataset generalization, which we assess against state-of-the-art approaches [134, 223]. Keep in mind that the generalization that has been presented can be used for inference as well as training.

Table 4.3: The comparison to RANSAC, algebraic triangulation [134], and VoxelPose [138] on Panoptic Studio (intra-dataset) [mm]. The numbers show the performance on novel camera views. Our number is obtained as an average over 12 non-diagonal values of Table 4.1.

Intra-dataset (CMU Panoptic Studio)			
RANSAC	Algebraic	VoxelPose	Ours
39.5	33.4	25.5	**25.4**

Volumetric Triangulation. Table 4.2 compares our proposed method to the state-of-the-art

Table 4.4: The evaluation of generalization performance compared to Remelli et al. [223] (lower is better). We measure the performance drop between the base test set and the novel test set for intra-dataset and inter-dataset configurations. Note that we do not compare on the same datasets, so we only measure the relative drop in percentages. Still, our approach demonstrates a significantly smaller performance drop compared to the competing method in all setups. The † presents the canonical fusion, and the ‡ presents the baseline approach in [223].

Intra-dataset			
Method (train dataset)	Base test	Novel test	Diff. ↓
Remelli et al. [223] (TC1)†	27.5 mm	38.2 mm	38.9%
Remelli et al. [223] (TC1)‡	39.3 mm	48.2 mm	22.6%
Ours (CMU1)	24.9 mm	25.8 mm	3.6%
Ours (CMU3)	25.0 mm	25.6 mm	2.4%
Ours (CMU4)	25.0 mm	25.6 mm	2.4%
Ours (CMU2)	25.6 mm	26.0 mm	**1.6%**
Inter-dataset			
Method (train dataset)	H36M	CMU1	Diff. ↓
Ours (H36M)	29.1 mm	33.5 mm	**15.1%**

3D pose estimation approach [134]. Iskakov et al. reported an average 34.0 mm error on Human3.6M test set when they trained on CMU Panoptic Studio (4-camera arrangement). Compared to them, we achieve 31.0 mm on our 4-camera arrangement (CMU3), demonstrating an improvement in inter-dataset generalization (see Table 4.1 for the comprehensive results).

Remelli et al. Our approach is compared with the approach by Remelli et al. [223] in Table 4.4. They explicitly address the generalization to novel views, just like us. They compare the test performances on cameras (1, 3, 5, 7) as a basic arrangement (TC1) and testing it on cameras (2, 4, 6, 8), as a novel arrangement, to show their intra-dataset generalization performance on Total Capture (TC2). Our model is not evaluated using Total Capture. Instead, to contrast with Remelli et al., we assess relative score differences and compare the performance of the CMU camera test sets. For intra-dataset configuration, our model performs consistently with a variety of camera configurations. Additionally, our inter-dataset performance between CMU Panoptic Studio and Human3.6M is 15.1%, which is still superior to the top score of Remelli et al. The inter-dataset experiment is the most challenging because it involves changing the camera configuration.

RANSAC. On Panoptic Studio, we outperform RANSAC by a significant margin. This can be explained by the fact that most cameras do not give the CMU dataset a complete image of a person, resulting in heavy occlusions and missing sections. In contrast to our model, which learns human pose prior (see Sec. 4.3.4), RANSAC is unable to evaluate the predicted 3D

Table 4.5: No additional training data setup. Overall comparison to the state-of-the-art on Human3.6M dataset. The proposed method outperforms most of the state-of-the-art methods. All values are showing MPJPE scores (mm).

Protocol 1, abs.	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg ↓
Tome et al. [225]	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8
Kadkh. et al. [224]	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
Cross-view [137]	28.9	32.5	26.6	28.1	28.3	29.3	28.0	36.8	41.0	30.5	35.6	30.0	28.3	30.0	30.5	31.2
Remelli et al. [223]	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	31.7	31.2	29.9	26.9	33.7	30.4	30.2
Epipolar [136]	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	31.7	28.2	26.4	23.6	28.3	23.5	**26.9**
Volumetric [134]	18.8	20.0	19.3	18.7	20.2	19.3	18.7	22.3	23.3	29.1	21.2	20.3	19.3	21.6	19.8	**20.8**
Ours [135] (h_{weight})	27.5	28.4	29.3	27.5	30.1	28.1	27.9	30.8	32.9	32.5	30.8	29.4	28.5	30.5	30.1	29.1

posture as a whole because it uses only reprojection errors of individual 3D joints as an inlier selection criterion.

Algebraic Triangulation. Originally, the algebraic triangulation [134] was recommended as an enhancement to RANSAC, and it estimates the weight for every joint location. The weight-based model does certainly perform better on Human3.6M and Panoptic Studio than RANSAC does. However, it has a number of shortcomings. Each view is first processed independently, and then each joint is separately triangulated. Because it ignores the entirety of the pose, the weight-based algebraic model shares the same flaw as RANSAC. On the other hand, our model successfully picks up on human position previous, allowing it to choose more realistic poses and making it more resistant to occlusions and missing body parts. You should be aware that algebraic triangulation does not test their weighted model on using camera locations other than the ones in the training set. As a result, Table 4.3 displays the outcome of the model without weights because this model is reliable when used with various camera systems. The actual result of the weighted model might differ, but it is hard to estimate by how much.

VoxelPose. In contrast to our 25.42 mm, VoxelPose [138] claims a 25.51mm MPJPE score on their intra-dataset experiment. We did not pretrain our 2D backbone on the Panoptic Studio dataset, which would have likely further improved our 2D keypoint estimation and, as a result, our final 3D pose predictions, even if we attain equivalent performances.

4.3.2 3D Pose Estimation on Base Dataset

Table 4.5 displays the comparison to the state-of-the-art. Notably, the volumetric triangulation approach [134] is not included since the Table only displays the methods that use Human3.6M for training and testing, with no additional training data. We acquire a 2.2 mm poorer MPJPE than the single-frame method with the highest performance, Epipolar Transformers [136], but we outperform the majority of other recent techniques.

Other than the evaluation of our best result ($\hat{h}_{weight}$), we also compare between different hypotheses:

- Weighted average hypothesis, $\hat{h}_{weight}$,

Table 4.6: Overall quantitative comparison between the hypotheses. The values are showing MPJPE scores in mm (the lower is better).

Hypothesis	Human3.6M ↓	Panoptic Studio ↓
$\hat{h}_{weight}$	29.1	**24.9**
$\hat{h}_{avg}$	31.2 +2.1	25.9 +1.0
$\hat{h}_{most}$	41.3 +12.2	25.0 +0.1
$\hat{h}_{least}$	74.5 +45.4	29.8 +3.9
$\hat{h}_{stoch}$	41.3 +12.2	26.5 +1.6
$\hat{h}_{random}$	45.0 +15.9	26.1 +1.2
h_{best}	22.3 -6.8	24.4 -0.5
h_{worst}	98.9 +69.8	31.0 +6.1
RANSAC	**27.4 -1.7**	39.5 +14.6

- Average hypothesis, $\hat{h}_{avg}$, obtained as an average of all hypotheses,
- Most and least probable hypotheses, $\hat{h}_{most}$ and $\hat{h}_{least}$, the hypotheses with maximal and minimal estimated score, s_{max} and s_{min},
- Stochastic hypothesis, $\hat{h}_{stoch}$, selected randomly, based on the estimated distribution $\theta_{\mathbf{H}}$,
- Random hypothesis, $\hat{h}_{random}$, selected randomly from an uniform distribution,
- Best and worst hypotheses[*], h_{best} and h_{worst}, with the lowest and the highest errors, e_{min} and e_{max}.

Additionally, we also compare ourselves with RANSAC as reported in [134] (see Subsec. 3.3.2).

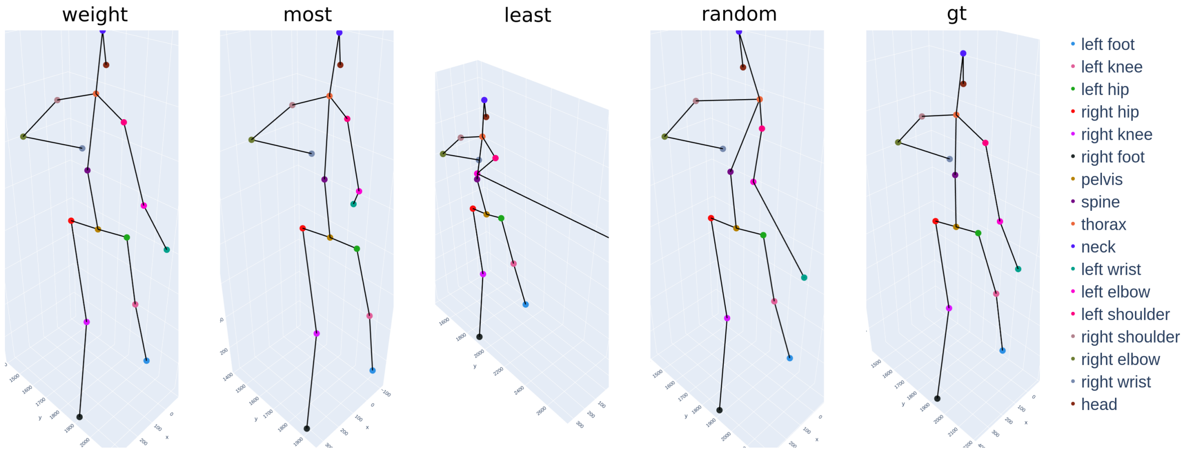


Figure 4.3: Qualitative comparison between four 3D pose hypotheses compared to ground truth (*gt*), on Human3.6M. The image is adapted from [135].

[*] Note that the best and the worst hypotheses are not available in inference (missing ^ sign), because they are determined using ground truth.

As published in [134], Table 4.6 compares the MPJPE scores of all previously described pose hypotheses on the two datasets to the RANSAC result. Although the RANSAC strategy outperforms our weighted average hypothesis, $\hat{h}_{weight}$, on Human3.6M, we demonstrate a considerable improvement on Panoptic Studio. Also keep in mind that RANSAC can compete with the majority of cutting-edge methods on Human3.6M, previously listed in Table 4.5.

Regarding other results, the average hypothesis, $\hat{h}_{avg}$ performs better than the stochastic, $\hat{h}_{stoch}$. The stochastic performs even worse than the random hypothesis on Panoptic Studio. The most probable hypothesis, $\hat{h}_{most}$, outperforms the average on Panoptic Studio. Note that the difference between best and worst hypothesis (h_{best}, h_{worst}) is significantly different on the two datasets. This suggests that the hypotheses generated on Panoptic Studio are more similar to each other and the distribution is less broad. The difference between the most and the least probable hypotheses ($\hat{h}_{most}$, $\hat{h}_{least}$) is reasonable on both datasets, which confirms that our model learned to differentiate between the poses.

4.3.3 Qualitative Results

Fig. 4.3 shows the qualitative performance comparison between several hypotheses. The least probable hypothesis, h_{least}, does not have visually plausible pose reconstruction, while the random hypothesis, $\hat{h}_{random}$, has some obvious errors in the upper body. The most probable hypothesis, $\hat{h}_{most}$ has minor reconstruction errors on the right arm and shoulder. The weighted hypothesis, $\hat{h}_{weight}$, is visually comparable to ground truth.

4.3.4 Human Pose Prior

In this subsection, we give an additional explanation of our pose estimation results in the light of human pose prior, i.e., the information about the human body based on which (part of) the decisions about the 3D pose hypothesis selection are made.

We demonstrate the successful pose prior learning of the pose scoring network, $f_{S,pose}$. There are previous works that attempt to learn human pose prior [234, 244, 245], but they do not quantitatively evaluate their methods. The idea of learning pose prior is to differentiate between the 3D poses that are more plausible and the poses that are less plausible with respect to several human body properties. The properties that can be extracted from the 3D pose are based on the body part lengths and between-joint angles. In this work, we focus on body part lengths, i.e. *left-right body symmetry*.

The body symmetry is measured for six different body left-right part pairs: upper arms, lower arms, shoulders, hips, upper legs, and lower legs. For each pair, l, we calculate the ratio, r_{il} between the left and right part, in each time frame, i. The final pose prior metric is a variance of the ratios over time:

$$S^2 = \frac{\sum_i (r_{il} - \bar{r}_l)^2}{T-1}, \tag{4.5}$$

where $\bar{r}_l$ is the mean ratio for the pair l, and T is the number of frames. The reason for using ratios instead of the differences between the body parts is that some people are naturally asymmetric, so the idea is only to measure the consistency over time.

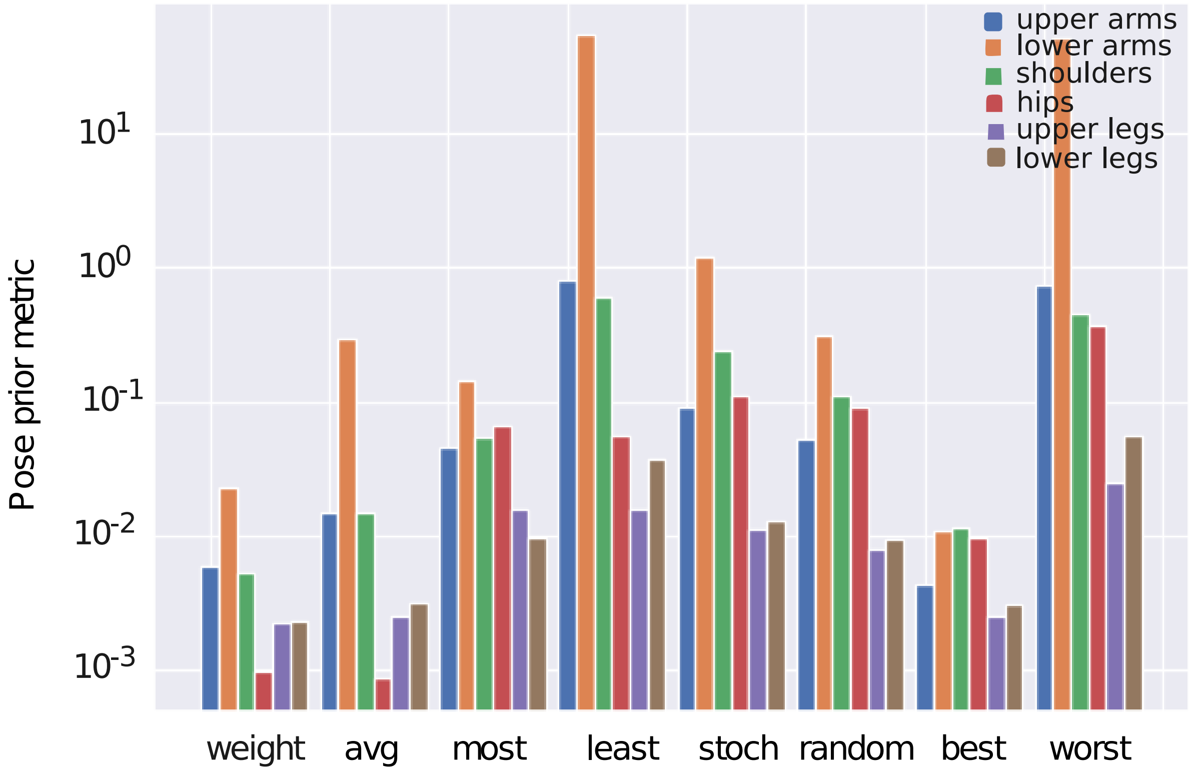


Figure 4.4: Evaluation of the human pose prior metric for different hypotheses, and six body part pairs (lower is better). The image is adapted from [135].

Fig. 4.4 shows the pose prior metrics for the subject 9 of the Human3.6M dataset, for different hypotheses. As expected, the values are generally the lowest for our best performing hypothesis, $\hat{h}_{weight}$, followed by the average hypothesis, $\hat{h}_{avg}$. The difference between the most probable and the least probable hypothesis ($\hat{h}_{most}$, $\hat{h}_{least}$) suggests that we successfully learned body pose prior, i.e. differentiate between the plausible poses with respect to the body symmetry consistency over time. Note that the best hypothesis, h_{best}, is comparable to $\hat{h}_{weight}$.

4.4 Final Remarks and Future Work

The proposed generalizable approach is a promising novel direction for 3D human pose estimation, as well as other related computer vision problems, such as the camera pose estimation. The results show convincing generalization capabilities between camera arrangements and datasets, outperforming previous methods. Using the proposed generalizable triangulation approach, it is now possible to transfer the performance of the base dataset to any novel multi-camera dataset, in inference. The model requires relatively little training data, which makes training faster and more convenient for smaller datasets. The overall performance is competitive with respect to human pose triangulation when evaluated on the same camera arrangement as the one in the training dataset. The next reasonable step is to exploit image features in an end-to-end learning fashion, which should further improve the performance and possibly outperform the state-of-the-art even on the base dataset. The current model supports only a single-person pose triangulation. To extend to multi-person, we need to solve the keypoint correspondence problem between the people.

Chapter 5

Learning Body Shape Estimation from Images

Capturing detailed human appearance and mesh regression are very active research topics. To solve for accurate human shape models, given in-the-wild visual observations, not only shape and pose must be taken into account, but also clothing. Estimating accurate human pose, shape, and clothing from 2D images is challenging due to (self-) occlusions, variety of body poses, shapes, and garments, all on top of the classical challenges such as 2D-to-3D scale ambiguity. Moreover, de-coupling human body from the garment geometry is particularly difficult and, in this dissertation, we focus on jointly estimating the parameters of pose, shape, and clothing style.

Remarkable progress has been made for estimating unclothed humans from images [128, 131, 186, 191, 201, 246, 247], recovering accurate parameters for body shape and pose in challenging examples. Also, several methods have been proposed for the reconstruction of humans in clothing [248, 249, 250, 251]. Although highly detailed on full-body images, these methods are still not robust on the examples with significant occlusions. Only recently, ClothWild has been proposed [252] for estimation of pose, shape, and clothing in challenging examples. They estimate body and pose using fixed, pretrained model [201], while garment displacements are calculated as an offset from the body based on DensePose semantic segmentation predictions [253]. We go a step further by learning to estimate pose, shape, and clothing style parameters from scratch, thus jointly taking all the parameters into account during training.

To learn a fully-parametric model, accurate training data for pose, shape, and clothes is required. Obtaining accurate body shapes is difficult as it requires 3D scanning of each subject, fitting a parametric human body model mesh on top of the scan, and then retargeting the mesh to each image frame of the corresponding subject. An example of such a dataset is 3DPW [254], which provides ground truth shapes for 18 different subjects throughout 62 video sequences. An alternative to obtaining accurate ground truth is using synthetic datasets [114, 125, 126,

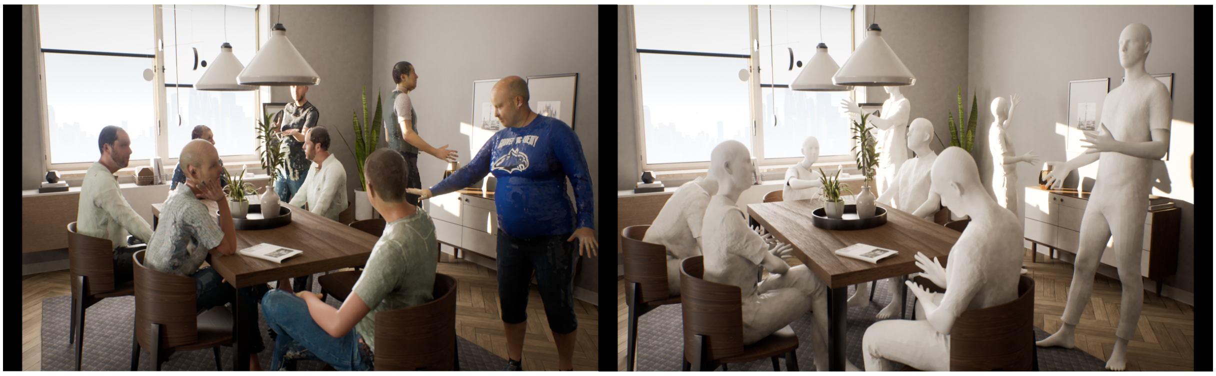


Figure 5.1: (Left). The ClothAGORA dataset consists of controllable characters with garment geometry in realistic 3D scenes. (Right). The proposed method recovers pose, shape, and clothing style parameters, which fully define the mesh geometry. Note that the right part is manually created and not estimated using our method.

127, 255, 256, 257], whose main challenge is to achieve realism in order to close or mitigate the synthetic-to-real gap in inference. SURREAL [125] renders textured SMPL meshes on random 2D background. Even though the textures mimic the clothes, the geometry of the mesh still corresponds to unclothed body. Recently, AGORA has been proposed [127], placing highly realistic commercial human scans into 3D scenes. AGORA contains diverse clothing and identities and has perfectly accurate ground truth. The main disadvantage is that the 3D scans are not controllable, i.e., their pose, shape, and clothes are fixed.

We propose two synthetic datasets populated by controllable characters - ClothSURREAL and ClothAGORA. Both image datasets consist of 2D renders of parametric clothed meshes. The garment displacements for the clothed meshes are estimated using TailorNet [115]. ClothSURREAL images consist of images of meshes [109] with clothing geometry displacements, rendered in front of random 2D backgrounds. ClothAGORA goes beyond by placing controllable clothed characters in 3D scenes (see Fig. 5.1, left, as an example). We propose ClothAGORA as a dataset, as well as the toolbox for creating a possibly infinite number of parametric characters. We also publish the tools to generate novel controllable characters that can then be used for training and evaluation of image-based estimation models. The two proposed datasets are used to train our model and demonstrate a state-of-the-art performance on a public benchmark.

Our model trains on synthetic data and generalizes to in-the-wild examples. This is achieved by learning from derived image features only - edges [170] and 2D keypoints [139, 258, 259], which has proven useful to mitigate synthetic-to-real gap in previous works [128, 131]. The features are passed through ResNet50 and the output are pose, shape, and *clothing style* parameters. The model is supervised using the parameters instead of directly using mesh geometry, which simplifies our architecture and makes our model the fastest among the state-of-the-art (inference time). Using the estimated parameters, the corresponding clothed 3D meshes can then be produced, using SMPL for the body and TailorNet for clothing displacements.

Our model achieves the best performance for estimating garment displacements from the body, i.e., compared to the recent previous work [252], although they do not jointly estimate all the parameters of pose, shape, and clothing style. Our method is the fastest in inference compared to other methods that estimate clothed humans (also the ones that estimate a single mesh for body and clothes), which we demonstrate on 3DPW test set and in-the-wild examples.

In summary:

1. We propose Garmentor, the first model to jointly estimate pose, shape, and garment style parameters from in-the-wild images, trained on image features - the model achieves the state-of-the-art performance for estimating clothed people;
2. We create two synthetic datasets containing controllable clothed characters - ClothSURREAL and ClothAGORA, along with providing the tools for extending and generating novel data.

5.1 Related Work

Recently, several clothed human estimation approaches are proposed. Most of the proposed approaches use SMPL, either as a final result or as a part of the estimation pipeline. We distinguish between the methods that: estimate pose and shape parameters of the SMPL model, recover clothed humans as a single mesh, and recover clothed humans as multiple meshes. Finally, we overview existing synthetic human datasets.

