ESTIMATING SHAPE AND POSE FROM IMAGES

Dissertation zur Erlangung des Doktorgrades der
Technischen Fakulät der
Albert-Ludwigs-Universität Freiburg im Breisgau

vorgelegt von

CHRISTIAN ZIMMERMANN

DEKAN:
Prof. Dr. Rolf Backofen

ERSTGUTACHTER UND BETREUER:
Prof. Dr. Thomas Brox
Albert-Ludwigs-Universität Freiburg

ZWEITGUTACHTER:
Prof. Dr. Otmar Hilliges
Eidgenössische Technische Hochschule Zürich

DATUM DER MÜNDLICHEN PRÜFUNG:
18.12.2020

ZUSAMMENFASSUNG

Diese Arbeit beschäftigt sich mit der Schätzung von Pose und Gestalt von artikulierten Objekten. Dieses Forschungsfeld hat in den letzten Jahren einen Wandel vollzogen: Es hat sich von modelbasierten Trackingverfahren hin zu diskriminative Methoden entwickelt. Diese basieren meist auf tiefen neuronalen Netzen und erlauben die Schätzung von Pose und Gestalt basierend auf einem einzigen Bild. Mit diesem methodischen Wechsel rücken neue Herausforderung in den Vordergrund, wie die Notwendigkeit von großen Mengen an annotierten Daten, um die diskriminativen Ansätze trainieren zu können. Diese Arbeit beschäftigt sich mit den Herausforderungen und entwickelt Strategien, um Trainingsdaten zu gewinnen und bestehende Daten für weitere Aufgaben nutzbar zu machen.

Zunächst wird die Anwendbarkeit von künstlich generierten Datensätzen geprüft, um das Problem der Hand Posen Schätzung von einem einzelnen Farbbild zu lernen. Hierzu wird eine tiefe Netzarchitektur entwickelt und auf dem erzeugten Datensatz trainiert. Der resultierende Algorithmus erlaubt es 3D Hand Pose von einem einzelnen Farbbild zu bestimmen und dadurch den Stand der Technik in der Detektion von Handzeichensprache zu verbessern.

Nachfolgend wird ein Ansatz vorgestellt, der es erlaubt Körperpose von Menschen anhand von RGBD Bildern zu schätzen. Dieser baut auf bestehenden RGB Datensätzen auf und benötigt daher nur eine geringe Menge annotierter RGBD Beispiele. Der entwickelte Ansatz wird dazu genutzt einem Roboter neue Aufgaben beizubringen, indem er einen menschlichen Lehrer während der Demonstration beobachtet und anschließend imitiert.

Als nächstes wird eine Methode vorgestellt, die es erlaubt Pose von Tieren unter Zuhilfenahme von mehreren Kameras zu bestimmen. Die Methode verfolgt einen ganzheitlichen Ansatz aller Kamera, was den vorgestellten Ansatz robuster und exakter arbeiten lässt als bisherige Methoden. Anwendung findet der Algorithmus im Rahmen biologischer Experimente mit Versuchstieren. Hierbei ermöglicht er deren Bewegungsschätzung und erlaubt es die Wirkung von optogenetischer Stimulation zu quantisieren.

Letztlich wird die vorige Methode um einen Ansatz zur Schätzung der Handgestalt erweitert. Hierbei wird ein parametrisches Handmodell verwendet, dass unter Zuhilfenahme von Posen- und Segmentierungsschätzung positioniert wird. Dies erlaubt es einen großen Datensatz mit echten RGB Bildern und zugehörigen Hand Annotation zu erstellen, in dessen Erstellung manueller Aufwand nur in geringem Umfang einfließen muss.

ABSTRACT

This thesis is set in the field of pose and shape estimation of articulated objects from image observations. With the recent shift in paradigm towards deep learning also the methods for pose estimation evolved from optimizing object models in a tracking fashion, towards powerful discriminative algorithms that make frame independent estimation possible. This directional change has introduced new challenges, such as the need for good training data to supervise deep learning methods. This thesis tackles some of the challenges and proposes ways to provide training data for discriminative methods or to extend existing data sources towards new settings.