Estimation of minimally-clothed meshes. The most recent human pose and shape estimation methods [191, 192, 247, 260, 261] are able to estimate body meshes in-the-wild, under diverse and loose clothing in-the-wild, along with expressive hands and faces. Even though these results are remarkable and particularly visually plausible, accurate body shape estimation has been tackled in a separate group of methods done by Sengupta et al. [128, 129, 131]. The success of these methods is in part achieved by training on diverse and even extreme set of input poses, shapes, global orientations, and occlusions. Inspired by their strategies, we extend their architecture to support clothing style estimation. Other than the parametric approaches, there are also many successful approaches that estimate 3D vertices directly [107, 185, 186, 246], some of which we compare to in Section 5.4.

Recovering clothed characters as single meshes. To add clothes on top of the body, many methods model clothing geometry as 3D offsets from corresponding body vertices [262, 263, 264, 265, 266, 267, 268]. They can be easily animated, but cannot model more complex garments such as skirts. The implicit methods such as PiFU [249, 250] have, in general, more details in their reconstructions. However, even the most recent methods [248, 251, 269] do not generalize well in-the-wild and usually require full-body images. In Section 5.4, we quantitatively compare with the most successful method in this group, ICON [248], and qualitatively also with PiFU-HD [250].

Using parametric clothing models for estimation. Only one previous approach, ClothWild, [252] uses the parametric clothing model [270] to estimate clothed people from images. The advantage of using the parametric clothing model is that it is specified only by the given, low-dimensional parameters. ClothWild estimates the pose and shape of the underlying body using the pretrained pose-shape model [201] and estimates human sex, visible clothing types, and garment displacements. The key to their success is in the fact that the displacements are determined by DensePose [253] semantic segmentation, which aligns the garment displacement vertices in 3D to the segmentation in 2D. Therefore, they do not directly estimate the style parameters. We propose a more straightforward computational strategy, where the estimation of all the parameters of the clothing model are done jointly and learned from scratch, making the strategy simpler, faster, and more successful on the test.

Synthetic human datasets. SURREAL is a synthetic human dataset that consists of rendered SMPL meshes in front of an images from LSUN dataset [174] that are used as random 2D backgrounds. The texture is applied on top of the meshes to mimic the clothes, but the geometry is unchanged. 3DPeople [160] addresses the problem of clothing geometry by creating a large synthetic dataset of dressed people in motion, using a similar 2D background strategy. Their models, however, are not parametric and, therefore, cannot be easily controllable. AGORA is a recent attempt to create highly realistic 3D synthetic human dataset. It contains several thousand different identities and poses, and it contains particularly diverse clothing. AGORA then fits SMPL-X meshes to each of these 3D scans and thus obtains the corresponding pose and shape parameter ground truth, which is useful both for training and evaluation. The 3D scans of people used by AGORA are commercially available through 3DPeople* [271], AXYZ [272], HumanAlloy [273], and RenderPeople [274] platforms. Our aim is to propose the first step towards tackling this issue by offering a dataset of free parametric clothed people populating the same realistic 3D scenes as the people from AGORA, while not being restricted by particular pose, shape, and clothes parameters. Compared to SURREAL, the proposed ClothSURREAL dataset also contains cloth geometry on top of the SMPL bodies.

5.2 Shape and Clothes Estimation Method

We propose a novel computational strategy for learning to estimate the parameters of human pose, shape, and clothing style from images (θ, β, γ). The main novelty is adding the clothing style parameters, γ_{est}, as output, as well as using the style parameters, γ_{gt}, to generate input data. By using the parameters both to generate input and to supervise output estimations, we have a

*Note that 3DPeople in this context is not the same as 3DPeople approach and dataset [160]

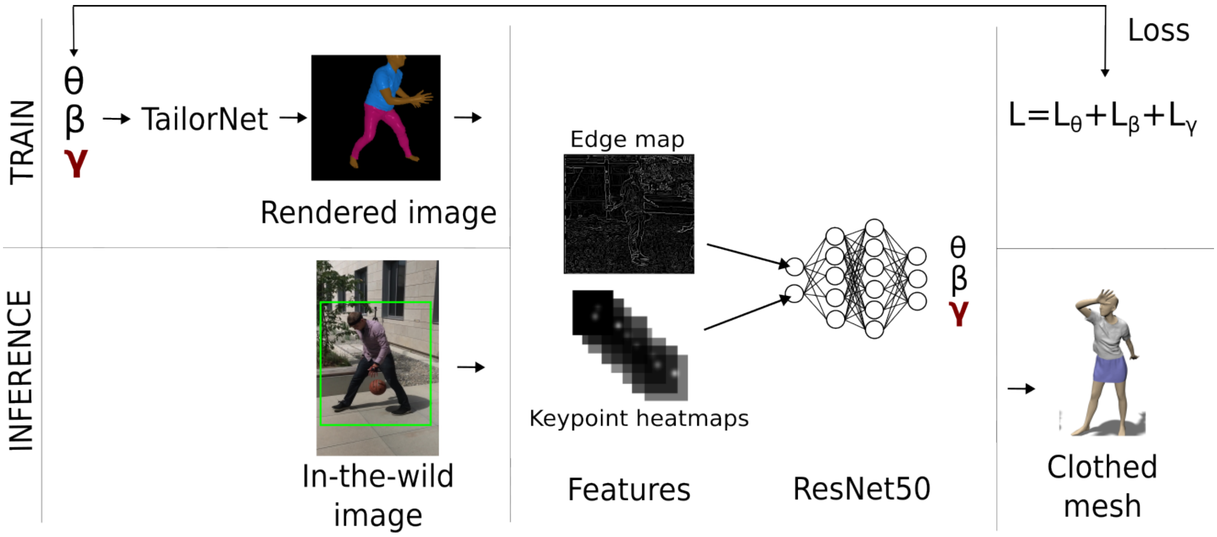


Figure 5.2: An overview of Garmentor. For training, we randomly sample pose, shape, and style parameters, $(\theta_{\mathrm{gt}}, \beta_{\mathrm{gt}}, \gamma_{\mathrm{gt}})$, and pass the parameters to the TailorNet model to generate the clothing mesh. The mesh is rendered and the image is augmented. Based on the image, the edge map and keypoint heatmaps are created and used as an input to ResNet50. The network estimates the pose, shape, and style parameters, $(\theta_{\mathrm{est}}, \beta_{\mathrm{est}}, \gamma_{\mathrm{est}})$. The ground truth set of parameters is used for supervision (see the loss function). In inference, an in-the-wild image is taken as input, and off-the-shelft feature extractors are applied to obtain input features for the network. Again, the parameters are estimated as output and the corresponding clothed mesh can be created for visual evidence.

complete control over the supervision of the estimation model, while being able to generate various pose, shape, and clothing style combinations.

The parameters $(\theta_{gt}, \beta_{gt}, \gamma_{gt})$ are given to TailorNet to estimate garment displacements, $\mathbf{D}(\theta_{gt}, \beta_{gt}, \gamma_{gt})$, while the underlying body, $B(\theta_{gt}, \beta_{gt})$, is calculated using SMPL. The input to ResNet50 [275] are derived image features - 2D heatmaps and an edge map extracted using a canny edge detector [170]. The network produces pose, shape, and style estimates, $(\theta_{est}, \beta_{est}, \gamma_{est})$, which are then supervised using known ground truth, $(\theta_{gt}, \beta_{gt}, \gamma_{gt})$.

In inference, an off-the-shelf 2D keypoint detector [139, 258, 259] and the same canny edge detector is used to obtain input features. The network outputs the parameters, $(\theta_{est}, \beta_{est}, \gamma_{est})$ which are then given to SMPL and TailorNet to produce the corresponding body, $B(\theta_{est}, \beta_{est})$, and garment displacements, $\mathbf{D}(\theta_{est}, \beta_{est}, \gamma_{est})$, which define the final clothed mesh, $M(\theta_{est}, \beta_{est})$. The high-level architecture consists of the following components (see Fig. 5.2 for visual reference): input features, deep learning network, loss components, and a backbone parametric model.

Input Features. The input features consist of 17 keypoint maps (COCO mapping), where each keypoint map is a synthetic heatmap that represents 2D normal distribution with a center in 2D joint location and a standard deviation as a dispersion around the center. The keypoints outside the image or occluded by the body are removed. In addition, an edge map is used instead of RGB image directly to mitigate the gap between the synthetic and real data. The example of input is shown in the upper left part of Fig. 5.2.

Deep Learning Network. The network is ResNet50 [275]. It is trained from scratch using described input features. The output from the network consists of pose, shape, and clothing style parameters, $(\theta_{est}, \beta_{est}, \gamma_{est})$. It is important to note that the style parameters, γ, have one additional dimension compared to vectors of pose and shape, (θ, β), because γ is a matrix consisting of style parameters for each of the four different garment types: T-shirt, shirt, pants, and short pants. Only two garment types are used for each example, such as T-shirt and pants, and the remaining values are ignored. Otherwise, the garment types that are not present in the current sample, such as a long-sleeve shirt or short pants, would unfairly increase the loss.

Loss Components. Our loss function is the following:

$$f_{loss} = L_\theta + L_\beta + L_\gamma + L_{J_{2D}} + L_{J_{3D}} + L_R, \tag{5.1}$$

where the loss components represents pose parameters', shape parameters', clothing style parameters', 2D joints', 3D joints', and global orientation's losses, respectively. Note that we omit the hyperparameter weights next to each of the loss components, which are specified at the end of the Chapter (Table 5.7). The $L_{J_{2D}}$, $L_{J_{3D}}$, and L_R loss components are calculated as mean squared errors (MSEs). The L_θ, L_β, and L_γ are calculated using negative log-likelihood. In particular, pose loss uses the matrix-Fisher distribution [276] over relative 3D joints rotations as described in [128]. The clothing style loss, L_γ, produces four loss sub-components for the corresponding garment types (T-shirt, shirt, short pants, and pants), but masks the ones that are not used in the current sample to avoid backpropagation through these branches.

Parametric Clothing Model. We use TailorNet as our parametric clothing model. Garment displacement vertices D are modeled as a function of the underlying SMPL body vertices, B. For a given pose, shape, and style parameters, (θ, β, γ), the garment vertices deform using the following function:

$$T^G(\theta, \beta, \gamma) = \mathbf{I}(B(\theta, \beta) + D(\theta, \beta, \gamma_c)), \tag{5.2}$$

followed by skinning. The indicator matrix, $\mathbf{I}$, is 1 for the corresponding body and garment vertices, (B_i, D_j). The displacement function, $D(\theta, \beta, \gamma_c)$, is learned for each separate garment type. The problem with this kind of modeling is that there are C models for each of C garment classes. We mitigate this problem by training all of the garment classes at once, as described in this section.

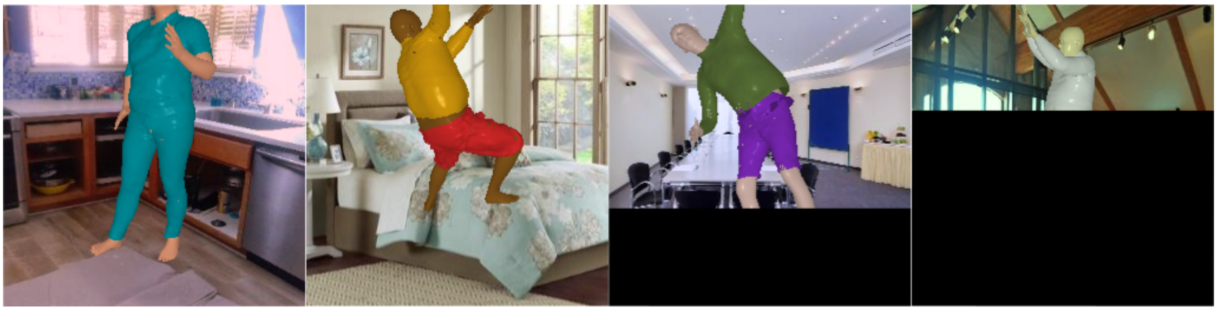


Figure 5.3: A few augmented samples of the augmented ClothSURREAL training data. Based on a small probability, significant parts of the images are removed.

5.3 Datasets

One of the main advantages of synthetic compared to real data is that accurate ground truth information is available and infinitely large datasets can be generated. The crucially useful information for learning accurate pose, shape, and clothes estimation, that are difficult to obtain on large real datasets, are the ground truth parameters of the characters shown in images. Even though there are synthetic datasets like SURREAL, 3DPeople, and AGORA, which contain useful information such as depth, normals, keypoints, silhouettes, and even pose and shape parameters, we propose novel datasets that also contain clothing information (see Sec. 5.3 for the description of ClothSURREAL used for training). Additionally, the characters in our datasets are parameterized which allows to generate infinite combinations of poses, shapes, and styles. We highlight this advantage in Figure 5.4.

5.3.1 ClothSURREAL Dataset

The proposed ClothSURREAL training dataset consists of 120000 male characters and 120000 female characters, having 30000 samples for each garment type (T-shirt, shirt, pants, and short pants). The poses, θ, are randomly sampled from AMASS dataset[†]. The shape and style parameters, (β, γ) are randomly sampled from normal distributions, $\mathcal{N}(0, 1.25)$ and $\mathcal{N}(0, 0.75)$. Using the specified parameters, TailorNet meshes are produced and rendered on an black background. Then, a random 2D background image is added around the rendered clothed body. Note that the background does not add additional semantic meaning to the image, but is instead used as a simple augmentation strategy to avoid overfitting on empty backgrounds.

ClothSURREAL images are additionally augmented. First, the body is randomly moved along (x, y) axes of the image, to mimic random camera movements. Then, we remove a significant part of the body (either upper, lower, right, or left) with a probability of 0.1, following the strategy in [128]. Finally, random background from LSUN dataset is added, following original

[†] Note that the global orientation of the body is contained within the first three out of 72 pose parameters, $R_{glob} = \theta_{(0,1,2)}$, given in angle-axis representation.

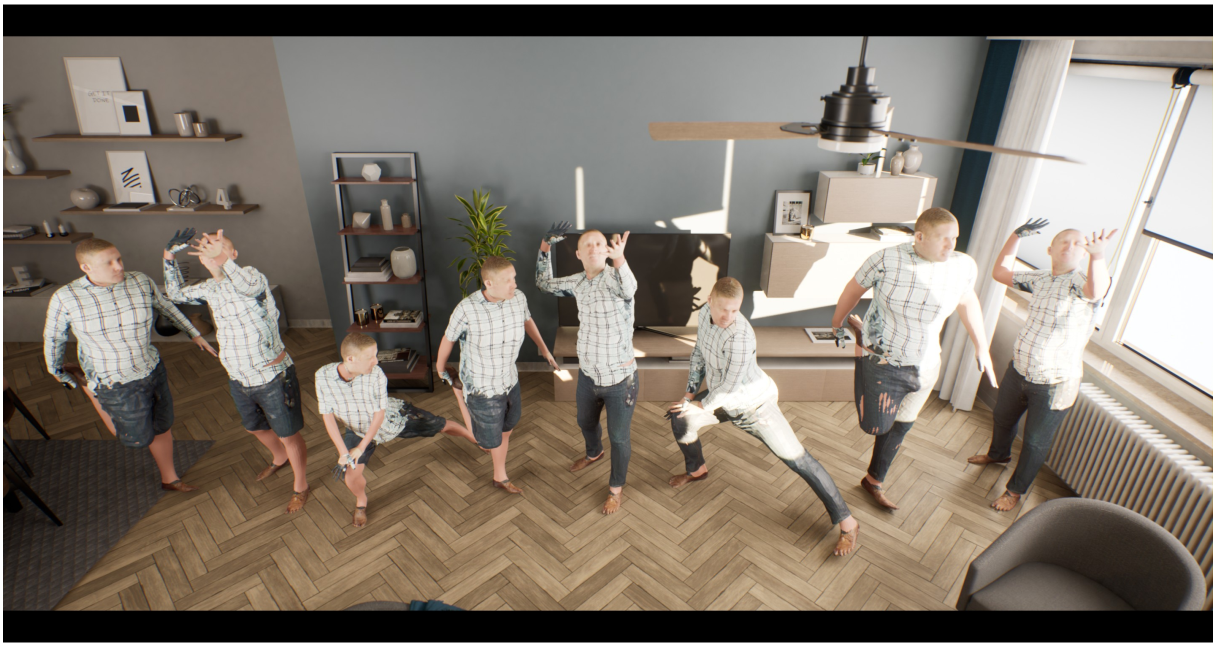


Figure 5.4: A highlight of our ClothAGORA dataset. Our meshes are fully controllable in terms of pose, shape, and clothing style.

SURREAL. It has been shown that these augmentation strategies improve the robustness on challenging in-the-wild examples for the estimation of undressed people [128, 129, 131]. The examples of our augmented training examples are shown in Fig. 5.3.

5.3.2 ClothAGORA Dataset

The ClothAGORA dataset consists of five 3D environments, one indoor and four outdoor, specified on AGORA's downloads website [277]. The environments are commercially available[‡] for use in Unreal Engine [175]. For each of the five environments, we generate 30 different scenes of 8 clothed characters. The pose, shape, and style of the characters are randomly sampled, resulting in 240 different characters for each of the 5 environments, which is 1200 characters in total, i.e., 150 scenes in total. To populate the scenes (assign global translations to each character), we use AGORA's parameters on each of the subjects' locations.

For each of the 150 scenes, we render an image from 10 different viewpoints, which makes 1500 high-quality renders. The renders can be used for the evaluation of our model by cropping each individual character as shown in Fig. 5.6 and resizing the image to (256, 256). The total number of cropped images in our dataset in 12000. For convenience, we provide a summary of the dataset in Table 5.1.

The textures for the clothed meshes are obtained from MGN dataset [278]. The original textures are made for SMPL mesh topology. In order to apply the textures to any given garment type, the area on the body where the skin is revealed, due to short sleeves or pants, is painted

[‡] One of the scenes is free (ArchViz Interior).

Table 5.1: A summary of our ClothAGORA dataset.

# Number	Description
5	3D environments (1 indoor, 4 outdoor)
30	3D scenes with specific characters
10	Camera viewpoints for particular scene
8	Different characters for particular scene
12000	Total number of single-image crops

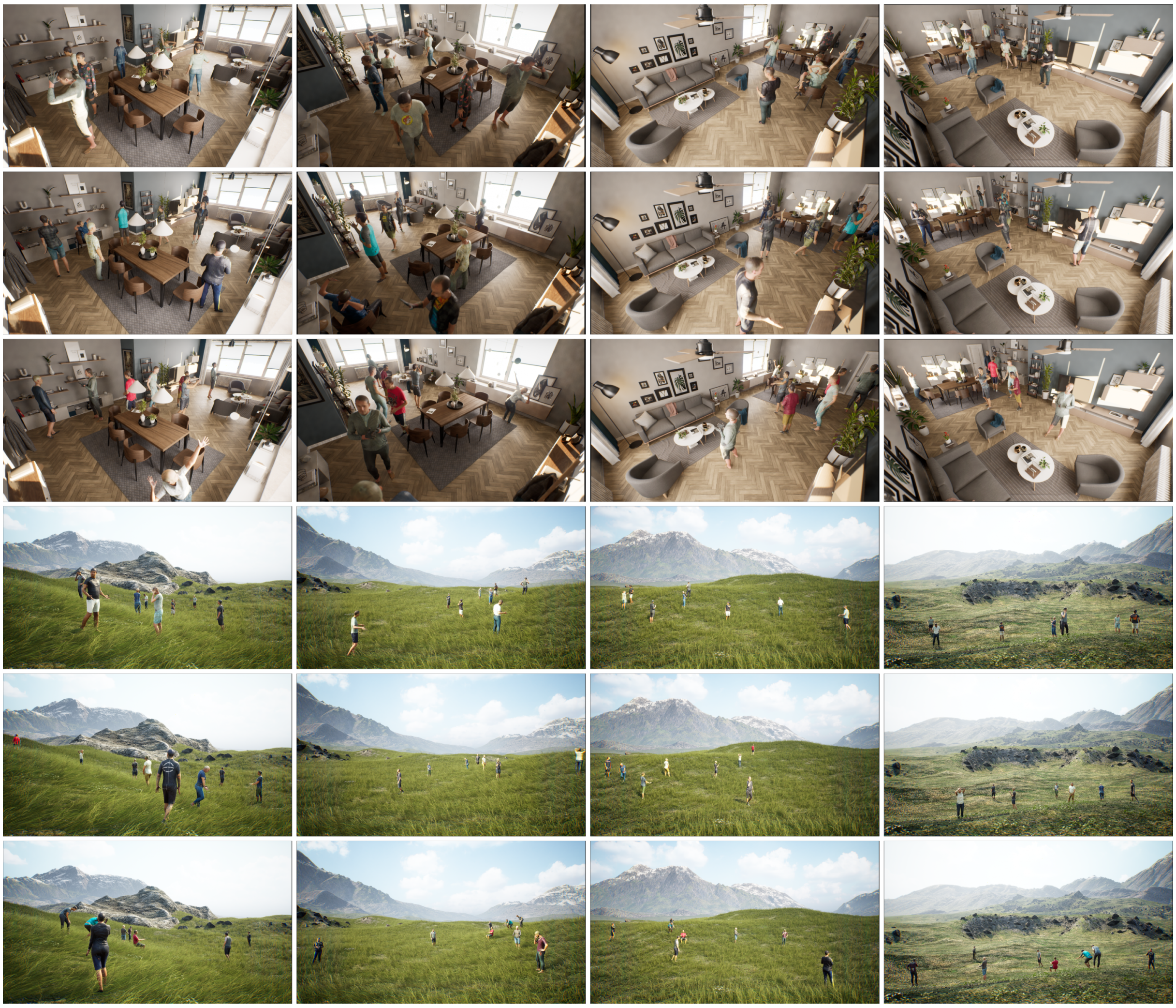


Figure 5.5: The examples of three indoor and outdoor scenes from our ClothAGORA dataset from four different viewpoints.

in skin color. We determine the skin color by sampling pixels from the texture maps in the area around the nose, where the skin colors is usually persistent.