First, the use of synthetic data is explored in the scope of hand pose estimation. An architecture for 3D hand pose estimation from a single image is proposed that, trained on the synthetic dataset, achieved state of the art in pose estimation and allowed surpassing previous approaches on sign language recognition.

Second, existing RGB datasets are leveraged to develop an approach that estimates metrically correct human pose from RGBD inputs, with only minimal need of labeled RGBD data. The approach outperformed comparable approaches and allowed teaching a robot new tasks from few human demonstrations.

Third, a new approach for estimation of 3D pose incorporating multiple camera views in a holistic fashion is presented. It is applied in a biological setting for motion capture of laboratory animals. More accurate and robust results are obtained than by using non-holistic prediction methods. Our approach can learn the task at hand from fewer labeled samples than state of the art methods.

Lastly, the multi-view pose estimation algorithm is adapted to hand pose and extended by a model-based fitting procedure that yields shape fits. This allowed to create a real-world dataset of RGB images with corresponding hand shape fits, which is an important mile stone for training and evaluating single view methods.

Dedicated to the most loving, supporting and proud parents one can imagine and any child can hope for.

for Rainer and Christa

ACKNOWLEDGMENTS

First of all, I'd like to thank my supervisor Thomas Brox, who gave me the opportunity to work in his lab. He's a great instructor for scientific work, has an amazing ability to find rewarding research fields and was a great partner to discuss and plan research with.

I feel fortunate to get in touch with numerous inspiring collaborators during my time as a Ph.D. student. Especially, I thank Tim Welschehold and Artur Schneider for their great partnership during our interdisciplinary research endeavors, and the great people from Adobe Research in San Jose, in particular Duygu Ceylan, for many fruitful inputs and practical guidance provided over the time of our cooperation. Furthermore, I'm grateful for the amazing work of secretaries and technicians at the universities' chair. You guys were always able to deal with any type of bureaucratic or computational issue, which enormously contributed to the work presented here by providing an environment where research is possible. I want to express my gratitude to all fellow Ph.D. candidates at my chair for the countless hallway, lunch and water cooler discussions that steered my research into the direction what culminated in this work. My special thanks in this regard goes to Nikolaus Mayer, Benjamin Ummenhofer, Max Argus, Philipp Schröppel and Maxim Tatarchenko for their unbreakable availability to discuss immature thoughts and directions as well as Nikolaus Mayer, Philipp Schröppel, Max Argus and Özgün Çiçek for proofreading the manuscripts that evolved into this thesis. For the continuous financial support through two research projects I'd like to say thank you to the Baden-Württemberg Stiftung for making my research possible.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications and form the basis for this thesis.

[1] Christian Zimmermann and Thomas Brox. "Learning to Estimate 3D Hand Pose from Single RGB Images." In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 4913–4921. DOI: 10.1109/ICCV.2017.525. URL: https://doi.org/10.1109/ICCV.2017.525.

[2] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. "3D Human Pose Estimation in RGBD Images for Robotic Task Learning." In: *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*. IEEE, 2018, pp. 1986–1992. DOI: 10.1109/ICRA.2018.8462833. URL: https://doi.org/10.1109/ICRA.2018.8462833.

[3] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max J. Argus, and Thomas Brox. "FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images." In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 813–822. DOI: 10.1109/ICCV.2019.00090. URL: https://doi.org/10.1109/ICCV.2019.00090.

[4] Christian Zimmermann, Artur Schneider, Mansour Alyahyay, Thomas S Brox, and Ilka Diester. "FreiPose: A Deep Learning Framework for Precise Animal Motion Capture in 3D Spaces." In: *bioRxiv* (2020).

CONTENTS

ACRONYMS

AI Artificial Intelligence

AR Augmented Reality

CNN Convolutional Neural Network

DLC DeepLabCut, Approach for 2D pose estimation by [75]

DoF Degrees of Freedom

EPE End Point error

RGB Red Green Blue, synonymous to colored images

RGBD Red Green Blue Depth, RGB plus depth information per pixel

PCK Percentage of Correct Keypoints

Part I

PROLOGUE

This thesis starts with a gentle introduction to the problems tackled in this work. The contributions of this work are listed and localized within the scientific spectrum. Important properties of the image formation process are recapitulated to show under which circumstances 3D reconstruction is possible. Lastly, common solutions for pose and shape estimation are presented with a broad scope.

1 INTRODUCTION

AI needs to understand the environment

Perceiving and understanding the surrounding world is natural for humans, but turns out to be one of the great challenges artificial intelligence is facing. With some type of sensor it is comparatively easy to turn the environment of a robot into computer-processable data. In this representation it is possible to evaluate whether a recorded value is larger than some threshold, but this is not relevant high level information that allows a robot to operate autonomously in loosely constrained environments. Relevant points in the context of robots interacting with humans could be:

- Is there a human in front of me? Far away or close to me?
- Where is the human located in the world? Which way is he facing?
- Where are objects of interest in the world? How is the human interacting with them?
- What is the human doing?

Without being able to answer high level questions like these it is impossible for artificial agents to act intelligently in the real world or to interact with humans in a natural way.

The unconstrained case is hard

The high level of abstraction in the aforementioned questions makes them hard, because information of individual pixels alone is not sufficient to answer them. The image needs to processed holistically, and possibly multiple frames need to be taken into account. This makes manually designing robust algorithms for these purposes hard, and even more difficult, the broader the scope of applicability is supposed to be. Solving high level problems algorithmically in a narrow laboratory setting might be possible, but engineering a general algorithm to work for a diverse range of settings is virtually impossible.

Machine learning and large datasets

This is exactly where machine learning has enabled large progress for the field of computer vision. Instead of manually designing algorithms to answer these questions, procedures were developed that allow machines to learn algorithms on their own. This change of paradigm from handcrafted towards learned features created the need for large, diverse datasets with annotation for the task in question. Given such a dataset, learned approaches achieve unprecedented accuracy in benchmarks for most classic computer vision tasks, including classification [43], segmentation [16], optical flow [49], and human pose estimation [14].

Acquisition of labeled data

Problems arise in fields where such datasets do not exist and data

acquisition is difficult. It is possible to ask human annotators to tell cat and non-cat depicting images apart or to draw boxes around all car objects. These problems are solvable through crowd sourcing, but there are many tasks where this becomes infeasible. Examples include dense prediction tasks like optical flow or abstract task like pose estimation that are difficult, take much time or expertise to annotate, and therefore yield erroneous annotations.

Importance of pose estimation

Understanding pose of articulated objects from images is essential information for many applications including self-driving cars, sign-language, action recognition, activity monitoring, sport analytics, human-computer interfaces, and virtual or augmented reality. In these cases measuring motion of the human body, or an animal body, either provides an expressive intermediate representation or is directly of interest.

Focus of this thesis

This thesis explores different methods for the acquisition of training data for tasks that are difficult to label, namely pose and shape estimation. Chapter 3 explores the use of synthetic data to learn 3D pose estimation of human hands from a single color image. Chapter 4 shows how to leverage existing RGB datasets for human pose estimation such that a small scale RGBD dataset is sufficient to create an algorithm that estimates 3D human pose from RGBD. Chapter 5 proposes a method for 3D pose estimation based on a multi-view camera setup. It serves as a motion capture system for laboratory animals and allows quantification of animal motion during biological experiments. Building on top of this motion capture method Chapter 6 deploys differential rendering for hand shape annotation of multi-view RGB recordings. This method is then used to create a dataset for training single-view hand shape estimation algorithms.

1.1 CONTRIBUTIONS

This thesis contributes to the state of art in the following ways:

- A large RGB dataset for hand pose estimation was created that allows training of deep neural networks and induced a noteworthy amount of follow-up research in this area (Chapter 3).
- The first CNN-based formulation using a clear distinction between extraction of 2D pose and a learned prior for 2D to 3D lifting for hand pose estimation was proposed (Chapter 3).
- A human pose estimation approach based on RGBD input with state-of-the-art accuracy that enabled robotic learning-from-demonstration was developed (Chapter 4).
- A novel multi-view CNN architecture incorporating camera geometry for learning pose estimation was presented (Chapter 5).