Along with the images, we publish the corresponding parameters and metadata, and the tools and instructions on how to extend the dataset. We believe that releasing such a tool is useful for the community and opens possibilities towards accurate and controllable characters

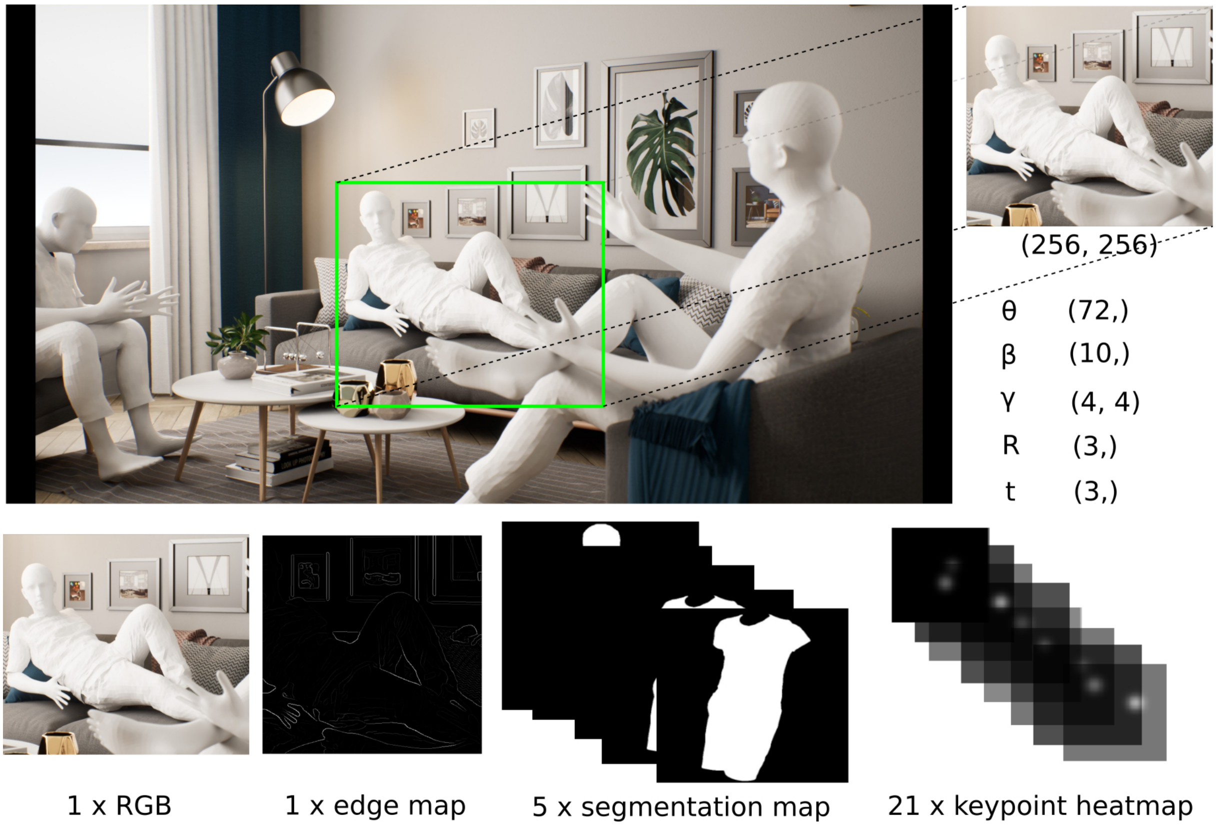


Figure 5.6: An example of a single input sample on ClothAGORA (non-textured). From a single render, we crop each person into images of size (256, 256); resize is applied if required. The person is specified using the set of parameters: pose θ, shape β, clothing styles γ, global orientation R, and global translation t. Using the crop, we prepare edges and keypoint heatmaps as derived image features.

in realistic 3D environments.

More examples of the ClothAGORA dataset are shown in Fig. 5.5.

5.3.3 Training and Inference

The model is first trained for 60 epochs on ClothSURREAL images and supervised using the corresponding pose, shape, and clothing style parameters, $(\theta_{gt}, \beta_{gt}, \gamma_{gt})$. Additionally, the model also uses 2D keypoint supervision, which has proven useful for improving pose estimation, especially in the first few epochs where the predictions tend to be converge to simple solutions and remain in the local optimum. After 60 epochs, the model is fine-tuned on ClothAGORA. The bounding boxes, such as the ones shown in Fig. 5.6, are used as input, i.e., the edges and keypoints are first extracted and then provided directly as input to the ResNet model. For supervision, the ground truth data also consist of pose, shape, and clothing style parameters, $(\theta_{gt}, \beta_{gt}, \gamma_{gt})$. We also experimented with using RGB images directly along with the cloth segmentation maps, but the results did not improve compared to using edges and keypoints only.

In inference, we take any in-the-wild image as input. Using off-the-shelf 2D keypoint detector and canny edge detector, we produce image features for ResNet50. The pose, shape,

Table 5.2: The comparison to the recent state-of-the-art method for the estimation of clothed people from images - Cloth-Wild on 3DPW test set [mm].

Method	CD ↓	CD-T ↓
Cloth-Wild [252]	**76.6**	34.8
Ours	98.9	**16.2**

and clothing style parameters, $(\theta_{est}, \beta_{est}, \gamma_{est})$, are estimated. The parameters are passed to the TailorNet model, which produces clothed mesh for visual evidence. More specifically, three individual components are produced: the unclothed body mesh, $B(\theta_{est}, \beta_{est})$, upper garment displacements, $D_{up}(\theta, \beta, \gamma_{up})$, and lower garment displacements, $D_{low}(\theta, \beta, \gamma_{low})$. In some cases, the clothing displacements intersect the body mesh, so we apply the intersection removal algorithm given by TailorNet. The intersection removal is applied three times: for body-lower garment, body-upper garment, and then combined lower garment and body with upper garment. The final mesh is a merge of the three processed components.

5.4 Experiments

5.4.1 Quantitative Evaluation

Evaluation metrics. For the underlying body shape evaluation, we use the per-vertex error on the Procrustes-aligned predictions (PVE-PA) and Procrustes-aligned per-vertex error in T-pose (PVE-T-PA). The per-vertex error is the average Euclidean distance between the corresponding vertices of the SMPL meshes. For the clothed meshes (upper and lower garment mesh and the body mesh), we do not know the correspondences between the predictions and ground truth. Therefore, we have to use a different distance measure. Following previous works [248, 252], we calculate the Chamfer distance measure (CD). In particular, we calculate Chamfer distance (CD) for posed predictions, and Chamfer in T-pose (CD-T) for unposed predictions. For 3D joint estimations, we use the standard mean per-joint precision error (MPJPE).

Estimation of clothed people. We use 3DPW [254] test set as an evaluation dataset. By following previous works [128, 131, 252], we sample each 25th frame for evaluation. We compare against Cloth-Wild, a method that relies on the parametric clothing model similar to us [252], and has previously demonstrated state-of-the-art performance for clothing geometry estimation. Table 5.2 shows that they are very competitive in terms of Chamfer distance (CD) on posed subjects, but are significantly outperformed when the subjects are in T-pose (CD-T).

Shape and pose estimation (undressed). The evaluation of shape estimation of the undressed bodies (by using the metrics on T-pose meshes) compared to ClothWild [252] and Sengupta et al. [128] is shown in Table 5.3. Our model performs comparable to Sengupta et al.,

Table 5.3: The comparison between the methods for the estimation of unclothed people from images on 3DPW test set [254] [mm].

Method	PVE-T-PA ↓	Chamfer-T-PA ↓
ClothWild [252]	21.5	24.8
HierProb3D [128]	**10.7**	**14.9**
Ours	12.1	16.2

Table 5.4: Body measurement estimation evaluation on 3DPW test set [mm].

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**Mean**
Ours	9.7	14.8	17.4	44.2	69.4	69.9	4.5	16.1	8.4	21.6	56.6	37.8	15.1	8.7	13.9	**27.2**

a strong competitor in the area of pose and shape estimation in-the-wild, while outperforming ClothWild. This results shows that our method is able to accurately estimate the underlying body shape of clothed people on a public benchmark, while also estimating the clothes geometry. Our estimations of body pose and global orientation still need additional training and augmentation in order to achieve more competitive performance, so we omit the comparisons for posed mesh estimates.

Body measurement estimation. Table 5.4 shows body measurement estimation results of Garmentor method on 3DPW test set. The body measurements are extracted from the estimated and ground truth SMPL meshes as described in Section 3.2.1. The error metric is mean absolute error (MAE) between the predicted and ground-truth measures.

Execution time.[§] In total, our method runs close to two seconds in inference on a consumer hardware, which is the fastest execution time compared to the most recent state-of-the-art methods, as shown in Table 5.5. Our execution time consists of several components, as highlighted in Table 5.6. The longest time it takes to resolve interpenetrations between the meshes. Secondly, edge detection takes almost half a second. Note that several components, such as Tailor-

[§] Note that our measured execution time for the previous method [252] is lower than the one reported in their paper (6.5s instead of 10.2s).

Table 5.5: The comparison between the methods that estimate a single clothed mesh and our method w.r.t. Chamfer distance between the predictions and ground truth on AGORA test set [127].

Method	Running time ↓
SMPLicit fit [270]	105.4s
ICON [248]	87.9s
PiFU-HD [250]	20.4s
Cloth-Wild [252]	6.5s
Ours	**1.7s**

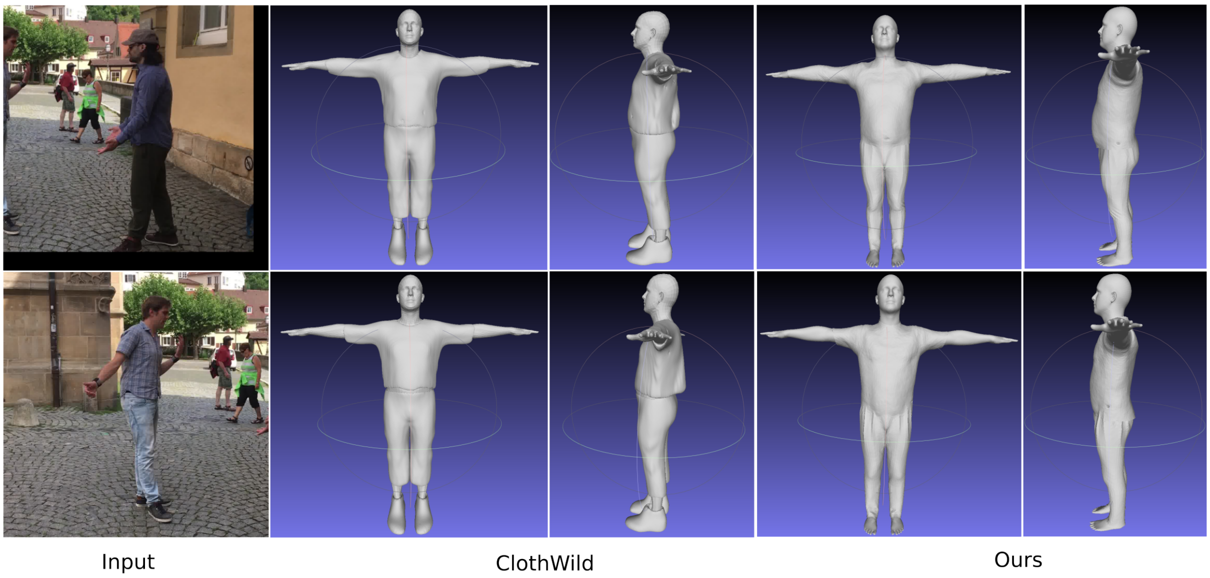


Figure 5.7: Qualitative comparison with ClothWild method in T-pose on 3DPW image examples.

Net, bounding box and keypoint detection run in real-time, while SMPL mesh generation and ResNet50 run close to real-time (∼25FPS).

5.4.2 Qualitative evaluation

In Fig. 4.3, we compare with ClothWild. Most notably, our garment meshes generally fit tighter to the body, i.e., ClothWild's meshes tend to be somewhat bulky. This is a consequence of using TailorNet [115] compared to SMPLicit [270] (used by ClothWild) as a clothing parametric model. Having tighter or bulkier clothes is not an advantage of none of the two methods in general. However, characteristic disadvantages of their meshes is that the pants are usually significantly shorter than average long pants, their meshes generic boots, and their shirts usually have larger displacements on the lower back part. All these details result in our final estimates being significantly more successful w.r.t. Chamfer-T distance, as shown by the quantitative evaluation. An advantage of ClothWild is that they do not differentiate between T-shirts and shirts in terms of garment classes (types); instead, their sleeve length is determined by the semantic segmentation predictions [253].

5.5 Discussion

5.5.1 Implementation Details

Network Architecture. ResNet18 does not have sufficient capacity to train the estimation of pose, shape, and clothing style parameters together. We notice that, after we include the clothing style parameters estimation, while using ResNet18, the PVE-PA error jumps four times

Table 5.6: The execution times of the particular components of our inference pipeline.

Component	Running time
Resolving interpenetrations	1.01s
Edge detection	0.49s
ResNet50	0.08s
SMPL mesh generation	0.07s
Keypoint detection	0.03s
Bounding box detection	0.02s
TailorNet (x2)	0.01s
Total time	**1.71s**

on validation set, comparing the first epochs between the models trained for pose-shape only and the ones trained for style also. However, when comparing the models trained for pose and shape estimation only, the performance on the test set is not significantly better using ResNet50 (up to 10%). In overall, we recommend ResNet50 as the backbone architecture of choice for the proposed task.

Calculating Chamfer distance. Note that our calculated mean Chamfer distance measure (CD) for the Cloth-Wild method is higher than the one reported in their paper [252]. We use the publicly available implementation of Chamfer distance [279], applied on the Procrustes-aligned meshes, without additional processing.

Training time. A single epoch (training and validation) using ResNet50 takes around 20 minutes on our NVidia Titan Xp. From our experience, it takes 100 or more epochs (>33 hours) to train the model that performs as good as possible on the test set. Our plan for future work is to continue training our final, representative model, which was trained for only 95 epochs.

Software packages and hyperparameters. Our learning model is written in PyTorch [173] and we use other Python libraries to generate ClothSURREAL, on Linux. For ClothAGORA, we additionally use Blender [280] and Unreal Engine [175], on Windows, to semi-automate the rendering and scene population. Our training requires 9GB of GPU memory (Titan Xp) for the batch size of 128 in case of ResNet18, i.e., around 6 GB for the batch size of 32 in case of ResNet50. We use around 6 GB of RAM on average for the training. A list of hyperparameters and additional implementation details are specified in the Appendix.

Texture application algorithm. The textures are obtained from the MGN dataset. In the original MGN work, the textures are applied directly to the SMPL body meshes [109]. To apply the same textures to our clothed SMPL models, we use given UV maps and textures from the dataset. We directly apply the given texture maps to our garments using the UV maps, since the topology of our garments correspond to the topology of the MGN garments (SMPL body displacements). To texture the body, however, we create an additional texture map, that replaces all the garments from the original texture map with an approximation of the skin color. We do so, to mitigate the problems of having visible cloth textures on the SMPL body mesh. We approximate the skin color of each texture map by averaging the color of a few sampled pixels located around the nose of the subject.

Interpenetration resolution algorithm. The interpenetration resolution algorithm is adapted from the one provided with the TailorNet's [115] public source code. The original algorithm resolves interpenetration between body and clothes by deforming the clothes mesh so that the vertices that are inside of the body end up at the nearest point on the surface of the body. We extend the procedure to resolve interpenetrations between the three meshes: upper garment, lower garment, and body mesh. First, an interpenetration resolution algorithm is applied to the lower garment and body mesh. Second, the algorithm is applied to the upper garment and the body. The body and lower garment meshes are then merged and, finally, the interpenetrations are solved between the merged lower-garment-plus-body mesh and the upper garment mesh. However, in some cases, the interpenetrations are not completely solved, which we have to investigate in future work.

5.6 Final Remarks and Future Work

Garmentor, the method for joint pose, shape, and clothes parameter estimation from images, is a promising approach towards complete human appearance understanding from single-view RGB images. The method is able to estimate the underlying body shape and pose, as well as the upper and lower garment on top of the body. We see body-from-clothes-separation approach as an interesting future research direction. In that sense, we propose the tools for automatic generation of novel poses, shapes, and clothes in realistic, synthetic 3D scenes, which can be used for training. We fully open source the tools to enhance further research.

Limitations and future work. While Garmentor is a high-level framework for learning any parametric human model, our current implementation depends on TailorNet, which brings some limitations. Garmentor does not support layered garments and the diversity of clothing is limited to four garment classes. Secondly, the current implementation of Garmetric depends on the performance of the pretrained feature extractors in inference, i.e., is not trained end-to-end. Except from end-to-end training, SMPLicit could be used as the parametric body model to compare to TailorNet. Also, TailorNet can be extended to support more garment classes. Extending the proposed datasets, especially ClothAGORA, is particularly exciting as it brings more realistic and free data to the community.

Potential negative societal impacts. Moving towards more realistic and controllable generation and reconstruction of people from images might have potential societal risks, such as

Table 5.7: The table of hyperparameters for training and inference of the Garmentor model. The *Train data* column specifies the hyperparameters used to generate ClothSURREAL data, and *Augm.* specifies augmentation.

	Hyperparameter	Value	Type	Description
Model	# Input channels	18	Int	Total # channels for input features (edges x 1, keypoints x 17)
	# ResNet layers	50	Int	Size of ResNet architecture (ResNet50)
	Embed. dim.	256	Int	The dimension of ResNet output before the final FCN
	# SMPL betas	10	Int	The number of shape principal components (10 out of 300)
	# Style params	4	Int	The number of style parameters (following TailorNet)
	# Garment types	4	Int	The number of different garment classes (types)
Train	Max. # Epochs	300	Int	The maximal number of training epochs
	Batch size	32	Int	Batch size in case of ResNet50 ($\sim$ 6GB GPU RAM)
	Learning rate	$10e^{-4}$	Float	Fixed learning rate
	Pin memory	True	Bool	PyTorch option; speeds up host-to-device data transfer
	# Workers	2	Int	PyTorch option; # subprocesses used for data loading
Loss	w_θ	80.0	Float	The weight next to L_θ loss component
	w_β	50.0	Float	The weight next to L_β loss component
	w_γ	50.0	Float	The weight next to L_γ loss component
	w_{2D}	$5.0e^3$	Float	The weight next to L_{2D} loss component
	w_R	$5.0e^3$	Float	The weight next to L_R loss component
	w_{3D}	$5.0e^3$	Float	The weight next to L_{3D} loss component
Train data	Focal length	300.0	Float	Focal length for the weak-perspective camera projection
	μ cam. trans.	$[0.0,-0.2,2.5]$	Float	Mean camera translation
	Shape (μ_β,σ_β)	$(0.0,1.25)$	Float	Mean and standard deviation of the shape parameters, $\mathcal{N}(\mu_\beta,\sigma_\beta)$
	Style $(\mu_\gamma,\sigma_\gamma)$	$(0.0,0.75)$	Float	Mean and standard deviation of the style parameters, $\mathcal{N}(\mu_\gamma,\sigma_\gamma)$
	Cam. XY σ	0.05	Float	The standard deviation of the XY camera offset
	Δ Z-range	$(-0.4,0.1)$	Float	The depth-range from which the camera location is sampled
Augm.	Occlude bottom prob.	0.02	Float	The probability of occluding the bottom part of the image
	Occlude top prob.	0.005	Float	The probability of occluding the top part of the image
	Occlude mid. prob.	0.05	Float	The probability of occluding the middle part of the image
Proxy features	Bbox. threshold	0.95	Float	The confidence threshold for the bounding box detection model
	Features size	(256, 256)	Int	The size of input features' map (edges and heatmaps)
	σ_{heatmaps}	4.0	Float	The standard deviation of the keypoint heatmaps, in pixels
	Non-max. supp.	True	Bool	Whether to apply non-maximum suppression on the edge detections
	Edge σ	1.0	Float	The stddev of Gaussian filter used for smoothness (edge)
	Edge Gauss. size	5.0	Float	The size of Gaussian filter, in pixels (edge)

creating convincing replicas of other people, which is already present for facial reconstruction. Until more clear regulations are established, we publish our source code and materials under an appropriate license.

Chapter 6

Discussion

This chapter discusses the overall dissertation. In particular, it comments on the performances of the proposed models w.r.t. body measurement estimation (the baseline described in Chapter 3 and the shape estimation model described in Chapter 5), it points out general limitations and the assumptions used during the evaluation of the proposed models, and, finally, future work is discussed. Note that the propose 3D pose estimation model from Chapter 4 does not directly measure the body, but can instead be used to improve the pose estimation for future work, in case shape estimation method from Chapter 5 would work from multiple views. We discuss these possibilities in Section 6.3.

Table 6.1: The comparison between the image-based model for shape estimation from images of clothed people (Chapter 5) and a linear regression baseline (Chapter 3) for the task of body measurement estimation of 15 standard body measurements defined in Chapter 1. The values are shown in millimeters.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**Mean**
Image-based	9.7	14.8	17.4	44.2	69.4	69.9	4.5	16.1	8.4	21.6	56.6	37.8	15.1	8.7	13.9	**27.2**
Linear base.	11.9	10.7	17.4	29.1	37.9	21.6	4.4	13.2	9.3	17.6	19.6	17.0	12.8	8.7	11.4	**16.2**

6.1 Comparing the Proposed Methods

Table 6.1 gathers the results previously reported in Table 3.6 and Table 5.4. It compares the shape estimation model evaluated on 3DPW dataset (image-based) with the linear baseline model evaluated on ANSUR dataset (estimating 15 body measurements using self-estimated body height and weight). The baseline model performs significantly better in overall, as well as for most of the individual body measurements, except for head (A) and forearm (I) circumference and shoulder-to-crotch (C) and ankle circumference (N), which are comparable.

These results show that a simple linear regression model which takes only the self-estimated height and weight as input outperforms the deep learning model trained for the estimation of body shape and clothes from images. The linear model thus poses a very challenging baseline, which is in accordance with Chapter 3, where other competitive deep learning approaches were also outperformed. On the other hand, the conditions in which the image-based model estimates body shape, i.e., body measurements, are very challenging due to variations in human pose, shape, clothing, and occlusions.

6.2 Limitations and Assumptions

There are several limitations and assumptions that apply to all the experiments demonstrated in the dissertation. In this Section, we briefly comment on the limitations and assumptions.

The lack of significant, accessible, and realistic benchmarks is the main drawback of body measurement estimates generally. Based on BODY-fit [177], the public benchmark employed in this study includes 3D template meshes matched to 3D scans. The SMPL fits are still an approximate representation of the original scans even though the fitted meshes represent the scanning dataset. Furthermore, the 3D scanning method is not flawless [29], thus the scans do not always accurately depict the original physical human bodies.

Therefore, the first assumption is that the manual body measurements are comparable to the body measurements obtained from the 3D scan, i.e., from the SMPL template mesh. In addition, all ANSUR body measurements do not necessarily physically correspond to the measurements from the SMPL model, such as shoulder-to-crotch or waist circumference, as specified in Section 3.3.1, but we assume they are highly linearly correlated and thus comparable. The height is measured as the difference between the top head point and the heel point on the y (height) axis. However, most of the datasets' subjects are expected to take approximately the A-pose, which is not fully erect. This might result in height being incorrectly estimated in some cases. We also assume that people wear simple or restrictive clothing and that hair artifacts won't have a substantial impact on bodily measurements like height and head circumference. For females, the issue with hair artifacts is particularly significant.

6.3 Future Work

In the future work, the models proposed in Chapters 3, 4, and 5 can be combined to further improve the performance of the final model. In this Section, we briefly discuss these possibilities.

Combining linear baseline and shape estimation model. Following on the performance of the baseline compared to the image-based shape estimation model, one might be tempted to combine the two approaches - estimate body measurements from images while also taking the self-estimated height and weight as additional input. This is actually possible and quite convenient using synthetic data. In fact, both ClothSURREAL and ClothAGORA datasets proposed in Chapter 5 can be extended in a way that they also contain information about the (self-estimation) body height and weight. However, testing such a model requires obtaining self-estimations for all of the subjects, as well as having their corresponding images. Note that no such public dataset exists so the method would either have to be tested on synthetic data as well, or only on a few subjects, otherwise, it would require significant dedication and time to gather more realistic test data.

Table 6.2: The comparison between the state-of-the-art monocular (upper part) and multi-view (lower part) 3D human pose estimation approaches on Human3.6M dataset. Even though the best monocular method performs comparably to the best multi-view method, note that the monocular method uses known 2D joint locations, which significantly improves its performance (see the performance of the same method in the first row while no ground-truth information is used). When comparing the monocular method that also estimate 2D joint locations and the best multi-view method, the difference is significant. The values are shown in millimeters.

Method	MPJPE	Using 2D ground-truth joints	Year
MixSTE [281]	39.8	No	2022
DiffPhy [282]	33.4	**No**	2022
PoseFormer [283]	31.3	Yes	2021
CrossFormer [284]	28.3	Yes	2022
MixSTE [281]	21.6	Yes	2022
Ours (Chapter 4)	29.1	No	2022
Epipolar Transformers [136]	19.0	No	2020
TesseTrack [285]	18.7	No	2021
Volumetric triangulation [134]	**17.7**	No	2019

Combining shape and pose estimation (multi-view). Another reasonable combination of approaches can be done with the presented model for 3D pose estimation from multiple views (Chapter 4) and shape estimation (Chapter 5). As expected, multi-view 3D pose estimation methods generally perform significantly better than monocular pose estimation due to the availability of information from different views at the same time, which helps to resolve ambiguities in 3D joint locations [44] (see Table 6.2 for the comparison of the state-of-the-art methods on a public benchmark [35]). Note that some monocular methods exploit 2D ground-truth joint location for their 3D pose estimations which helps them to significantly improve the result. Compared to the best monocular method that does exploit 2D ground truth, the generalizable approach, proposed and described in Chapter 4, still achieves better performance. Thus, it makes sense to combine the proposed pose estimation approach with the proposed shape estimation approach in a multi-view setting.

Chapter 7

Conclusion

The final chapter first provides separate conclusions on the three proposed approaches, and then the overall conclusion is given.

Linear regression baseline. The proposed regression approach is a straightforward but remarkably accurate tool for the quick and automatic estimate of body dimensions, without the need to take off clothing or take pictures. It shows that self-reported height and weight can predict body dimensions just as well as or even better than cutting-edge deep learning techniques. The baseline is ideal for a variety of applications, including virtual reality, ergonomics, and virtual try-on.

Human pose estimation method. The proposed generalizable approach is a promising novel direction for 3D human pose estimation. In particular, the demonstrated results show convincing generalization capabilities between different camera arrangements and datasets, outperforming previous methods. This allows the model to be trained on one or more camera arrangements and then be used in inference on different camera arrangements, which was not possible before or it was rather degraded compared to the performance on training data-like camera arrangement. The model requires relatively little training data, which makes training faster and more convenient for smaller datasets. The overall performance is competitive for 3D human pose estimation. By combining these two steps, it is possible to transfer the performance of the base dataset to any novel multi-camera dataset, in inference.

Human shape estimation method. The method for shape and clothes parameter estimation from images of clothed people, is a promising approach towards complete human appearance understanding from single-view RGB images. The method is able to estimate the underlying body shape and pose, as well as the upper and lower garment on top of the body. The body-from-clothes-separation approach is an interesting future research direction. In that sense, the tools for the automatic generation of novel poses, shapes, and clothes in realistic, synthetic 3D scenes, are proposed. As the approach is using the statistical body model (SMPL), the body measurements can be directly extracted from the estimated shapes.

The presented methods for human pose, shape, and body measurement estimation are significant steps toward automatic body measurement from images, without constraints. In particular, the person can wear loose clothes, take arbitrary body poses, and have various shapes and body measurements. The methods and their possible combinations (Chapter 6) open many possibilities toward accurate body measurements and understanding and analysis of the body from sensor data such as RGB camera. These results are important because the RGB cameras are omnipresent which allows to gather various data more easily.

Appendix A

Mobile Implementations

A.1 Catadioptric Adapter for Smartphones

We present a 3D printed adapter with planar mirrors for stereo reconstruction using front and back smartphone camera [286]. The adapter presents a practical and low-cost solution for enabling any smartphone to be used as a stereo camera, which is currently only possible using high-end phones with expensive 3D sensors. Using the prototype version of the adapter, we experiment with parameters like the angles between cameras and mirrors and the distance to each camera (the stereo baseline). We find the most convenient configuration and calibrate the stereo pair. Based on the presented preliminary analysis, we identify possible improvements in the current design. To demostrate the working prototype, we reconstruct a 3D human pose using 2D keypoint detections from the stereo pair and evaluate extracted body lengths. The result shows that the adapter can be used for anthropometric measurement of several body segments.

Stereo vision is a well-known approach for 3D reconstruction. It is a popular as it only requires two cameras and imposes relatively few constraints, such as textured scene and reasonably-wide baseline, compared to other 3D reconstruction techniques [29]. Stereo is important for numerous applications, such as 3D scanning [29], cultural heritage replications [287], SLAM [288], etc.

Regarding smartphones, 3D reconstruction is becoming more and more accessible with embedded ToF sensors on Androids and iPhones, but these technologies are still not affordable for the mass. Alternatively, an attempt towards stereo was made on a smartphone equipped with multiple back cameras [289]; however, the baseline between these cameras is very small, which degrades the reconstruction performance. Finally, there are works which take advantage of built-in video projectors [290, 291]; still, most smartphones and tablets are not equipped with them. We therefore present a low-cost solution for stereo on a smartphone — proposing a novel design of a 3D printed adapter with mirrors (catadioptric stereo) depicted in Fig. A.4, front and back camera, on standard smartphone into common field-of-view (FOV).

A.1.1 Catadioptric Stereo

In general, catadioptric systems consist of mirrors and lenses [292]. By using mirrors, a common part of the scene can be imaged from multiple views, which allows 3D reconstruction [44]. Most of the previously presented catadioptric systems use a single camera with multiple planar mirrors [293, 294, 295, 296, 297, 298]. Several works use prisms in combination with planar mirrors [299, 300]. The remaining works analyze the use of hyperbolic [301] and parabolic mirrors [298, 302].

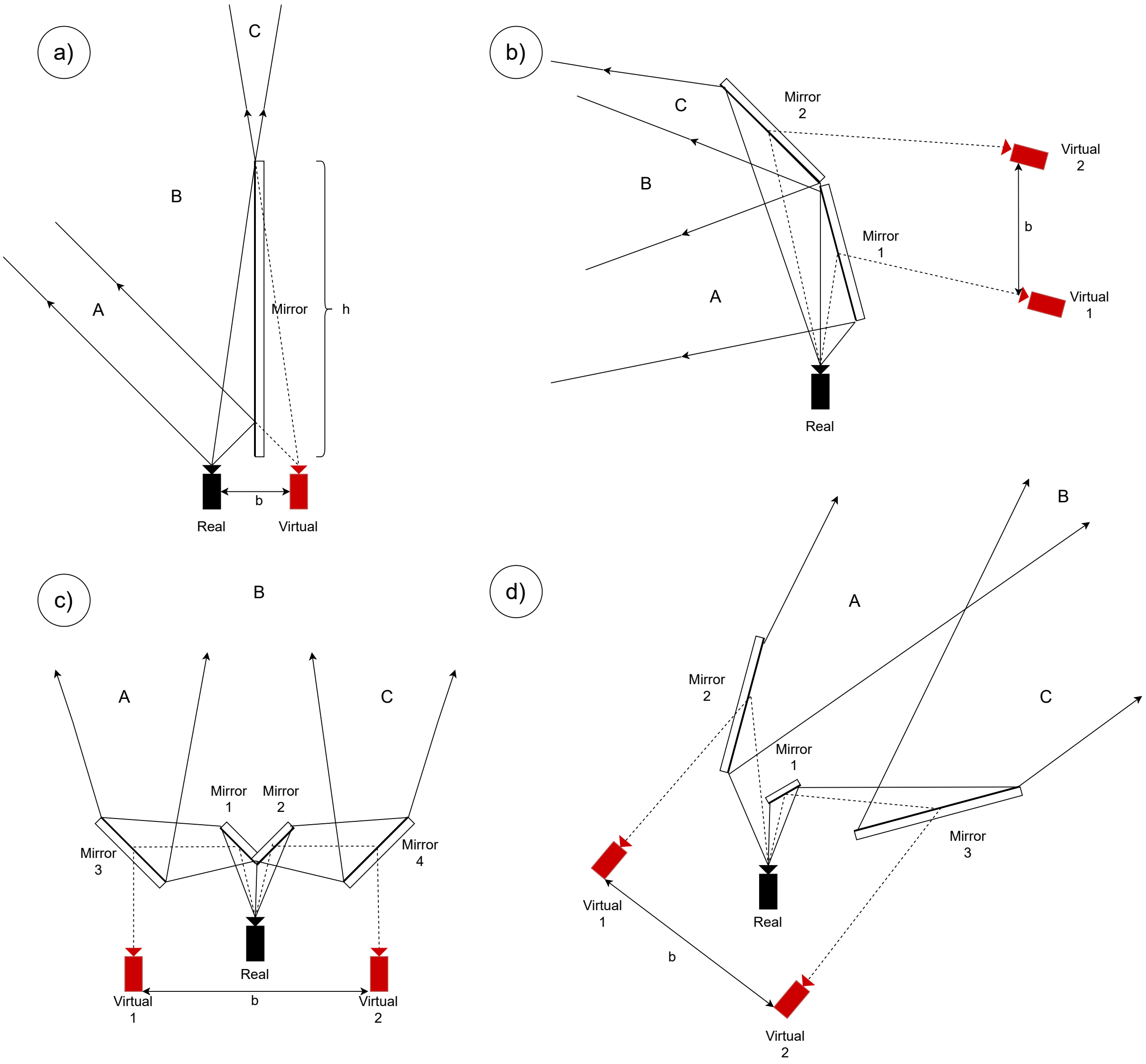


Figure A.1: Four previously proposed planar-mirror catadioptric configurations. The Figure shows real (black) and corresponding virtual (red) cameras, mirror placement, baselines, and three areas, A, B, and C, for each setup. Letter B represents a common area, visible by the stereo pair, while A and C are unused areas. Adapted from [286].

Regarding planar mirrors, Fig. A.1 shows four different catadioptric stereo configurations, previously described by Gluckman and Nayar [292]. All four systems use a single camera. The simplest configuration (Fig. A.1a) creates a single virtual camera, using the reflection in the mirror. A stereo pair consists of a real and a virtual camera. By moving a camera from the mirror, the baseline increases, but the common FOV between the images proportionally decreases. A two-mirror configuration (Fig. A.1b) produces two virtual cameras and these cameras comprise a stereo pair. The third configuration ((Fig. A.1c) uses four mirrors, as shown in a lower part of Fig. A.1. The advantage of third configuration over the second is that symmetric mirror setup produces virtual cameras without relative rotation, which is more suitable for stereo reconstruction, i.e. it does not require rectification [44]. However, the third configuration is not practical for 3D reconstruction of nearby objects, as the common FOV is relatively far from the virtual cameras. The fourth configuration (Fig. A.1d) fixes the above issue. Our system uses two planar mirrors and two cameras.

To the best of our knowledge, no peer-reviewed work* has been dedicated to extending multi-camera systems with mirrors or prisms to enable stereo on a standard smartphone. The analysis of such catadioptric systems makes sense today, with the advent of high-resolution front-back smartphone camera configurations.

In the remainder of Appendix A, we describe previous work on using mirrors for enabling stereo, then we present our adapter and its parameters. In the experimental section, we analyze the effect of the parameters to the FOV, calibration procedure, and reconstruction results.

A.1.2 Catadioptric Adapter

The 3D printed adapter consists of two mirrors to enable two cameras to record a common part of the scene (see Fig. A.4). To accomodate for different camera placements and to be able to experiment with different baselines and FOVs, we have several degrees of freedom:

- baseline (Fig. A.3),
- mirror angle (Fig. A.2 and A.3),
- vertical and horizontal tuning (Fig. A.3).

Horizontal and vertical tuning allow moving the mirrors to compensate for various camera positions on different smartphones. Baseline and angular tuning affect the common FOV of the virtual cameras (see the remainder of the section). The current version of phone cadle is designed specifically for Xiaomi Mi A2 smartphone, based on its dimensions, to perfectly sit into the mask and stay fixed during reconstruction for multiple experiments.

Baseline and Angular Tuning. When building catadioptric stereo, some parts of the images are unusable, on two levels. First, the mirror of each individual camera does not cover the whole FOV, and, second, the common FOV of the virtual cameras is smaller than the original, single-

*Note that there exist a webpage where multi-camera catadioptric systems are described [303].

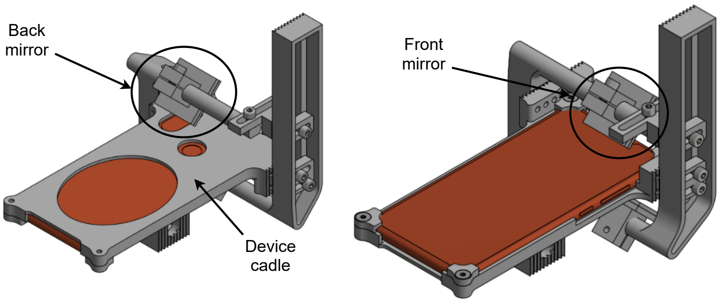


Figure A.2: The adapter design in a regular 3D view. The Figure features front and back mirrors and a smartphone mask, to keep the device fixed. The smartphone is almost normally usable for image recording, as the buttons on the side are available. Adapted from [286].

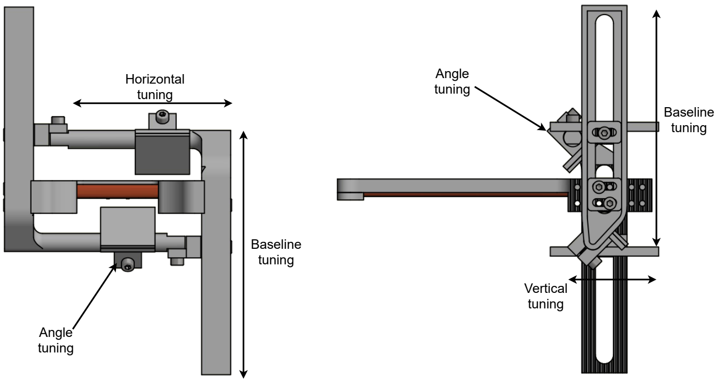


Figure A.3: The adapter design in forward and side view. The Figure features the parameters for adjusting mirror position (vertical, horizontal, and angular). Adapted from [286].

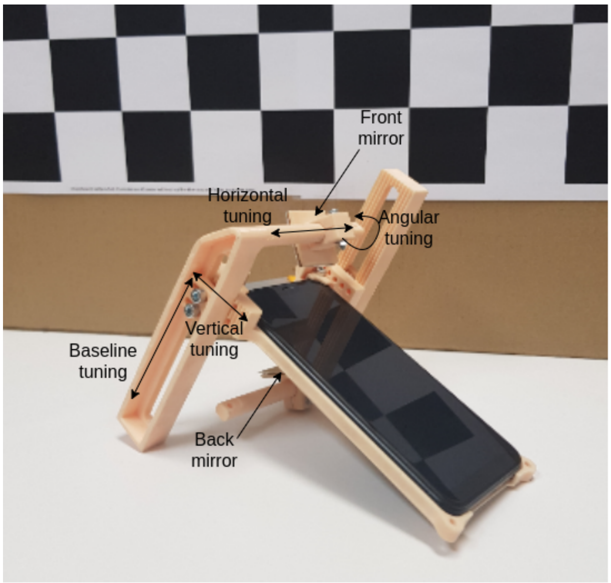


Figure A.4: The 3D printed adapter, pointing out front and back mirrors and baseline tuning. Adapted from [286].

camera FOV. In this Section, we analyze how the baseline, angles, and mirror sizes affect the common FOV of the virtual cameras. More specifically, we derive two quantities:

- the percentage of the common FOV, $\%_{\text{FOV}}$, retained compared to the original, single-

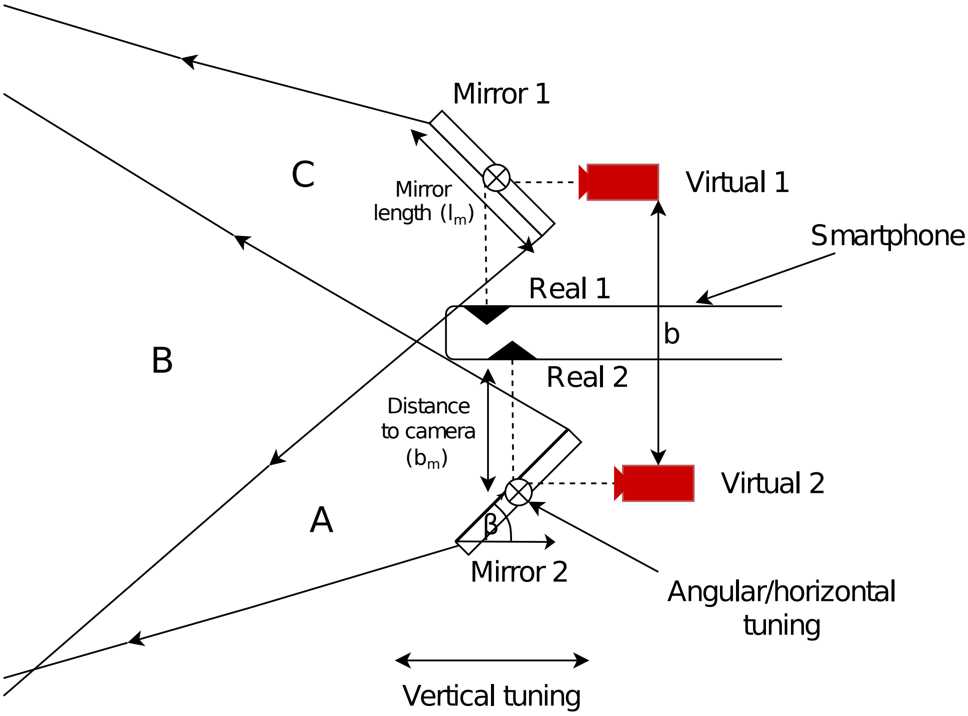


Figure A.5: The designed catadioptric stereo using front and back smartphone camera. Letter B shows the common FOV of the two virtual cameras. To adjust for the desired stereo pair properties and for different real camera positions, the mirrors can be moved in all three axes. Adapted from [286].

camera FOV, and

- the minimal distance between the virtual camera and the person, $d_{\min}$, needed to fit the average person ($h_{\text{avg}} = 1.8\,\text{m}$) into the common FOV.

First, we derive the angle α_{virtual} for a single camera (Fig. A.6). Then, we use the angle α_{virtual} to calculate the minimal distance, $d_{\min}$, needed to record a person, and the retained common FOV, $\%_{\text{FOV}}$. We take into account the following parameters (see Fig. A.5):

- distance b_m between the mirror and the camera,
- mirror length l_m,
- angle of the mirror β.

Note that changing the distance b_m directly affects the baseline b (see Fig A.5). Mirror length l_m, for squared mirrors such ours, results in a $P = l_m^2$ cm surface.

Individual FOV analysis. The FOV of the virtual camera, α_{virtual}, can be derived as a sum of left and right angles, $\alpha_{\text{virtual}} = \alpha_L + \alpha_R$, as shown in Fig. A.6. The left angle, α_L, spans between the middle of mirror and its left end, i.e. right end in case of the right angle, α_R. The angles can be calculated using the tangent function, as follows:

$$\tan \alpha_L = \frac{l_{m\_\text{proj}}}{2b_{m\_\text{down}}} \tag{A.1}$$

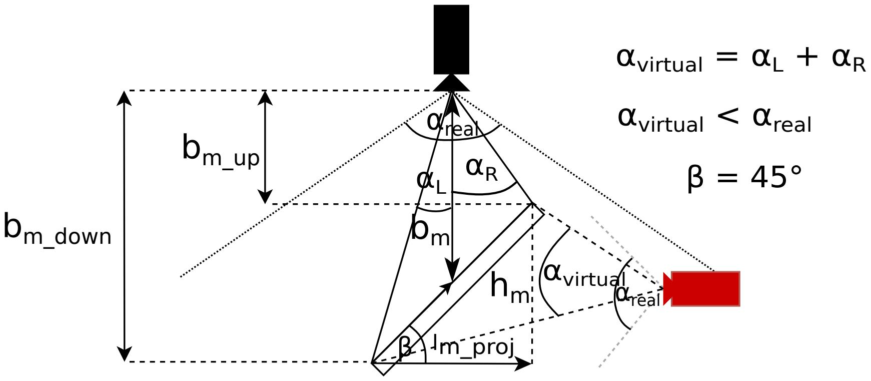


Figure A.6: The analysis of the virtual camera FOV with respect to the distance between the real camera and the mirror, d_m, mirror length l_m, and mirror angle β. Note that $\beta = 45°$ is shown for simplicity, so that the optical axis of the virtual camera is perpendicular to the real. Adapted from [286].

$$\tan\alpha_R = \frac{l_{m\_\text{proj}}}{2b_{m\_\text{up}}} \tag{A.2}$$

The l_{m_proj} is the mirror length projected on the line perpendicular to the optical axis of the real camera. The values b_{m_up} and b_{m_down} are the distances of the upper and the lower corner of the mirror. The mirror projection length is simply $l_{m\_\text{proj}} = l_m \cos\beta$, and the distances are $b_{m\_\text{up}} = b_m - \frac{h_m}{2}$ and $b_{m\_\text{down}} = b_m + \frac{h_m}{2}$. Mirror height h_m can be calculated as $h_m = l_m \sin\beta$. Virtual camera angle is therefore:

$$\alpha_{\text{virtual}} = \tan^{-1}\left(\frac{l_m\cos\beta}{2b_m + l_m\sin\beta}\right) + \tan^{-1}\left(\frac{l_m\cos\beta}{2b_m - l_m\sin\beta}\right) \tag{A.3}$$

The percentage of the retained FOV for an individual camera is $\%_{\text{FOV}} = \frac{\alpha_{\text{virtual}}}{\alpha_{\text{real}}}$.

Common FOV analysis. In the second part, we analyze what happens when mirror angle $\beta > 45°$ with respect to the common FOV (similar analysis can be done for $\beta < 45°$). We also find the minimal distance needed to record an average-height person. For simplicity, we assume that mirror angles β for both mirrors are the same (therefore, $\alpha_{in} = \alpha'_{in}$, as seen in Fig. A.7). We also assume that the real cameras are one beneath the other with respect to the smartphone, even though the physical cameras are not. The latter assumption does not significantly change the analysis, as the distances between the cameras are relatively small compared to the distance to the object, as will be shown in the remainder of the section.

Fig. A.7 shows a virtual camera configuration. The angle α_{in} can be calculated as $\alpha_{in} = 180° - \beta - \delta$. (see Fig. A.7, right). The unknown angle is $\delta = 180° - \alpha_R - \gamma$, where $\gamma = 90° - \beta$. Finally, the inner angle is $\alpha_{in} = 90° + \alpha_R - 2\beta$.