- An estimation procedure of hand shape from multi-view imagery was developed (Chapter 6).
- The first large-scale real-world dataset with shape annotation that allows to train monocular hand shape estimation methods was created (Chapter 6).

1.2 OVERVIEW OF PROBLEMS

Overview of problems

This thesis addresses multiple tasks, which include:

- Detection: Where is the object of interest? E. g., bounding box detection.
- Classification: What is the object of interest doing? E. g., sign or action recognition.
- Pose Estimation: Whats the location, orientation and articulation of the object? E. g., human body pose estimation
- Shape Estimation: What does the surface of the object look like? E. g., hand shape estimation

It is immediately clear that some problems from this list are harder to solve than others and that some problems are extensions of others. For example, if the hand shape is known it is trivial to extract pose information from a posed shape model or if the pose is known then the detection problem is also solved. Furthermore, gesture classification tasks become much simpler once pose is estimated.

Problems tackled in this thesis

Most of this work deals with pose estimation in a 3D setting (Chapter 3, Chapter 4, Chapter 5). Chapter 6 tackles shape estimation and the other tasks are covered as preliminary tasks. Detection is needed in Chapter 5, and Chapter 3 uses gesture classification as an application.

1.3 CHALLENGES

There are several reasons why estimating pose or shape from images is difficult. Occlusion is a big challenge, not only can the object of interest be occluded by something else, but the object will also occlude itself. For occluded parts no information can be obtained from the image. This requires to hallucinate the missing parts needing prior knowledge. Occlusion not only makes inference of the occluded parts harder, but also complicates acquiring annotation for these cases. This can create a strong bias in manually labeled datasets that can not deal with occlusions appropriately.

Occlusion

Ambiguity

Secondly, the solution is often ambiguous. There is an ambiguity that arises from the missing depth information, e. g., is it a small

object close-by or a larger one further away. Another ambiguity can come from the appearance of objects; For example fingers of a human hand can be difficult to tell apart from each other, because they look very similar.

Variation

Different sources of variation make it difficult to develop robust approaches. One source is appearance which can be caused by clothing, lighting, shadowing, skin color, viewpoint or the camera used for recording. Additionally, there can be a lot of scale variation between subjects, i. e., large and small people to estimate pose of, which is challenging for 3D estimation methods. Ideally, an algorithm performs equally well throughout all settings, but this is difficult to achieve.

High dimensionality

Lastly, the problem is hard because estimating pose is of high dimensionality. Estimating pose of a rigid object is already considered to be a difficult problem, which means determining 6 parameters that describe rotation and translation. Whereas, it is common to use hand skeletons with 27 *DoF*, inspired by biomechanical constraints [69], which makes solving the problem a lot harder, because more parameters have to be estimated. Additionally, optimization methods tend to get less effective when applied in higher dimensional spaces.

1.4 ROLE OF DATA

CNNs depend on data

Data has always played an important role for computer vision, but its use has shifted from being a pure test bench for hand-crafted algorithms towards deriving large amounts of the algorithm from data. In current deep learning approaches most of the algorithm is learned from data given a random initialization. This makes that class of algorithms depending heavily on the data used for training. It was shown that deep learning based approaches are very good at identifying shortcuts and peculiarities of the data, that help them with the task on the training set, but this does not guarantee any generalization [143]. So careful compilation of the training data at a sufficiently large scale and in an unbiased way gains importance [77].

Ways to generate training data

When data is needed for supervised training of machine learning algorithms there are two major approaches. One possibility is to record data in the real world and add the annotation in a post processing step. The other one is to use computer graphics to render images that mimic the real world. In this case the annotation is available almost for free, because rendered scenes consist of graphical models in a known geometric configuration. This thesis covers both approaches and each one has its up- and downsides that are discussed now.