The minimal distance between the virtual camera and the person is:

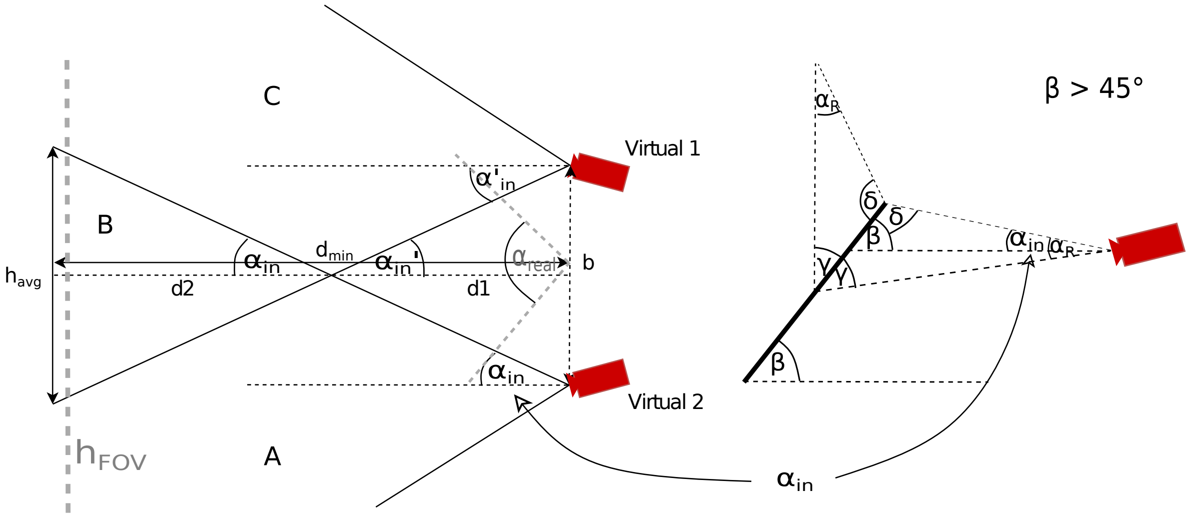


Figure A.7: The analysis of virtual cameras and their common FOV, assuming mirror angle to be $\beta > 45°$. Adapted from [286].

$$d_{\text{min}} = d_1 + d_2 = \frac{b + h_{\text{avg}}}{2 \tan \alpha_{in}}. \tag{A.4}$$

The height visible from the single, real camera can be calculated as $h_{\text{FOV}} = 2 d_{\text{min}} \tan \frac{\alpha_{\text{real}}}{2}$. Assuming that original single-camera FOV is $\alpha_{\text{real}} = 80°$ and $h_{\text{avg}} = 1.8\,\text{m}$, then the percentage of the retained, the percentage of common FOV between the virtual cameras is:

Selecting parameters. Based on the analysis and our empirical observations, we decided to set both mirrors to (roughly) $\beta = 55°$, $b_m = 2.5\,\text{cm}$ apart from the device, and use a 3 cm x 3 cm ($l_m = 3\,\text{cm}$) mirror surfaces. Based on the equations in the previous section, each virtual camera FOV is reduced to $\alpha_{\text{virtual}} = 47.08°$ (59%), where $\alpha_R = 34.09°$. The inner angle is $\alpha_{in} = 14.08°$. For the given parameters, we expect that the baseline is $b \approx 5\,\text{cm}$. Therefore, the minimal distance for recording an average-height person, $h_{\text{avg}} = 1.8\,\text{m}$, is $d_{\text{min}} = 3.69\,\text{m}$, and the retained common FOV is $\%_{\text{FOV}} = 29.4\%$. The main reason for such a small common FOV are the small mirror sizes.

A.1.3 Evaluation

We describe and evaluate virtual stereo pair calibration and demonstrate 3D human pose reconstruction.

Calibration. For stereo pair calibration of two virtual cameras, we use a standard Zhang's calibration [304] in MATLAB. An example of a calibration pair of views is shown in Fig. A.8. Before the calibration, we flip and rotate the images so that both virtual cameras are upright and horizontally aligned.

We record 14 different pairs of views for the calibration. During and after the calibration,

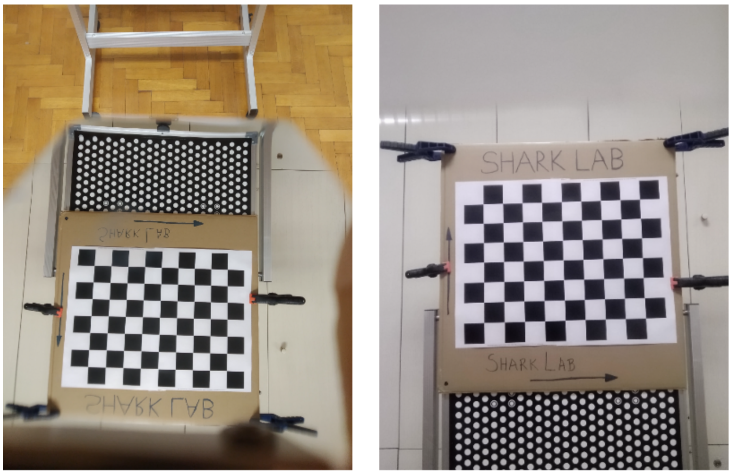


Figure A.8: Example of the original calibration chessboard pair of views (before image flipping and rotation), recorded using back and front camera, respectively. Adapted from [286].

Table A.1: Average distances (mm) between the reconstructed planes and the calibration planes, for each of the 14 views.

Axis/Plane	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Mean
X	3.06	2.34	3.98	2.69	2.07	2.26	2.13	4.05	3.72	2.86	4.20	4.92	5.08	5.53	**3.49**
Y	3.11	2.37	3.98	2.72	2.07	2.27	2.13	4.07	3.72	2.86	4.21	4.96	5.11	5.54	**3.51**
Z	4.23	3.67	4.58	2.66	1.67	2.12	1.60	3.29	4.64	3.68	4.98	4.50	5.06	6.31	**3.79**

the automatic adjustments done by the smartphone, such as autofocus, are turned off. The internal parameters are verified by comparing focal lengths to device's specifications. Qualitative calibration results are shown in Fig. A.9. Regarding extrinsics, there is a slight relative rotation between the mirrors, because the angles are not perfectly set to 45°. The distance between the virtual camera centers is 5.4 cm, as expected based on the analyses in Sec. A.1.2.

For quantitative evaluation, we calculate the distances between the reconstructed points of the calibration plate and the ideal calibration planes (Table A.1). Average distances are within few millimeters, which is reasonable. Mean reprojections errors per each image are shown in Fig. A.10. The overall mean error is 1.23 pixels. The calibration might be further improved by using more input pairs for the calibration.

3D Human Pose Reconstruction. To demonstrate the working prototype, we reconstruct a 3D human pose. To detect keypoints (human joints), we use OpenPose [140]. OpenPose detects 25 corresponding 2D keypoints on both front and back images, as shown in left and right part of Fig. A.11. The keypoints are then triangulated to produce the 3D skeleton shown in the middle of Fig. A.11.

To evaluate the reconstruction, we compare several body lengths of the skeleton with our manual measurements: lower arm, upper arm, shoulder width, hips width, upper leg, and lower leg. The measurements, differences and average error are shown in Table A.2. The average error is around 2 cm, which is an acceptable error for many applications, including anthropometry

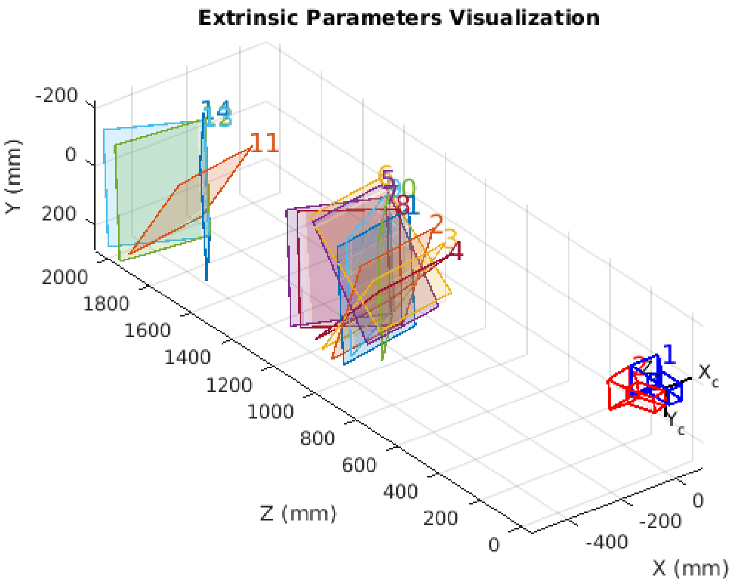


Figure A.9: Qualitative evaluation of the calibration. Red and blue cameras represent back and front camera, respectively. The planes on the left represent 14 pairs of views used for the calibration. Most of the views were recorded closer to the calibration chessboard, while others are about 1 m further. Adapted from [286].

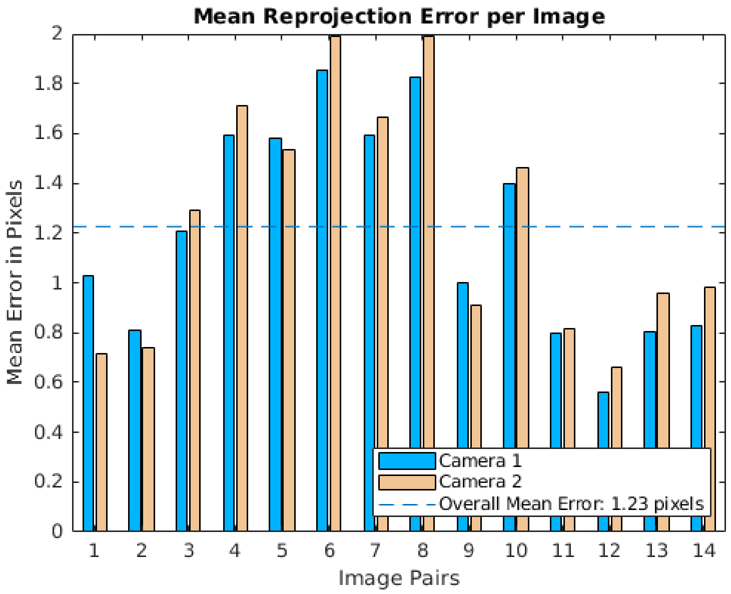


Figure A.10: Quantitative evaluation of the calibration. Camera 1 and 2 are back and front camera, respectively. Adapted from [286].

[29]. Notably, most of the measurements have an error below or around 1 cm, while shoulder width and lower leg length is much higher. The error in lower leg length might appear due to an error in knee keypoint reconstruction, seen in the middle of Fig. A.11. Similar reconstruction failure appears in the right shoulder and elbow, but the error did not propagate to the upper arm segment.

Table A.2: Quantitative reconstruction results for the recorded person. Six body measurements are evaluated: lower and upper arm, shoulders and hips width, upper and lower leg, respectively. The first row shows the reconstructed value, the second shows our manual measurement, and the third is the difference between the two. All numbers are shown in cm.

	L.arm	U.arm	Should.	Hips	U.leg	L.leg	Mean
3D	21.3	24.1	26.1	21.3	38.3	48.6	-
Meas.	21.2	24.7	31.0	22.1	38.9	43.5	-
Diff.	0.1	0.6	4.9	1.2	0.6	5.1	**2.1**

We attribute most of the reconstruction errors to the fact that the baseline between the virtual cameras is relatively small compared to the distance from which the subject is recorded. Small baseline results in large depth deviations for each correspondent pixel location error.

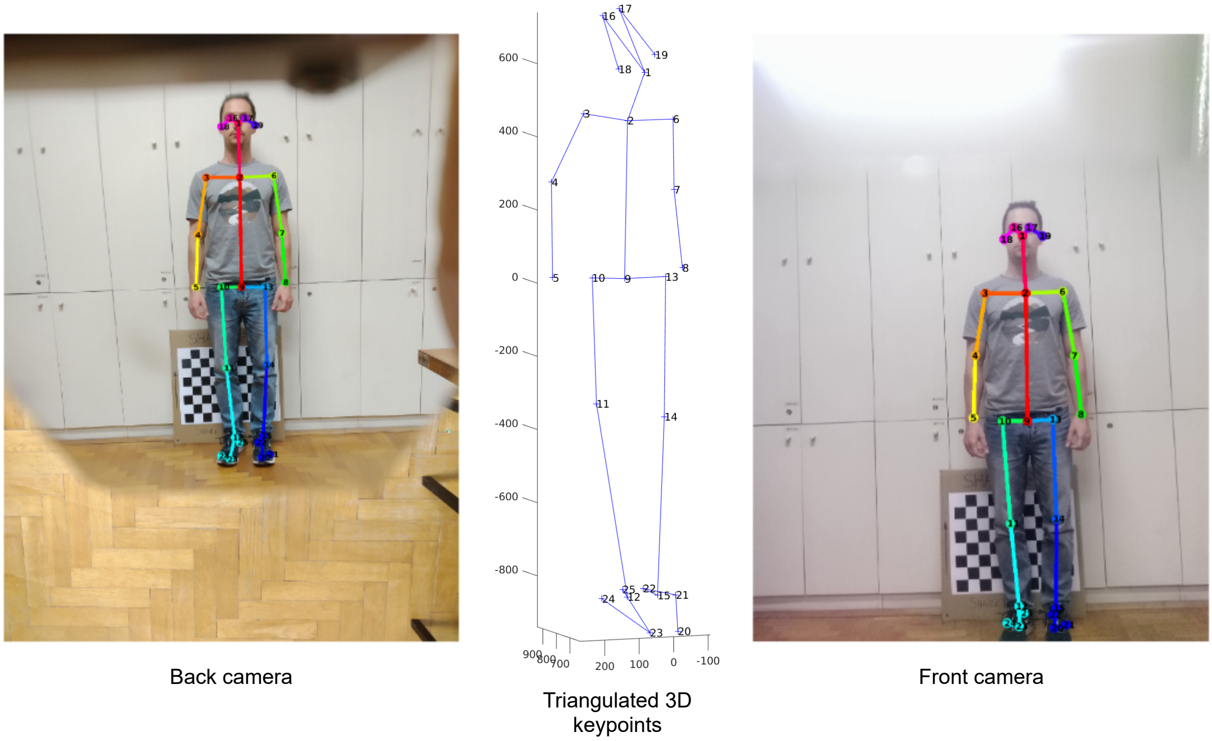


Figure A.11: Qualitative reconstruction results. Back and front images, along with the OpenPose keypoints, are shown in the left and right part of the Figure. In the middle, the triangulated 3D skeleton is shown in blue. Adapted from [286].

Conclusion. The design, analysis, and reconstruction using a prototype catadioptric stereo adapter for front and back smartphone camera is presented. The front-back camera stereo design is compared to previously proposed planar-mirror catadioptric system designs and the analysis of virtual camera FOV based on several degrees-of-freedom is described. The system is successfully calibrated and evaluated based on 3D human pose reconstruction. Taking into account the reconstruction results, we conclude that the reconstruction is successful and the adapter can be used for anthropometric measurements of body lengths. For future work, we propose using wider baseline between the virtual cameras, which requires proportionally larger mirror surfaces, as shown in the analysis. By using a wider baseline, 3D reconstructions should be further improved.

Appendix B

3D Scanners

B.1 ToF Scanners

Most of the commercial human body scanners, such as the SizeStream SS20, Styku S100, and TC^2-30R, are based on indirect ToF methods. In general, ToF as a standalone solution is unable to provide high-quality 3D human body scans due to its lower resolution. Hence, it is usually used in combination with RGB cameras. Noticeably, a bigger percentage of stationary scanners, such as the TC^2-19R, Naked scanner, and BodyGee Orbiter, come with a turntable on which subjects take a standard scanning position. This alleviates the problem of light interference caused by having multiple cameras. Note that all mini scanners are ToF-based and therefore used for 3D data acquisition in mobile applications (see Appendix B).

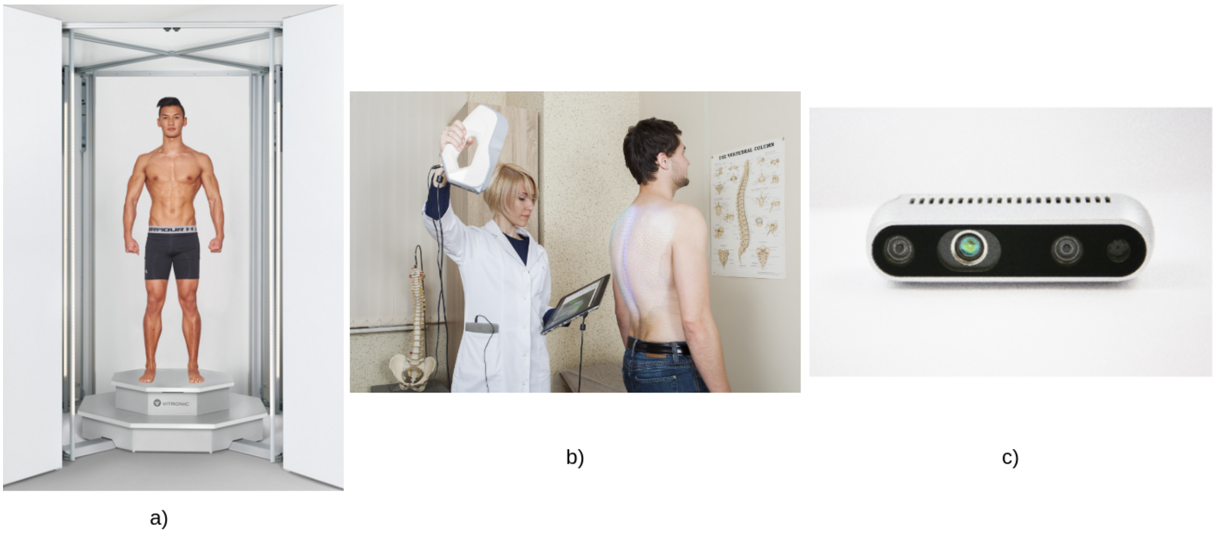


Figure B.1: Three types of 3D scanners in terms of mobility and size: stationary (a), handheld (b) and mini-scanners (c). Adapted from [29].

B.2 Structured Light Scanners

Commercial SL scanners either rotate around a person or have a fixed multi-sensor configuration that surrounds them. Stationary scanners, such as the HP Pro S3, 4DDynamics EOS, TC^2-105, or Hexagon Aicon Primescan, rotate around the body to obtain a whole 3D scan. Another way to move around the body is to use handheld scanners, such as the Artec Eva, TechMed3D BodyScan Scanner, Mantis Vision F6 Smart or ScanTech Axe B17. Stationary scanners with fixed sensor positions, such as the Artec Shapify Booth, botscan Neo, botscan OptaONE+, TC^2-105, and 4D Dynamics IIID Body Scan, showcase a booth filled with cameras and projectors in fixed positions that surround the scanned subject. Solutions to avoid light interference [305] from multiple projectors have been proposed, but in practice, every projector illuminates the subject in its designated time interval. Hence, the acquisition time is prolonged and proportional to the number of scanners.

B.3 Photogrammetry-Based Scanners

Commercial 3D scanners use either a rotating monocular system or multiple fixed cameras. For example, Texel Portal MX, Fit3D, and BodyGee Orbiter rotate a person that is standing on a platform, while Texel Portal BX circles around a static body. A few examples of fixed-camera scanning systems are Bootscan Neo, TC^2-21B, and 3IOSK by Mantis Vision, which uses from several to more than 50 RGB cameras to obtain the reconstruction. There are several advantages of fixed multi-camera over single-camera scanners. The first advantage is reduced scanning time, because neither the cameras nor the person need to move. The second advantage is the ability to scan people in motion over a period of time, also called 4D scanning (Move4D scanner by IBV). Thirdly, it is possible to reconstruct multiple people at once, if the scanning area is large enough to avoid occlusions, for example, as in Panoptic Studio [306].

Based on the images and the reconstruction described in this section, a mobile device camera can be considered a special case of a monocular PS-based scanner, where a camera is moved around a person to record a video or take individual images.

The following is the discussion based on the Appendix of a detailed report on 3D commercial scanners [29].

The work presents an overview of the commercial 3D scanners that have the ability to scan human bodies, excluding scanners that are not fit for the task, such as the Revopoint Tanso S1 [307], used to reconstruct smaller objects. We provide more than 80 currently available 3D scanners manufactured by more than 50 companies, as well as their taxonomy regarding several key characteristics: their mobility, method of reconstruction, price, resolution, accuracy, number of sensors, dimensions, provided texture, scanning time and provided anthropometric software. Additionally, we comment on their effect on human body scanning.

We observe an equal amount of stationary (booth-like) and handheld scanners, whereas only a few mini scanners are on the market. While handheld scanners offer a quicker scanning setup time in new environments, stationary scanners are more ideal for fixed scenarios, omitting (almost) entirely the setup process. Naturally, the mobility of a scanner is correlated with its dimensions. Stationary scanners are large and bulky, while mini scanners are compact and portable. Hence, mini and handheld scanners offer better applicability to the task of the distributed data collection process [1] since they present higher portability. On the other hand, stationary scanners offer faster scanning times, in the range of seconds, while handheld scanners offer scanning times in the range of minutes, presenting a trade-off between their dimensions and applicability. Since breathing and fidgeting causes human bodies to move during the scanning process, faster scanning times are more desirable. Nevertheless, the performance of handheld scanners does not seem to lag behind stationary 3D scanners, as seen by their accuracy.

The mobility and scanning time of a scanner seem to mostly drive its price. Smaller scanners tend to be cheaper, while scanners offering faster scanning times tend to be pricier, indicating that the market is still more appreciative towards stationary scanners. Most of the scanners use structured light (SL) to reconstruct the human body since it offers the best reconstruction accuracy compared to other two popular technologies - photogrammetry and time-of-flight. Additionally, they present the lowest resolution, followed by passive stereo (PS) and time-of-flight (ToF), respectively. Hence, they allow dense 3D human body reconstructions, appropriate for the anthropometric application. To this end, we additionally report if the scanner comes with an anthropometry software that can automatically extract body measurements from a 3D scan. While texture does not directly impact the scanning process, arguments have been made in favor of the greater usecase for textured 3D human body models [308].

The market is moving towards handheld and mini scanners. Mini scanners are particularly important for the future of tablet and smartphone scanning, because they can be attached to or even embedded into devices. For example, the Occipital sensors can be attached to a smartphone device, while the Apple iPhone 12 has an embedded LIDAR sensor (see Appendix B). Mini scanners are usually ToF-based [29]. As the computing capabilities of mobile devices improve further and ToF-based mini scanners increase their resolution, we expect that mobile devices will become more reliable and accurate 3D scanners.

Appendix C

Practical Recommendations

Based on the presented technologies, the proposed measurement framework, and the previous discussion, we finally provide practical recommendations for body measurement, as shown in Fig. C.1. First, the scanner classification is introduced. Next, specific pipelines are proposed with respect to their input. Finally, the requirements for the applications are described along with the introduced scanner types and pipeline recommendations.

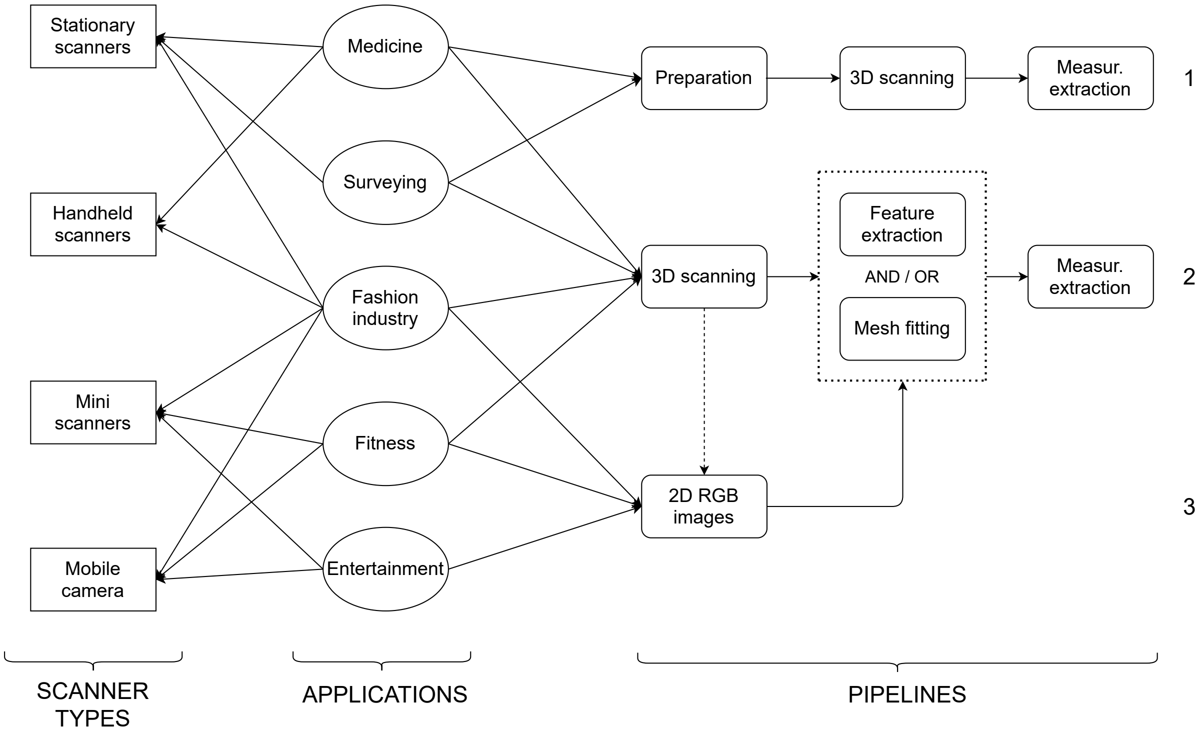


Figure C.1: The diagram of practical body measurement recommendations. Adapted from [29].

C.1 Scanner types

We classify scanners based on their mobility/size, into: (a) stationary; (b) handheld; (c) mini; and (d) mobile camera*. Stationary scanners (see Fig. B.1a) are usually installed in a fixed location, e.g. a lab or a medical facility. They are usually SL or PS based. Compared to other scanner types they are the most accurate and reliable and are therefore typically used to obtain ground truth data, e.g. stationary scanners were used to create 3D body scanning datasets like CAESAR [32], SIZE-UK [33], Scan DB [18], and FAUST [19]. Handheld scanners (see Fig. B.1b) are designed to be moved around the imaged body area by hand. Most of the existing handheld 3D scanners are SL based. Mini-scanners (see Fig. B.1c) are embedded in or attached to mobile devices like smartphones and tablets to enable 3D data acquisition. Most mini-scanners are ToF or SL based. Finally, we distinguish mobile RGB cameras as a separate scanner type, because they are wide-spread and convenient for non-demanding users, and usually rely simply on monocular measurement estimation techniques†. The four scanner types represent the data acquisition techniques for body measurement, as shown in Fig. C.1.

C.2 Measurement Pipelines

We propose and distinguish three possible pipelines for body measurement, as shown in the right part of Fig. C.1. The first pipeline, sufficient for majority of applications, consists of: preparation, 3D scanning, and measurement extraction. The second pipeline is more flexible and consists of: 3D scanning (without prior subject preparation), feature extraction along with or without mesh fitting, and measurement extraction. In both pipelines 2D images acquired using RGB cameras are often useful for improving the reconstruction [309]. Finally, the third and usually the least precise pipeline only takes 2D RGB images as input. These images are then used for feature extraction, mesh fitting, and measurement extraction.

C.3 Body Measurement Applications

We recommend specific measurement pipelines and scanner types for different anthropometric applications: medicine, surveying, fashion industry, fitness, and entertainment.

For medical applications [5, 6], it is usually desirable that high-quality body measurements are obtained. Therefore, 3D scanning using stationary or handheld scanners, along with the preparation stage (marker placement), is recommended (see the first pipeline in Fig. C.1). The measurements can then be directly extracted from the 3D scan (as decribed in Sec. 3.2.1).

*For more details on the currently available scanners on the market see Appendix A.

†For more details on mobile devices and applications for body measurement assessment, see Appendix B.

The second application is surveying, a systematical measurement of a population sample for the purpose of analyzing and tracking the properties of human bodies over time [1, 8]. High-quality surveys sometimes release their data publicly [32, 33], which serves for the creation and improvement of statistical models [31, 34, 108, 109, 111, 112]. Surveying is usually done using stationary scanners and the markers are sometimes placed on the body to improve and simplify the measurement [8].

For fashion industry applications (garment and clothing design), all of the four data acquisition techniques are used. For individually designed garments, stationary scanners are preferable [1]. For less reliable measurements and mass-produced clothes, other data acquisition techniques are sufficient.

For fitness and entertainment applications (gaming, AR, VR, etc.), low-budget solutions using mini scanners and mobile cameras are ideal for individual users. For fitness applications, the body measurements are used for tracking physical progress over time. As seen in the Appendix B, there are a few fitness-based mobile applications that estimate body measurements. Most of them use one or two RGB images from different views. For gyms or fitness centers, stationary 3D scanners might be more convenient. Regarding entertainment, 3D human pose [134, 153] in an AR setup allows the creation of a rigged character [113]; therefore only a rough estimate of body measurements is needed.

Bibliography

[1] Zakaria, N., Gupta, D., Anthropometry, Apparel Sizing and Design, ser. The Textile Institute Book Series. Elsevier Science, 2019, dostupno na: https://books.google.hr/books?id=s0K0DwAAQBAJ

[2] Allen, B., Curless, B., Popovic, Z., "The space of human body shapes: reconstruction and parameterization from range scans", ACM Trans. Graph., Vol. 22, 2003, str. 587–594.

[3] Robinette, K. M., Daanen, H., Paquet, E., "The caesar project: a 3-d surface anthropometry survey", in Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062), 1999, str. 380–386.

[4] Bougourd, J., Treleaven, P., "Uk national sizing survey–sizeuk", in International Conference on 3D Body Scanning Technologies, Lugano, Switzerland, 2010, str. 19–20.

[5] Heymsfield, S., Bourgeois, B., Ng, B., Sommer, M., Li, X., Shepherd, J., "Digital anthropometry: a critical review", European Journal of Clinical Nutrition, Vol. 72, 05 2018, str. 680–687.

[6] Đ onlić, M., "Three-dimensional analysis of back surface under dynamic conditions in scoliosis diagnostics", Doktorski rad, University of Zagreb, Faculty of Electrical Engineering and Computing, 2019.

[7] Casadei, K., Kiel, J., Anthropometric Measurement. StatPearls Publishing, Treasure Island (FL), 2020, dostupno na: http://europepmc.org/books/NBK537315

[8] Giancola, S., Valenti, M., Sala, R., "A survey on 3d cameras: Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies", in SpringerBriefs in Computer Science, 2018.

[9] D. Camba, J., Leon, A., Cantero, J., Saorín, J., Contero, M., "Application of low-cost 3d scanning technologies to the development of educational augmented reality content", in Frontiers in Education Conference, FIE, 10 2016, str. 1–6.

[10] "Basic human body measurements for technological design - Part 1: Body measurement definitions and landmarks", International Organization for Standardization, Standard, 2017.

[11] "3-D scanning methodologies for internationally compatible anthropometric databases - Part 1: Evaluation protocol for body dimensions extracted from 3-D body scans", International Organization for Standardization, Standard, 2018.

[12] Koepke, N., Zwahlen, M., Wells, J., Bender, N., Henneberg, M., Rühli, F., Staub, K., "Comparison of 3d laser-based photonic scans and manual anthropometric measurements of body size and shape in a validation study of 123 young swiss men", PeerJ, Vol. 5, 2017.

[13] Medina-Inojosa, J., Somers, V., Ngwa, T., Hinshaw, L., López-Jiménez, F., "Reliability of a 3d body scanner for anthropometric measurements of central obesity.", Obesity, open access, Vol. 2 3, 2016.

[14] Kouchi, M., Mochimaru, M., "Errors in landmarking and the evaluation of the accuracy of traditional and 3d anthropometry", Applied Ergonomics, Vol. 42, No. 3, 2011, str. 518–527, dostupno na: https://www.sciencedirect.com/science/article/pii/S0003687010001614

[15] Lescay, R. N., Becerra, A., González, A. H., "Anthropometry. comparative analysis of technologies for the capture of anthropometric dimensions", in Revista EIA, 2017.

[16] Daanen, H., van de Water, G. J., "Whole body scanners", Displays, Vol. 19, No. 3, 1998, str. 111–120, dostupno na: http://www.sciencedirect.com/science/article/pii/S0141938298000341

[17] Olds, T., Honey, F., "The use of 3d whole-body scanners in anthropometry", Kinanthropometry IX, 01 2005, str. 1–12.

[18] Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.-P., "A Statistical Model of Human Pose and Body Shape", Computer Graphics Forum, 2009.

[19] Bogo, F., Romero, J., Loper, M., Black, M. J., "FAUST: Dataset and evaluation for 3D mesh registration", in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, Jun. 2014.

[20] Zhang, C., Pujades, S., Black, M., Pons-Moll, G., "Detailed, accurate, human shape estimation from clothed 3D scan sequences", in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC, USA: IEEE Computer Society, July 2017, str. 5484–5493, spotlight.

[21] Pujades, S., Mohler, B. J., Thaler, A., Tesch, J., Mahmood, N., Hesse, N., Bülthoff, H. H., Black, M. J., “The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements”, IEEE Transactions on Visualization and Computer Graphics, Vol. 25, 2019, str. 1887–1897.

[22] Gordon, C. C., Blackwell, C. L., Bradtmiller, B., Parham, J. L., Barrientos, P., Paquette, S., Corner, B. D., Carson, J., Venezia, J., Rockwell, B. M., Mucher, M., Kristensen, S., “2012 anthropometric survey of u.s. army personnel: Methods and summary statistics”, 2015.

[23] Kouchi, M., Mochimaru, M., Tsuzuki, K., Yokoi, T., “Random errors in anthropometry”, Journal of Human Ergology, Vol. 25, No. 2, 1996, str. 155–166.

[24] Jamison, P. L., Zegura, S. L., “A univariate and multivariate examination of measurement error in anthropometry”, American Journal of Physical Anthropology, Vol. 40, No. 2, 1974, str. 197–203, dostupno na: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.1330400206

[25] Kouchi, M., Mochimaru, M., Tsuzuki, K., Yokoi, T., “Interobserver errors in anthropometry.”, Journal of human ergology, Vol. 28 1-2, 1999, str. 15–24.

[26] Ulijaszek, S. J., Kerr, D. A., “Anthropometric measurement error and the assessment of nutritional status”, British Journal of Nutrition, Vol. 82, No. 3, 1999, str. 165–177.

[27] Branson, R. S., Branson, R. S., Vaucher, Y. E., Harrison, G. G., Harrison, G. G., Vargas, M., Thies, C., “Inter- and intra-observer reliability of skinfold thickness measurements in newborn infants”, Human Biology, Vol. 54, No. 1, 1982, str. 137–143, dostupno na: http://www.jstor.org/stable/41463360

[28] Utermohle, C. J., Zegura, S. L., Heathcote, G. M., “Multiple observers, humidity, and choice of precision statistics: factors influencing craniometric data quality.”, American journal of physical anthropology, Vol. 61 1, 1983, str. 85–95.

[29] Bartol, K., Bojanić, D., Petković, T., Pribanić, T., “A review of body measurement using 3d scanning”, IEEE Access, Vol. 9, 2021, str. 67 281–67 301.

[30] Lu, J.-M., Wang, M.-J. J., “Automated anthropometric data collection using 3d whole body scanners”, Expert Syst. Appl., Vol. 35, No. 1–2, Jul. 2008, str. 407–414, dostupno na: https://doi.org/10.1016/j.eswa.2007.07.008

[31] Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., “Total capture: 3d human pose estimation fusing video and inertial sensors”, in 2017 British Machine Vision Conference (BMVC), 2017.

[32] "Caesar i, the most comprehensive source for body measurement data", http://store.sae.org/caesar/, accessed: 2020-11-20.

[33] "Uk national sizing survey", http://www.size.org/, accessed: 2020-11-20.

[34] Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J., "Scape: Shape completion and animation of people", ACM Trans. Graph., Vol. 24, No. 3, July 2005, str. 408–416, dostupno na: https://doi.org/10.1145/1073204.1073207

[35] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, No. 7, July 2014, str. 1325–1339.

[36] Smith, B. M., Chari, V., Agrawal, A., Rehg, J. M., Sever, R., "Towards accurate 3d human body reconstruction from silhouettes", in 2019 International Conference on 3D Vision (3DV), 2019, str. 279–288.

[37] Norton, K., Olds, T., "Anthropometrica : a textbook of body measurement for sports and health courses", 1996.

[38] Bragança, S., Arezes, P., Carvalho, M., Ashdown, S. P., Castellucci, I., Leão, C., "A comparison of manual anthropometric measurements with kinect-based scanned measurements in terms of precision and reliability", Work, Vol. 59, No. 3, Apr. 2018, str. 325–339, dostupno na: https://doi.org/10.3233/WOR-182684

[39] Kuehnapfel, A., Ahnert, P., Loeffler, M., Broda, A., Scholz, M., "Reliability of 3d laser-based anthropometry and comparison with classical anthropometry", Scientific Reports, Vol. 6, No. 1, May 2016, dostupno na: https://doi.org/10.1038/srep26672

[40] Yan, S., Wirta, J., Kämäräinen, J.-K., "Anthropometric clothing measurements from 3d body scans", Machine Vision and Applications, Vol. 31, 2020, str. 1–11.

[41] Daneshmand, M., Helmi, A., Avots, E., Noroozi, F., Alisinanoglu, F., Arslan, H., Gorbova, J., Haamer, R. E., Ozcinar, C., Anbarjafari, G., "3d scanning: A comprehensive survey", ArXiv, Vol. abs/1801.08863, 2018.

[42] El-Hakim, S., Beraldin, J.-A., Blais, F., "A comparative evaluation of the performance of passive and active 3-d vision systems", Proceedings of SPIE - The International Society for Optical Engineering, 05 2003.

[43] Remondino, F., Guarnieri, A., Vettore, A., "3d modeling of close-range objects: Photogrammetry or laser scanning", Proc SPIE, Vol. 5665, 12 2004, str. 216–225.

[44] Hartley, R., Zisserman, A., Multiple View Geometry in Computer Vision, 2nd ed. USA: Cambridge University Press, 2003.

[45] Loop, C., Zhang, Z., "Computing rectifying homographies for stereo vision", Vol. 1, 02 1999, str. -131 Vol. 1.

[46] Hartley, R. I., Sturm, P., "Triangulation", Computer Vision and Image Understanding, Vol. 68, No. 2, November 1997, str. 146–157, dostupno na: http://www.sciencedirect.com/science/article/B6WCX-45MFXC9-2/2/016263b42cb54133a006806cf452c40a

[47] Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., "A multi-view stereo benchmark with high-resolution images and multi-camera videos", in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 2538–2547.

[48] Furukawa, Y., Hernández, C., Multi-View Stereo: A Tutorial, 2015.

[49] Özyeşil, O., Voroninski, V., Basri, R., Singer, A., "A survey of structure from motion.", Acta Numerica, Vol. 26, 2017, str. 305–364.

[50] Schönberger, J. L., Frahm, J.-M., "Structure-from-motion revisited", in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, str. 4104–4113.

[51] Bianco, S., Ciocca, G., Marelli, D., "Evaluating the performance of structure from motion pipelines", Journal of Imaging, Vol. 4, 08 2018, str. 98.

[52] Lowe, D. G., "Distinctive image features from scale-invariant keypoints", Int. J. Comput. Vision, Vol. 60, No. 2, Nov. 2004, str. 91–110, dostupno na: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[53] Rublee, E., Rabaud, V., Konolige, K., Bradski, G., "Orb: An efficient alternative to sift or surf", in 2011 International Conference on Computer Vision, 2011, str. 2564–2571.

[54] Rosten, E., Drummond, T., "Machine learning for high-speed corner detection", in Computer Vision – ECCV 2006, Leonardis, A., Bischof, H., Pinz, A., (ur.). Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, str. 430–443.

[55] Alcantarilla, P., Nuevo, J., Bartoli, A., "Fast explicit diffusion for accelerated features in nonlinear scale spaces", in BMVC, 2013.

[56] Triggs, B., McLauchlan, P. F., Hartley, R. I., Fitzgibbon, A. W., "Bundle adjustment — a modern synthesis", in Vision Algorithms: Theory and Practice, Triggs, B., Zisserman, A., Szeliski, R., (ur.). Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, str. 298–372.

[57] Park, S. I., Hodgins, J. K., “Capturing and animating skin deformation in human motion”, ACM Trans. Graph., Vol. 25, No. 3, Jul. 2006, str. 881–889, dostupno na: https://doi.org/10.1145/1141911.1141970

[58] Liu, Y., Pears, N., Rosin, P. L., Huber, P., (ur.), 3D Imaging, Analysis and Applications. Springer International Publishing, 2020, dostupno na: https://doi.org/10.1007/978-3-030-44070-1

[59] Hafeez, J., Kwon, S., Lee, S., Hamacher, A., “3d surface reconstruction of smooth and textureless objects”, in 2017 International Conference on Emerging Trends Innovation in ICT (ICEI), 2017, str. 145–149.

[60] Hafeez, J., Hamacher, A., Kwon, S., Lee, S., “Performance evaluation of patterns for image-based 3d model reconstruction of textureless objects”, in 2017 International Conference on 3D Immersion (IC3D), 2017, str. 1–5.

[61] Hosseininaveh Ahmadabadian, A., Karami, A., Yazdan, R., “An automatic 3d reconstruction system for texture-less objects”, Robotics and Autonomous Systems, Vol. 117, 2019, str. 29–39, dostupno na: http://www.sciencedirect.com/science/article/pii/S0921889017307431

[62] Salvi, J., Mouaddib, E., Batile, J., “An overview of the advantages and constraints of coded pattern projection techniques for autonomous navigation”, in Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robot and Systems. Innovative Robotics for Real-World Applications. IROS ’97, Vol. 3, 1997, str. 1264–1271 vol.3.

[63] Lanman, D., Taubin, G., “Build your own 3d scanner: 3d photography for beginners”, ACM SIGGRAPH 2009 Courses, SIGGRAPH ’09, 01 2009, str. 8.

[64] Van der Jeught, S., Dirckx, J. J., “Real-time structured light profilometry: a review”, Optics and Lasers in Engineering, Vol. 87, 2016, str. 18–31, digital optical & Imaging methods in structural mechanics, dostupno na: http://www.sciencedirect.com/science/article/pii/S0143816616000166

[65] Bell, T., Li, B., Zhang, S., Structured Light Techniques and Applications. American Cancer Society, 2016, str. 1–24, dostupno na: https://onlinelibrary.wiley.com/doi/abs/10.1002/047134608X.W8298

[66] Abdelhafiz, A., “Integrating digital photogrammetry and terrestrial laser scanning”, Doktorski rad, 02 2009.

[67] Ebrahim, M., “3d laser scanners: History, applications, and future”, 10 2014.

[68] Boehler, W., Vicent, M., Marbs, A., “Investigating laser scanner accuracy”, Proc. CIPA XIXth Int. Symposium, Vol. 34, 01 2003.

[69] Van Gestel, N., Cuypers, S., Bleys, P., Kruth, J.-P., “A performance evaluation test for laser line scanners on cmms”, Optics and Lasers in Engineering, Vol. 47, No. 3, 2009, str. 336–342, optical Measurements, dostupno na: http://www.sciencedirect.com/science/article/pii/S0143816608001292

[70] D’Apuzzo, N., “3D body scanning technology for fashion and apparel industry”, in Videometrics IX, Beraldin, J.-A., Remondino, F., Shortis, M. R., (ur.), Vol. 6491, International Society for Optics and Photonics. SPIE, 2007, str. 203 – 214, dostupno na: https://doi.org/10.1117/12.703785

[71] Ribo, M., Brandner, M., “State of the art on vision-based structured light systems for 3d measurements”, in International Workshop on Robotic Sensors: Robotic and Sensor Environments, 2005., 2005, str. 2–6.

[72] Geng, J., “Structured-light 3d surface imaging: a tutorial”, Adv. Opt. Photon., Vol. 3, No. 2, Jun 2011, str. 128–160, dostupno na: http://aop.osa.org/abstract.cfm?URI=aop-3-2-128

[73] Salvi, J., Fernandez, S., Pribanić, T., Llado, X., “A state of the art in structured light patterns for surface profilometry”, Pattern Recognition, Vol. 43, No. 8, 2010, str. 2666–2680, dostupno na: http://www.sciencedirect.com/science/article/pii/S003132031000124X

[74] Salvi, J., Pagès, J., Batlle, J., “Pattern codification strategies in structured light systems”, Pattern Recognit., Vol. 37, 2004, str. 827–849.

[75] Pages, J., Salvi, J., Garcia, R., Matabosch, C., “Overview of coded light projection techniques for automatic 3d profiling”, in 2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422), Vol. 1, 2003, str. 133–138 vol.1.

[76] Kawasaki, H., Furukawa, R., Sagawa, R., Yagi, Y., “Dynamic scene shape reconstruction using a single structured light pattern”, 06 2008.

[77] Gan, Z., Visual sensing and its applications intergration of laser sensors to industrial robots, ser. Advanced topics in science and technology in China. Hangzhou: Zhejiang University PressHeidelberg, 2011, includes bibliographical references and index.

[78] Fofi, D., Sliwa, T., Voisin, Y., “A comparative survey on invisible structured light”, SPIE Electronic Imaging-Machine Vision Applications in Industrial Inspection XII, San José, USA, Vol. 5303, 05 2004, str. 90–97.

[79] Pribanić., T., Petković, T., Bojanić., D., Bartol, K., Gupta., M., “Scene adaptive structured light 3d imaging”, in Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,, INSTICC. SciTePress, 2020, str. 576–582.

[80] Đonlić, M., Petković, T., Pribanić, T., “3d surface profilometry using phase shifting of de bruijn pattern”, in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, str. 963–971.

[81] Sun, Z., Qian, G., Peng, Z., Dai, W., Sun, D., Zhang, G., Zhang, N., Xu, J., Wang, R., Li, C., “Orthogonal coded multi-view structured light for inter-view interference elimination”, in 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2020, str. 181–184.

[82] Sato, K., Inokuchi, S., “Three-dimensional surface measurement by space encoding range imaging.”, Journal of Robotic Systems, Vol. 2, No. 1, 1985, str. 27–39, cited By 105, dostupno na: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0022023379&partnerID=40&md5=4a7bced01abf552ec1bcae1095a67430

[83] Quan, C., He, X., Wang, C., Tay, C., Shang, H., “Shape measurement of small objects using lcd fringe projection with phase shifting”, Optics Communications, Vol. 189, No. 1, 2001, str. 21–29, dostupno na: http://www.sciencedirect.com/science/article/pii/S0030401801010380

[84] Pages, J., Collewet, C., Chaumette, F., Salvi, J., “An approach to visual servoing based on coded light”, in Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006., 2006, str. 4118–4123.

[85] Pagès, J., Salvi, J., Collewet, C., Forest, J., “Optimised de bruijn patterns for one-shot shape acquisition”, Image and Vision Computing, Vol. 23, No. 8, 2005, str. 707–720, dostupno na: http://www.sciencedirect.com/science/article/pii/S0262885605000508

[86] Li, L., “Time-of-flight camera - an introduction”, 2014.

[87] Horaud, R., Hansard, M. E., Evangelidis, G. D., Ménier, C., “An overview of depth cameras and range scanners based on time-of-flight technologies”, Machine Vision and Applications, Vol. 27, 2016, str. 1005–1020.

[88] Foix, S., Alenya, G., Torras, C., “Lock-in time-of-flight (tof) cameras: A survey”, IEEE Sensors Journal, Vol. 11, No. 9, 2011, str. 1917–1926.

[89] Piron, F., Morrison, D., Yuce, M. R., Redouté, J. M., “A review of single-photon avalanche diode time-of-flight imaging sensor arrays”, IEEE Sensors Journal, 2020, str. 1–1.

[90] Pribanić, T., Petković, T., Bojanić, D., Bartol, K., Gupta, M., “Smart time-multiplexing of quads solves the multicamera interference problem”, in Proc. 3DV, 2020.

[91] Bellisai, S., Villa, F., Tisa, S., Bronzi, D., Zappa, F., “Indirect time-of-flight 3D ranging based on SPADs”, in Quantum Sensing and Nanophotonic Devices IX, Razeghi, M., Tournie, E., Brown, G. J., (ur.), Vol. 8268, International Society for Optics and Photonics. SPIE, 2012, str. 282 – 289, dostupno na: https://doi.org/10.1117/12.908222

[92] He, Y., Chen, S., “Recent advances in 3d data acquisition and processing by time-of-flight camera”, IEEE Access, Vol. 7, 2019, str. 12 495–12 510.

[93] Payne, A. D., Jongenelen, A. P., Dorrington, A. A., Cree, M. J., Carnegie, D. A., “Multiple frequency range imaging to remove measurement ambiguity”, Conference held at Vienna, Austria, dostupno na: https://hdl.handle.net/10289/4032 Conference Contribution. Jul 2009.

[94] Gokturk, S. B., Yalcin, H., Bamji, C., “A time-of-flight depth sensor - system description, issues and solutions”, in 2004 Conference on Computer Vision and Pattern Recognition Workshop, 2004, str. 35–35.

[95] He, Y., Liang, B., Zou, Y., He, J., Yang, J., “Depth errors analysis and correction for time-of-flight (tof) cameras”, Sensors, Vol. 17, 01 2017, str. 92.

[96] Stoykova, E., Alatan, A. A., Benzie, P., Grammalidis, N., Malassiotis, S., Ostermann, J., Piekh, S., Sainov, V., Theobalt, C., Thevar, T., Zabulis, X., “3-d time-varying scene capture technologies—a survey”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, No. 11, 2007, str. 1568–1586.

[97] Wermke, F., Wübbenhorst, T., Meffert, B., “Interference avoidance for two time-of-flight cameras using autonomous optical synchronization”, 2020 6th International Conference on Control, Automation and Robotics (ICCAR), 2020, str. 586–595.

[98] Streeter, L., Kuang, Y. C., “Metrological aspects of time-of-flight range imaging”, IEEE Instrumentation Measurement Magazine, Vol. 22, No. 2, 2019, str. 21–26.

[99] Hansard, M., Lee, S., Choi, O., Horaud, R., Time-of-Flight Cameras. Springer London, 2013, dostupno na: https://doi.org/10.1007/978-1-4471-4658-2

[100] “Sizeusa dataset.”, https://www.tc2.com/size-usa.html, accessed: 2021-03-17.

[101] Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., “Total capture: 3d human pose estimation fusing video and inertial sensors”, in BMVC, 2017.

[102] Malleson, C., Volino, M., Gilbert, A., Trumble, M., Collomosse, J., Hilton, A., “Real-time full-body motion capture from video and imus”, in 2017 Fifth International Conference on 3D Vision (3DV), 2017.

[103] “4d human body motion scanning”, https://www.ibv.org/en/technologies/4d-human-body-motion-scanning/, accessed: 2020-03-12.

[104] Loper, M. M., Mahmood, N., Black, M. J., “MoSh: Motion and shape capture from sparse markers”, ACM Transactions on Graphics, (Proc. SIGGRAPH Asia), Vol. 33, No. 6, Nov. 2014, str. 220:1–220:13, dostupno na: http://doi.acm.org/10.1145/2661229.2661273

[105] Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., Black, M. J., “AMASS: Archive of motion capture as surface shapes”, in International Conference on Computer Vision, Oct. 2019, str. 5442–5451.

[106] Pons-Moll, G., Romero, J., Mahmood, N., Black, M. J., “Dyna: A model of dynamic human shape in motion”, ACM Transactions on Graphics, (Proc. SIGGRAPH), Vol. 34, No. 4, Aug. 2015, str. 120:1–120:14.

[107] Kolotouros, N., Pavlakos, G., Daniilidis, K., “Convolutional mesh regression for single-image human shape reconstruction”, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 4496–4505.

[108] Hirshberg, D., Loper, M., Rachlin, E., Black, M., “Coregistration: Simultaneous alignment and modeling of articulated 3D shape”, in European Conf. on Computer Vision (ECCV), ser. LNCS 7577, Part IV. Springer-Verlag, October 2012, str. 242–255.

[109] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M. J., “SMPL: A skinned multi-person linear model”, ACM Trans. Graphics (Proc. SIGGRAPH Asia), Vol. 34, No. 6, October 2015, str. 248:1–248:16.

[110] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., Black, M. J., “Expressive body capture: 3d hands, face, and body from a single image”, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 10 967–10 977.

[111] Osman, A. A. A., Bolkart, T., Black, M. J., “STAR: A sparse trained articulated human body regressor”, in European Conference on Computer Vision (ECCV), 2020, str. 598–613, dostupno na: https://star.is.tue.mpg.de

[112] Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B., “Building statistical shape spaces for 3d human modeling”, Pattern Recognition, 2017.

[113] Baran, I., Popović, J., “Automatic rigging and animation of 3d characters”, ACM Trans. Graph., Vol. 26, No. 3, Jul. 2007, str. 72–es, dostupno na: https://doi.org/10.1145/1276377.1276467

[114] Zhang, Y., Hassan, M., Neumann, H., Black, M. J., Tang, S., “Generating 3d people in scenes without people”, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, str. 6193–6203.

[115] Patel, C., Liao, Z., Pons-Moll, G., “Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style”, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun 2020.

[116] Sigal, L., Balan, A., Black, M. J., “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion”, International Journal of Computer Vision, Vol. 87, No. 1, Mar. 2010, str. 4–27.

[117] Anguelov, D., Srinivasan, P., Pang, H.-C., Koller, D., Thrun, S., Davis, J., “The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces”, in Proceedings of the 17th International Conference on Neural Information Processing Systems, ser. NIPS’04. Cambridge, MA, USA: MIT Press, 2004, str. 33–40.

[118] Kwok, T.-H., Yeung, K.-Y., Wang, C., “Volumetric template fitting for human body reconstruction from incomplete data”, Journal of Manufacturing Systems, Vol. 33, 2014, str. 678–689.

[119] Prokudin, S., Lassner, C., Romero, J., “Efficient learning on point clouds with basis point sets”, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, str. 4331–4340.

[120] Wang, M.-J. J., Wu, W.-Y., Lin, K.-C., Yang, S., Lu, J.-M., “Automated anthropometric data collection from three-dimensional digital human models”, The International Journal of Advanced Manufacturing Technology, Vol. 32, 2007, str. 109–115.

[121] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., “Microsoft coco: Common objects in context”, ArXiv, Vol. abs/1405.0312, 2014.

[122] Bojanić, D., Bartol, K., Pribanić, T., Petković, T., Donoso, Y. D., Mas, J. S., “On the comparison of classic and deep keypoint detector and descriptor methods”, in 2019 11th

International Symposium on Image and Signal Processing and Analysis (ISPA), 2019, str. 64–69.

[123] Armeni, I., Sax, A., Zamir, A. R., Savarese, S., “Joint 2D-3D-Semantic Data for Indoor Scene Understanding”, ArXiv e-prints, Feb. 2017.

[124] Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E., “Arkitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data”, 2021, dostupno na: https://arxiv.org/pdf/2111.08897.pdf

[125] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., Schmid, C., “Learning from synthetic humans”, in CVPR, 2017.

[126] Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M. J., “Populating 3D scenes by learning human-scene interaction”, in Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Jun. 2021.

[127] Patel, P., Huang, C.-H. P., Tesch, J., Hoffmann, D. T., Tripathi, S., Black, M. J., “AGORA: Avatars in geography optimized for regression analysis”, in Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Jun. 2021.

[128] Sengupta, A., Budvytis, I., Cipolla, R., “Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild”, in International Conference on Computer Vision, October 2021.

[129] Sengupta, A., Budvytis, I., Cipolla, R., “Probabilistic estimation of 3d human shape and pose with a semantic local parametric model”, in BMVC, 2021.

[130] Sengupta, A., Budvytis, I., Cipolla, R., “Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild”, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, str. 16 089–16 099.

[131] Sengupta, A., Budvytis, I., Cipolla, R., “Synthetic training for accurate 3d human pose and shape estimation in the wild”, in British Machine Vision Conference (BMVC), September 2020.

[132] Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K., “Probabilistic modeling for human mesh recovery”, in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[133] Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z., “Hierarchical kinematic human mesh recovery”, in Computer Vision – ECCV 2020, Vedaldi, A., Bischof,

H., Brox, T., Frahm, J.-M., (ur.). Cham: Springer International Publishing, 2020, str. 768–784.

[134] Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y., “Learnable triangulation of human pose”, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, str. 7717–7726.

[135] Bartol, K., Bojanić, D., Petković, T., Pribanić, T., “Generalizable human pose triangulation”, in Proceedings of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Jun. 2022.

[136] He, Y., Yan, R., Fragkiadaki, K., Yu, S.-I., “Epipolar transformers”, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, str. 7776–7785.

[137] Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W., “Cross view fusion for 3d human pose estimation”, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, str. 4341–4350.

[138] Tu, H., Wang, C., Zeng, W., “Voxelpose: Towards multi-camera 3d human pose estimation in wild environment”, in European Conference on Computer Vision (ECCV), 2020.

[139] Xiao, B., Wu, H., Wei, Y., “Simple baselines for human pose estimation and tracking”, in ECCV, 2018.

[140] Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., “Realtime multi-person 2d pose estimation using part affinity fields”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 1302–1310.

[141] Kocabas, M., Athanasiou, N., Black, M. J., “VIBE: Video inference for human body pose and shape estimation”, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, str. 5252–5262.

[142] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M. J., “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image”, in Computer Vision – ECCV 2016, ser. Lecture Notes in Computer Science. Springer International Publishing, October 2016.

[143] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., Black, M. J., “Expressive body capture: 3d hands, face, and body from a single image”, in Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[144] Kolotouros, N., Pavlakos, G., Black, M. J., Daniilidis, K., “Learning to reconstruct 3d human pose and shape via model-fitting in the loop”, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, str. 2252–2261.

[145] Vukotić, V., “Raspoznavanje objekata dubokim neuronskim mrežama”, Doktorski rad, University of Zagreb. Faculty of Electrical Engineering and Computing, Jul. 2014.

[146] Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A. L., Wang, X., “Multi-context attention for human pose estimation”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 5669–5678.

[147] Newell, A., Yang, K., Deng, J., “Stacked hourglass networks for human pose estimation”, in ECCV, 2016.

[148] Yang, W., Li, S., Ouyang, W., Li, H., Wang, X., “Learning feature pyramids for human pose estimation”, 2017 IEEE International Conference on Computer Vision (ICCV), 2017, str. 1290–1299.

[149] Chen, Y., Shen, C., Wei, X.-S., Liu, L., Yang, J., “Adversarial posenet: A structure-aware convolutional network for human pose estimation”, in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, str. 1221–1230.

[150] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., “2d human pose estimation: New benchmark and state of the art analysis”, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, str. 3686–3693.

[151] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., “Cascaded pyramid network for multi-person pose estimation”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, str. 7103–7112.

[152] “Coco: Coco leader board. *https://cocodataset.org/*”, https://cocodataset.org/#home, accessed: 2022-08-26.

[153] Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M., “3d human pose estimation in video with temporal convolutions and semi-supervised training”, in Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[154] Bulat, A., Kossaifi, J., Tzimiropoulos, G., Pantic, M., “Toward fast and accurate human pose estimation via soft-gated skip connections”, ArXiv, Vol. abs/2002.11098, 2020.

[155] Artacho, B., Savakis, A., “Unipose: Unified human pose estimation in single images and videos”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[156] Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C., “Distribution-aware coordinate representation for human pose estimation”, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, str. 7091–7100.

[157] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C., “Single-shot multi-person 3d pose estimation from monocular rgb”, in 3D Vision (3DV), 2018 Sixth International Conference on. IEEE, sep 2018, dostupno na: http://gvv.mpi-inf.mpg.de/projects/SingleShotMultiPerson

[158] Sun, K., Xiao, B., Liu, D., Wang, J., “Deep high-resolution representation learning for human pose estimation”, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 5686–5696.

[159] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., “Monocular 3d human pose estimation in the wild using improved cnn supervision”, in 3D Vision (3DV), 2017 Fifth International Conference on. IEEE, 2017, dostupno na: http://gvv.mpi-inf.mpg.de/3dhp_dataset

[160] Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F., “3DPeople: Modeling the Geometry of Dressed Humans”, in International Conference in Computer Vision (ICCV), 2019.

[161] Ascenso, G., Yap, M. H., Allen, T., Choppin, S. S., Payton, C., “A review of silhouette extraction algorithms for use within visual hull pipelines”, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 8, No. 6, 2020, str. 649–670, dostupno na: https://doi.org/10.1080/21681163.2020.1790040

[162] Lin, K., Wang, L., Luo, K., Chen, Y., Liu, Z., Sun, M.-T., “Cross-domain complementary learning using pose for multi-person part segmentation”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. PP, 05 2020, str. 1–1.

[163] Li, P., Xu, Y., Wei, Y., Yang, Y., “Self-correction for human parsing”, IEEE transactions on pattern analysis and machine intelligence, Vol. PP, 2020.

[164] Chen, L.-C., Collins, M. D., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J., “Searching for efficient multi-scale architectures for dense image prediction”, in NeurIPS, 2018.

[165] Fang, H., Lu, G., Fang, X., Xie, J., Tai, Y.-W., Lu, C., “Weakly and semi supervised human body part parsing via pose-guided knowledge transfer”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, str. 70–78.

[166] Xia, F., Wang, P., Chen, X., Yuille, A., “Joint multi-person pose estimation and semantic part segmentation”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 6080–6089.

[167] Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., Feng, J., “Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing”, in Proceedings of the 26th ACM International Conference on Multimedia, ser. MM ’18. New York, NY, USA: Association for Computing Machinery, 2018, str. 792–800, dostupno na: https://doi.org/10.1145/3240508.3240509

[168] Liang, X., Gong, K., Shen, X., Lin, L., “Look into person: Joint body parsing & pose estimation network and a new benchmark”, IEEE Transactions on Pattern Analysis & Machine Intelligence, Vol. 41, No. 04, apr 2019, str. 871–885.

[169] Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., Yuille, A., “The role of context for object detection and semantic segmentation in the wild”, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[170] Canny, J., “A computational approach to edge detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 6, 1986, str. 679–698.

[171] Davis, L. S., “A survey of edge detection techniques”, Computer Graphics and Image Processing, Vol. 4, No. 3, 1975, str. 248–270, dostupno na: https://www.sciencedirect.com/science/article/pii/0146664X7590012X

[172] Samet, H., Tamminen, M., “Efficient component labeling of images of arbitrary dimension represented by linear bintrees”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, No. 4, 1988, str. 579–586.

[173] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., “Pytorch: An imperative style, high-performance deep learning library”, in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, str. 8024–8035, dostupno na: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[174] Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J., “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop”, arXiv preprint arXiv:1506.03365, 2015.

[175] Epic Games, “Unreal engine”, https://www.unrealengine.com, accessed: 2019-04-25.

[176] Bartol, K., Bojanić, D., Petković, T., Peharec, S., Pribanić, T., “Linear regression vs. deep learning: A simple yet effective baseline for human body measurement”, Sensors, Vol. 22, No. 5, Feb 2022, str. 1885, dostupno na: http://dx.doi.org/10.3390/s22051885

[177] Yan, S., Wirta, J., Kämäräinen, J.-K., “Silhouette body measurement benchmarks”, in 2020 25th International Conference on Pattern Recognition (ICPR), 2021, str. 7804–7809.

[178] You, H., Ryu, T., “Development of a hierarchical estimation method for anthropometric variables”, International Journal of Industrial Ergonomics, Vol. 35, No. 4, 2005, str. 331–343, dostupno na: https://www.sciencedirect.com/science/article/pii/S0169814104001908

[179] Brolin, E., Högberg, D., Hanson, L., Örtengren, R., “Adaptive regression model for synthesizing anthropometric population data”, International Journal of Industrial Ergonomics, Vol. 59, 2017, str. 46–53, dostupno na: https://www.sciencedirect.com/science/article/pii/S0169814117301178

[180] PHEASANT, S. T., “A technique for estimating anthropometric data from the parameters of the distribution of stature”, Ergonomics, Vol. 25, No. 11, 1982, str. 981–992, pMID: 7173162, dostupno na: https://doi.org/10.1080/00140138208925059

[181] Kawakami, R., Miyachi, M., Tanisawa, K., Ito, T., Usui, C., Midorikawa, T., Torii, S., Ishii, K., Suzuki, K., Sakamoto, S., Higuchi, M., Muraoka, I., Oka, K., “Development and validation of a simple anthropometric equation to predict appendicular skeletal muscle mass”, Clinical Nutrition, Vol. 40, No. 11, 2021, str. 5523–5530, dostupno na: https://www.sciencedirect.com/science/article/pii/S0261561421004490

[182] Lu, J.-M., Wang, M.-J. J., “Automated anthropometric data collection using 3d whole body scanners”, Expert Systems with Applications, Vol. 35, No. 1, 2008, str. 407–414, dostupno na: https://www.sciencedirect.com/science/article/pii/S095741740700259X

[183] Kaashki, N. N., Hu, P., Munteanu, A., “Deep learning-based automated extraction of anthropometric measurements from a single 3-d scan”, IEEE Transactions on Instrumentation and Measurement, Vol. 70, 2021, str. 1–14.

[184] Wang, M.-J., Wu, W.-Y., Lin, K.-C., Yang, S.-N., Lu, J.-M., “Automated anthropometric data collection from three-dimensional digital human models”, International Journal of Advanced Manufacturing Technology, Vol. 32, 02 2007, str. 109–115.

[185] Lin, K., Wang, L., Liu, Z., “End-to-end human pose and mesh reconstruction with transformers”, in CVPR, 2021.

[186] Lin, K., Wang, L., Liu, Z., “Mesh graphormer”, in ICCV, 2021.

[187] Choi, H., Moon, G., Lee, K. M., “Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose”, ArXiv, Vol. abs/2008.09047, 2020.

[188] Tsoli, A., Loper, M., Black, M., “Model-based anthropometry: Predicting measurements from 3d human scans in multiple poses”, in 2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014, 03 2014.

[189] Balan, A. O., Sigal, L., Black, M. J., Davis, J. E., Haussecker, H. W., “Detailed human shape and pose from images”, in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, str. 1–8.

[190] Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K., “Learning to estimate 3d human pose and shape from a single color image”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, str. 459–468.

[191] Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M. J., “Monocular expressive body regression through body-driven attention”, in European Conference on Computer Vision (ECCV), 2020, str. 20–40, dostupno na: https://expose.is.tue.mpg.de

[192] Kocabas, M., Huang, C.-H. P., Hilliges, O., Black, M. J., “PARE: Part attention regressor for 3d human body estimation”, ArXiv, Vol. abs/2104.08527, 2021.

[193] Iqbal, U., Xie, K., Guo, Y., Kautz, J., Molchanov, P., “Kama: 3d keypoint aware body mesh articulation”, ArXiv, Vol. abs/2104.13502, 2021.

[194] Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C., “HybrIK: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation”, in CVPR, 2021.

[195] Choi, H., Moon, G., Lee, K. M., “Beyond static features for temporally consistent 3d human pose and shape from a video”, in Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[196] Joo, H., Neverova, N., Vedaldi, A., “Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation”, ArXiv, Vol. abs/2004.03686, 2020.

[197] Doersch, C., Zisserman, A., “Sim2real transfer learning for 3d pose estimation: motion to the rescue”, ArXiv, Vol. abs/1907.02499, 2019.

[198] Guan, S., Xu, J., Wang, Y., Ni, B., Yang, X., “Bilevel online adaptation for out-of-domain human mesh reconstruction”, ArXiv, Vol. abs/2103.16449, 2021.

[199] Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L. J., “Humor: 3d human motion model for robust pose estimation”, ArXiv, Vol. abs/2105.04668, 2021.

[200] Chen, Y.-C., Piccirilli, M., Piramuthu, R., Yang, M.-H., “Self-attentive 3d human pose and shape estimation from videos”, Computer Vision and Image Understanding, 2021, str. 103305, dostupno na: https://www.sciencedirect.com/science/article/pii/S1077314221001491

[201] Moon, G., Lee, K. M., “Pose2pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation”, ArXiv, Vol. abs/2011.11534, 2020.

[202] Yi, X., Zhou, Y., Xu, F., “Transpose: Real-time 3d human translation and pose estimation with six inertial sensors”, ACM Trans. Graph., Vol. 40, 2021, str. 86:1–86:13.

[203] Zhou, Y., Habermann, M., Habibie, I., Tewari, A., Theobalt, C., Xu, F., “Monocular real-time full body capture with inter-part correlations”, in CVPR, 2021.

[204] Madadi, M., Bertiche, H., Escalera, S., “SMPLR: Deep learning based smpl reverse for 3d human pose and shape recovery”, Pattern Recognition, Vol. 106, 2020, str. 107472, dostupno na: https://www.sciencedirect.com/science/article/pii/S0031320320302752

[205] Kanazawa, A., Black, M. J., Jacobs, D. W., Malik, J., “End-to-end recovery of human shape and pose”, in Computer Vision and Pattern Recognition (CVPR), 2018.

[206] Smith, B. M., Chari, V., Agrawal, A., Rehg, J. M., Sever, R., “Towards accurate 3d human body reconstruction from silhouettes”, in 2019 International Conference on 3D Vision (3DV), 2019, str. 279–288.

[207] Dibra, E., Jain, H., Öztireli, C., Ziegler, R., Gross, M., “Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks”, in 2016 Fourth International Conference on 3D Vision (3DV), 2016, str. 108–117.

[208] Dibra, E., Jain, H., Öztireli, C., Ziegler, R., Gross, M., “Human shape from silhouettes using generative hks descriptors and cross-modal neural networks”, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 5504–5514.

[209] Dibra, E., Öztireli, C., Ziegler, R., Gross, M., “Shape from selfies: Human body shape estimation using cca regression forests”, in Computer Vision – ECCV 2016, Leibe, B., Matas, J., Sebe, N., Welling, M., (ur.). Cham: Springer International Publishing, 2016, str. 88–104.

[210] Boisvert, J., Shu, C., Wuhrer, S., Xi, P., “Three-dimensional human shape inference from silhouettes: reconstruction and validation”, Machine Vision and Applications, Vol. 24, 2011, str. 145–157.

[211] Chen, Y., Kim, T.-K., Cipolla, R., “Inferring 3d shapes and deformations from single views”, in Computer Vision – ECCV 2010, Daniilidis, K., Maragos, P., Paragios, N., (ur.). Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, str. 300–313.

[212] Xi, P., Lee, W.-S., Shu, C., “A data-driven approach to human-body cloning using a segmented body database”, in 15th Pacific Conference on Computer Graphics and Applications (PG’07), 2007, str. 139–147.

[213] Yan, S., Kämäräinen, J.-K., “Learning anthropometry from rendered humans”, ArXiv, Vol. abs/2101.02515, 2021.

[214] Walpole, R. E., Myers, R. H., Myers, S. L., Ye, K., Probability & statistics for engineers and scientists, 8th ed. Upper Saddle River: Pearson Education, 2007.

[215] Lien, S.-l., Kajiya, J. T., “A symbolic method for calculating the integral properties of arbitrary nonconvex polyhedra”, IEEE Computer Graphics and Applications, Vol. 4, No. 10, 1984, str. 35–42.

[216] Krzywicki, H. J., Chinn, K. S., “Human body density and fat of an adult male population as measured by water displacement.”, The American journal of clinical nutrition, Vol. 20 4, 1967, str. 305–10.

[217] Yan, S., Wirta, J., Kämäräinen, J.-K., “Anthropometric clothing measurements from 3d body scans”, Machine Vision and Applications, Vol. 31, 01 2020.

[218] “Human solutions gmbh. anthroscan. http://www.”, https://www.human-solutions.com/, accessed: 2020-11-20.

[219] Heinze, G., Wallisch, C., Dunkler, D., “Variable selection – a review and recommendations for the practicing statistician”, Biometrical Journal, Vol. 60, No. 3, 2018, str. 431–449, dostupno na: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700067

[220] Cawley, J., Maclean, J., Hammer, M., Wintfeld, N., “Reporting error in weight and its implications for bias in economic models”, Economics & Human Biology, Vol. 19, No. C, 2015, str. 27–44.

[221] Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T. S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., “Panoptic studio: A massively

multiview system for social interaction capture", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[222] Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., "Convolutional pose machines", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, str. 4724–4732.

[223] Remelli, E., Han, S., Honari, S., Fua, P., Wang, R. Y., "Lightweight multi-view 3d pose estimation through camera-disentangled representation", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, str. 6039–6048.

[224] Kadkhodamohammadi, A., Padoy, N., "A generalizable approach for multi-view 3d human pose regression", Mach. Vis. Appl., Vol. 32, 2021, str. 6.

[225] Tomè, D., Toso, M., Agapito, L., Russell, C., "Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture", 2018 International Conference on 3D Vision (3DV), 2018, str. 474–483.

[226] Fischler, M. A., Bolles, R. C., "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography", Commun. ACM, Vol. 24, No. 6, Jun. 1981, str. 381–395, dostupno na: https://doi.org/10.1145/358669.358692

[227] Simon, T., Joo, H., Matthews, I., Sheikh, Y., "Hand keypoint detection in single images using multiview bootstrapping", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 4645–4653.

[228] Sun, J. J., Zhao, J., Chen, L.-C., Schroff, F., Adam, H., Liu, T., "View-invariant probabilistic embedding for human pose", in Computer Vision – ECCV 2020, Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., (ur.). Cham: Springer International Publishing, 2020, str. 53–70.

[229] Shuai, H., Wu, L., Liu, Q., "Adaptively multi-view and temporal fusing transformer for 3d human pose estimation", ArXiv, Vol. abs/2110.05092, 2021.

[230] Kocabas, M., Karagoz, S., Akbas, E., "Self-supervised learning of 3d human pose using multi-view geometry", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 1077–1086.

[231] Moon, G., Yu, S.-I., Wen, H., Shiratori, T., Lee, K. M., "Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image", ArXiv, Vol. abs/2008.09309, 2020.

[232] Bultmann, S., Behnke, S., “Real-time multi-view 3d human pose estimation using semantic feedback to smart edge sensors”, in Robotics: Science and Systems XVII, Virtual Event, July 12-16, 2021, Shell, D. A., Toussaint, M., Hsieh, M. A., (ur.), 2021, dostupno na: https://doi.org/10.15607/RSS.2021.XVII.040

[233] Bouazizi, A., Wiederer, J., Kressel, U., Belagiannis, V., “Self-supervised 3d human pose estimation with multiple-view geometry”, ArXiv, Vol. abs/2108.07777, 2021.

[234] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., “3d pictorial structures for multiple human pose estimation”, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, str. 1669–1676.

[235] Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C., “Dsac — differentiable ransac for camera localization”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, str. 2492–2500.

[236] Brachmann, E., Rother, C., “Learning less is more - 6d camera localization via 3d surface regression”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, str. 4654–4662.

[237] Brachmann, E., Rother, C., “Neural-guided ransac: Learning where to sample model hypotheses”, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, str. 4321–4330.

[238] Yi, K. M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P., “Learning to find good correspondences”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, str. 2666–2674.

[239] Schulman, J., Heess, N., Weber, T., Abbeel, P., “Gradient estimation using stochastic computation graphs”, in NIPS, 2015.

[240] Jang, E., Gu, S., Poole, B., “Categorical reparameterization with gumbel-softmax”, ArXiv, Vol. abs/1611.01144, 2017.

[241] Maddison, C. J., Mnih, A., Teh, Y., “The concrete distribution: A continuous relaxation of discrete random variables”, ArXiv, Vol. abs/1611.00712, 2017.

[242] Maddison, C. J., Tarlow, D., Minka, T., “A* sampling”, in NIPS, 2014.

[243] Xiang, D., Joo, H., Sheikh, Y., “Monocular total capture: Posing face, body, and hands in the wild”, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 10 957–10 966.

[244] Drover, D., Rohith, M., Chen, C.-H., Agrawal, A., Tyagi, A., Huynh, C. P., “Can 3d pose be learned from 2d projections alone?”, in ECCV Workshops, 2018.

[245] Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W., “Deep kinematics analysis for monocular 3d human pose estimation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[246] Corona, E., Pons-Moll, G., Alenyà, G., Moreno-Noguer, F., “Learned vertex descent: A new direction for 3d human model fitting”, ArXiv, Vol. abs/2205.06254, 2022.

[247] Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M. J., “Collaborative regression of expressive bodies using moderation”, in International Conference on 3D Vision (3DV), 2021.

[248] Xiu, Y., Yang, J., Tzionas, D., Black, M. J., “ICON: Implicit Clothed humans Obtained from Normals”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, str. 13 296–13 306.

[249] Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H., “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization”, in The IEEE International Conference on Computer Vision (ICCV), October 2019.

[250] Saito, S., Simon, T., Saragih, J., Joo, H., “Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization”, in CVPR, 2020.

[251] Zheng, Z., Yu, T., Liu, Y., Dai, Q., “Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, 2022, str. 3170–3184.

[252] Moon, G., Nam, H., Shiratori, T., Lee, K. M., “3d clothed human reconstruction in the wild”, in European Conference on Computer Vision (ECCV), 2022.

[253] Güler, R. A., Neverova, N., Kokkinos, I., “Densepose: Dense human pose estimation in the wild”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, str. 7297–7306.

[254] von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G., “Recovering accurate 3d human pose in the wild using imus and a moving camera”, in European Conference on Computer Vision (ECCV), sep 2018.

[255] Bertiche, H., Madadi, M., Escalera, S., “Cloth3d: Clothed 3d humans”, in European Conference on Computer Vision. Springer, 2020, str. 344–359.

[256] Madadi, M., Bertiche, H., Bouzouita, W., Guyon, I., Escalera, S., “Learning cloth dynamics: 3d + texture garment reconstruction benchmark”, in Proceedings of the NeurIPS 2020 Competition and Demonstration Track, PMLR, Vol. 133, 2021, str. 57–76.

[257] Zhao, K., Wang, S., Zhang, Y., Beeler, T., , Tang, S., “Compositional human-scene interaction synthesis with semantic control”, in European conference on computer vision (ECCV), 2022.

[258] Sun, K., Xiao, B., Liu, D., Wang, J., “Deep high-resolution representation learning for human pose estimation”, in CVPR, 2019.

[259] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., “Deep high-resolution representation learning for visual recognition”, TPAMI, 2019.

[260] Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z., “Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop”, in Proceedings of the IEEE International Conference on Computer Vision, 2021.

[261] Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y., “Pymaf-x: Towards well-aligned full-body model regression from monocular images”, arXiv preprint arXiv:2207.06400, 2022.

[262] Alldieck, T., Magnor, M. A., Bhatnagar, B. L., Theobalt, C., Pons-Moll, G., “Learning to reconstruct people in clothing from a single rgb camera”, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 1175–1186.

[263] Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G., “Detailed human avatars from monocular video”, in International Conference on 3D Vision, Sep 2018, str. 98–109.

[264] Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M., “Tex2shape: Detailed full human body geometry from a single image”, in IEEE International Conference on Computer Vision (ICCV), 2019.

[265] Lazova, V., Insafutdinov, E., Pons-Moll, G., “360-degree textures of people in clothing from a single image”, 2019 International Conference on 3D Vision (3DV), 2019, str. 643–653.

[266] Zhu, H., Zuo, X., Yang, H., Wang, S., Cao, X., Yang, R., “Detailed avatar recovery from single image”, in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.

[267] Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R., “Detailed human shape estimation from a single image by hierarchical mesh deformation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, str. 4491–4500.

[268] Xiang, D., Prada, F., Wu, C., Hodgins, J. K., “MonoClothCap: Towards Temporally Coherent Clothing Capture from Monocular RGB Video”, 2020 International Conference on 3D Vision (3DV), 2020, str. 322-332.

[269] Bhatnagar, B. L., Sminchisescu, C., Theobalt, C., Pons-Moll, G., “Combining implicit function learning and parametric models for 3d human reconstruction”, in European Conference on Computer Vision (ECCV). Springer, August 2020.

[270] Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F., “Smplicit: Topology-aware generative model for clothed people”, in CVPR, 2021.

[271] “3DPeople”, https://3dpeople.com/en/, accessed: 2022-10-22.

[272] “AXYZ”.

[273] “HumanAlloy”, https://humanalloy.com/, accessed: 2022-10-22.

[274] “RenderPeople”, https://renderpeople.com/, accessed: 2022-10-22.

[275] He, K., Zhang, X., Ren, S., Sun, J., “Deep residual learning for image recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, str. 770–778.

[276] Mohlin, D., Sullivan, J., Bianchi, G., “Probabilistic orientation estimation with matrix fisher distributions”, in Advances in Neural Information Processing Systems, Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., (ur.), Vol. 33. Curran Associates, Inc., 2020, str. 4884–4893, dostupno na: https://proceedings.neurips.cc/paper/2020/file/33cc2b872dfe481abef0f61af181dfcf-Paper.pdf

[277] “Agora downloads”, https://agora.is.tue.mpg.de/download.php, accessed: 2022-11-11.

[278] Bhatnagar, B. L., Tiwari, G., Theobalt, C., Pons-Moll, G., “Multi-garment net: Learning to dress 3d people from images”, in IEEE International Conference on Computer Vision (ICCV). IEEE, Oct 2019.

[279] “Chamfer distance Gist”, https://gist.github.com/sergeyprokudin/, accessed: 2022-11-12.

[280] Community, B. O., Blender - a 3D modelling and rendering package, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018, dostupno na: http://www.blender.org

[281] Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J., "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, str. 13 232–13 242.

[282] Gartner, E., Andriluka, M., Coumans, E., Sminchisescu, C., "Differentiable dynamics for articulated 3d human motion reconstruction", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, str. 13 180–13 190.

[283] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z., "3d human pose estimation with spatial and temporal transformers", Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[284] Hassanin, M., Khamiss, A., Bennamoun, Boussaid, F., Radwan, I., "Crossformer: Cross spatio-temporal transformer for 3d human pose estimation", ArXiv, Vol. abs/2203.13387, 2022.

[285] Reddy, N. D., Guigues, L., Pischulini, L., Eledath, J., Narasimhan, S., "Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking", in Proceedings of (CVPR) Computer Vision and Pattern Recognition. IEEE, June 2021.

[286] Bartol, K., Bojanić, D., Petković, T., Pribanić, T., "Catadioptric stereo on a smartphone", in 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA), 2021, str. 189–194.

[287] Wachowiak, M., Karas, B., "3d scanning and replication for museum and cultural heritage applications", Journal of the American Institute for Conservation, Vol. 48, 08 2009, str. 141–158.

[288] Mur-Artal, R., Tardós, J. D., "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras", IEEE Transactions on Robotics, Vol. 33, 2017, str. 1255–1262.

[289] Masiero, A., Fissore, F., Piragnolo, M., Guarnieri, A., Pirotti, F., Vettore, A., "Initial evaluation of 3d reconstruction of close objects with smartphone stereo vision", ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLII-1, 09 2018, str. 289–293.

[290] Pribanić, T., Petković, T., Đonlić, M., "3d registration based on the direction sensor measurements", Pattern Recognition, Vol. 88, 2019, str. 532–546, dostupno na: https://www.sciencedirect.com/science/article/pii/S0031320318304291

[291] Pribanić, T., Petković, T., Đonlić, M., Angladon, V., Gasparini, S., "3d structured light scanner on the smartphone", in Image Analysis and Recognition, Campilho, A., Karray, F., (ur.). Cham: Springer International Publishing, 2016, str. 443–450.

[292] Gluckman, J., Nayar, S., "Rectified catadioptric stereo sensors", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, 2002, str. 224–236.

[293] Goshtasby, A., Gruver, W. A., "Design of a single-lens stereo camera system", Pattern Recognition, Vol. 26, No. 6, 1993, str. 923–937, dostupno na: https://www.sciencedirect.com/science/article/pii/0031320393900585

[294] Inaba, M., Hara, T., Inoue, H., "A stereo viewer based on a single camera with view-control mechanisms", in Proceedings of 1993 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '93), Vol. 3, 1993, str. 1857–1865 vol.3.

[295] Southwell, D., Basu, A., Fiala, M., Reyda, J., "Panoramic stereo", in Proceedings of 13th International Conference on Pattern Recognition, Vol. 1, 1996, str. 378–382 vol.1.

[296] Nene, S., Nayar, S., "Stereo with mirrors", in Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), 1998, str. 1087–1094.

[297] Gluckman, J., Nayar, S., "Planar catadioptric stereo: geometry and calibration", in Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Vol. 1, 1999, str. 22–28 Vol. 1.

[298] Mouaddib, E., Sagawa, R., Echigo, T., Yagi, Y., "Stereovision with a single camera and multiple mirrors", in Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005, str. 800–805.

[299] Somanath, G., Rohith, M., Kambhamettu, C., "Single camera stereo system using prism and mirrors", in ISVC, 2010.

[300] Lee, D. H., Kweon, I. S., Cipolla, R., "A biprism-stereo camera system", in Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Vol. 1, 1999, str. 87 Vol. 1.

[301] Chaen, A., Yamazawa, K., Yokoya, N., Takemura, H., "Acquisition of three-dimensional information using omnidirectional stereo vision", in Computer Vision — ACCV'98,

Chin, R., Pong, T.-C., (ur.). Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, str. 288–295.

[302] Kang, S. B., "Catadioptric self-calibration", in Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), Vol. 1, 2000, str. 201–207 vol.1.

[303] Simanek, D., "Mirror and prism methods for 3d macro photography", dostupno na: https://www.lockhaven.edu/~dsimanek/3d/stereo/3dgallery16.htm

[304] Zhang, Z., "A flexible new technique for camera calibration", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, 2000, str. 1330–1334.

[305] Wang, J., Zhang, C., Zhu, W., Zhang, Z., Xiong, Z., Chou, P., "3d scene reconstruction by multiple structured-light based commodity depth cameras", 03 2012, str. 5429–5432.

[306] Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B. C., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., "Panoptic studio: A massively multiview system for social interaction capture", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, 2019, str. 190–204.

[307] "Revopoint tanso s1", https://www.revopoint3d.com/portable-3d-scanner-tanso-s1/, accessed: 2021-03-16.

[308] Saint, A., Ahmed, E., Shabayek, A. E. R., Cherenkova, K., Gusev, G., Aouada, D., Ottersten, B., "3dbodytex: Textured 3d body dataset", in 2018 International Conference on 3D Vision (3DV), 2018, str. 495–504.

[309] Park, J., Zhou, Q., Koltun, V., "Colored point cloud registration revisited", in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, str. 143–152.

Biography

Kristijan Bartol received the B.Sc. degree in computing, in 2016, and the M.Sc. degree in computer science, in 2019. He is currently pursuing the Ph.D. degree in computing with the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His current research interests include computer vision, human pose and shape estimation, and deep learning. He is the co-author of two scientific papers published in international journals of Q1 and Q2 categories and the co-author of a total of twelve papers published in conference proceedings with international review. Several papers are presented at the most important conferences in the field of computer vision (namely CVPR, BMVC, and ICPR) where the CVPR work was selected for an oral presentation at the main conference in New Orleans in 2022. As part of his doctoral research, he attended scientific training at INRIA Institute in Grenoble, in the MORPHEO group, and at the Technical University in Dresden, at the Chair of Computer Graphics and Visualization.

List of Journal Papers

1. Bartol, Kristijan; Bojanić, David; Petković, Tomislav; Pribanić, Tomislav. A review of body measurement using 3D scanning. IEEE Access. 2021. str. 67281-67301, doi: 10.1109/ACCESS.2021.3076595
2. Bartol, Kristijan; Bojanić, David; Petković, Tomislav; Peharec, Stanislav; Pribanić, Tomislav. Linear Regression vs. Deep Learning: A Simple Yet Effective Baseline for Human Body Measurement. Sensors 2022, 22, doi: 10.3390/s22051885

List of Conference Papers

1. Jertec, Andrej; Bojanić, David; Bartol, Kristijan; Pribanić, Tomislav; Petković, Tomislav; Petrak, Slavenka. On using PointNet Architecture for Human Body Segmentation. 11th International Symposium on Image and Signal Processing and Analysis (ISPA). 2019, pp. 253-257 doi:10.1109/ISPA.2019.8868844
2. Bojanić, David; Bartol, Kristijan; Pribanić, Tomislav; Petković, Tomislav; Diez Donoso, Yago; Salvi Mas, Joaquim. On the Comparison of Classic and Deep Keypoint Detector

and Descriptor Methods. 11th International Symposium on Image and Signal Processing and Analysis (ISPA). 2019, pp. 64-69 doi:10.1109/ISPA.2019.8868792

3. Pribanić, Tomislav; Bojanić, David; Bartol, Kristijan; Petković, Tomislav. Can OpenPose Be Used as a 3D Registration Method for 3D Scans of Cultural Heritage. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12667. Springer International Publishing, 2021, pp. 83-96, doi:10.1007/978-3-030-68787-8_6
4. Pribanić, Tomislav; Petković, Tomislav; Bojanić, David; Bartol, Kristijan; Gupta, Mohit. Smart Time-Multiplexing of Quads Solves the Multicamera Interference Problem. International Conference on 3D Vision (3DV). Conference Publishing Services, 2020, str. 811-819 doi:10.1109/3DV50981.2020.00091
5. Bartol, Kristijan; Bojanić, David; Petković, Tomislav; Pribanić, Tomislav; Donoso, Yago. Towards Keypoint Guided Self-Supervised Depth Estimation. Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - (Volume 4). SCITEPRESS - Science and Technology Publications, 2020, pp. 583-589, doi:10.5220/0009190005830589
6. Bojanić, David; Bartol, Kristijan; Petković, Tomislav; Pribanić, Tomislav. A Review of Rigid 3d Registration Methods. Book of Proceedings of 13th International Scientific – Professional Symposium TEXTILE SCIENCE & ECONOMY. 2020, pp. 286-296.
7. Bartol, Kristijan; Bojanić, David; Petković, Tomislav; D'Apuzzo, Nicola; Pribanić, Tomislav. A Review of 3D Human Pose Estimation from 2D Images. Proceedings of 3DBODY.TECH 2020 11th International Conference and Exhibition on 3D Body Scanning and Processing Technologies. 2020, doi:10.15221/20.29
8. Bojanić, David; Bartol, Kristijan; Petković, Tomislav; D'Apuzzo, Nicola; Pribanić, Tomislav. Evaluation of 3D Registration Deep Learning Methods using Iterative Transformation Estimations. Proceedings of 3DBODY.TECH 2020 11th International Conference and Exhibition on 3D Body Scanning and Processing Technologies. 2020, 31, 10 doi:10.15221/20.31
9. Pribanić, Tomislav; Petković, Tomislav; Bojanić, David; Bartol, Kristijan; Gupta, Mohit. Scene Adaptive Structured Light 3D Imaging. Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Volume 4). SCITEPRESS - Science and Technology Publications, 2020, pp. 576-582, doi:10.5220/0009189905760582
10. Bartol, Kristijan; Bojanić, David; Petković, Tomislav; Pribanić, Tomislav. Catadioptric Stereo on a Smartphone. 12th International Symposium on Image and Signal Processing and Analysis (ISPA). 2021, pp. 189-194, doi: 10.1109/ISPA52656.2021.9552146
11. Bartol, Kristijan; Bojanić, David; Petković, Tomislav; Pribanić, Tomislav. Generaliz-

able Human Pose Triangulation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022., pp. 11018-11027

12. Bojanić, David; Bartol, Kristijan; Forest, Josep; Gumhold, Stefan; Petković, Tomislav; Pribanić, Tomislav. Challenging the Universal Representation of Deep Models for 3D Point Cloud Registration. The 33rd British Machine Vision Conference, BMVC (Workshops), 2022

Životopis

Kristijan Bartol je 2016. godine dobio diplomu inženjera računarstva, a diplomu magistra računarstva dobio je u 2019. godini. Trenutno pohađa doktorski studij na Sveuličilištu u Zagrebu, na Fakultetu Elektrotehnike i računarstva. Njegovi trenutni istraživački interesi uključuju računalni vid, procjenu držanja i oblik tijela te duboko učenje. Suautor je dva znanstvena rada objavljena u međunarodnim časopisima Q1 i Q2 kategorije. Suautor je ukupno dvanaest radova objavljenih u zbornicima konferencija s međunarodnom recenzijom. Nekoliko radova objavljeno je na najvažnijim konferencijama u području računalnog vida (na CVPR-u , BMVC-u i ICPR-u) od kojih je rad na CVPR-u odabran za usmenu prezentaciju na glavnoj konferenciji u New Orleansu 2022. U sklopu doktorskog istraživanja, bio je na znanstvenom usavršavanju na INRIA institutu u Grenoblu, u grupi MORPHEO te na Tehničkom sveučilištu u Dresdenu, na katedri za računalnu grafiku i vizualizaciju.