Real-world data

When recordings of the real world are used many types of priors and characteristics are fulfilled automatically. For example, the images will contain acquisition artifacts that are typical for real cameras. When imagining humans their poses and interaction between subjects is realistic per definition. In the case of hand pose estimation the

grasps are plausible and anatomically feasible. Realistic hand-object-interaction is achieved effortless. On the other hand real recordings usually contain little variation, because it costs time and effort to introduce many subjects, diverse clothing, different skin color or ethnicity. Additional limiting factors of variation are the viewpoints covered, the background scene and the types of cameras used. These factors might introduce a strong bias towards what a dataset contains. This can get amplified by further limitations that arise from the annotation method deployed. The method may show systematic failure modes for certain configuration, which then get discarded during manual validation and result in "blind spots" of the respective data pool. When annotating real datasets one attempts to automate as many steps of the creation pool as possible, but commonly manual effort can not be eliminated completely which makes real datasets hard to scale.

Synthetic data

On the other hand, introducing much variation or scaling up the dataset in terms of samples is effortless, when computer graphical models and rendering pipelines are used. The synthetic or rendered datasets are trivially scaled up by introducing randomness in every run. This means to sample lighting, appearance, viewpoint and motion repeatedly, and the amount of dataset samples can virtually be indefinite. This is also where the crux lies: One needs to sample valid configurations or parameters to perform a rendering run. While it still might be fine to engineer distributions for lighting and viewpoint, this is not trivial for pose or motion because complex correlations exist between parts of articulated objects. The case where two hands interact with each other, or a hand with an object, are settings where it is definitely impossible to sample a valid configuration from an engineered distribution representing the complete space of feasible poses. The generation of data is usually imperfect such that there is a remaining gap between data from the simulation and real domain. An alternative to sampling from a continuous distribution is to extract a valid configuration from real data and transfer it to the simulation, i. e., sample from a discrete set of samples. This allows to augment the given set of samples, which is frequently performed in fields where a large corpus of real motion capture data is available. For this kind of transfer the simulated data is still limited to some extent by the real data, but variation of appearance, viewpoint and lighting can be increased through the additional samples.

2 PRELIMINARIES

Before delving deep into the contribution of this thesis, some general remarks about the class of problems will be made. This allows to locate this work in the research context and give an understanding how common approaches operate.

2.1 ESTIMATING 3D GEOMETRY FROM 2D OBSERVATIONS

Focus lies on RGB cameras

This thesis is set in the fairly common setting of an RGB matrix camera that is observing the surrounding world. Its prevalence arises from the low cost of this camera type in conjunction with the easy-to-understand acquisition result for humans. Color cameras are widely spread to mobile phones, cars, webcams and robotic systems.

Projective loss of information

For matrix cameras the image formation process can approximately be modeled using the pinhole camera model. The underlying idea is that light from the environment passes through a small opening on one side of the camera and hits the planar sensor on the other side. The light is absorbed and accumulated by the sensor over some period of time to form the image. Figure 2.1 shows the mathematically equivalent scenario, where the image plane is mirrored with respect to the pinhole. Then the image plane is depicted in front of the camera.

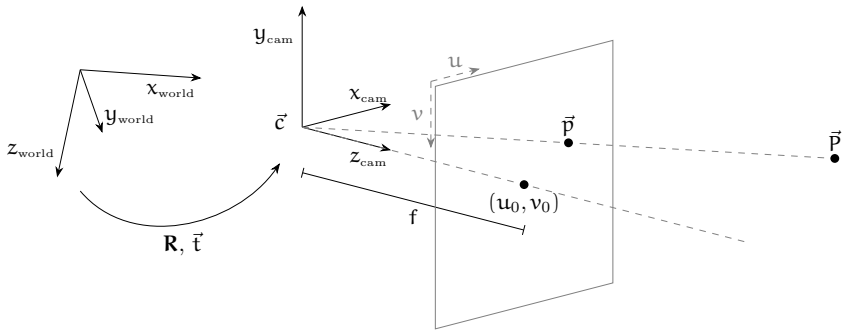

Figure 2.1: **Pinhole camera model.** A world point $\vec{P}$ is projected towards the optical center $\vec{c}$ and intersects the image plane at $\vec{p}$. The resulting image is influenced by the focal length f which is illustrated by the distance between image plane and optical center. Practically, the focal length describes the scale between real-world and image coordinates. One can see that the location of $\vec{P}$ on the ray is irrelevant for the resulting projection point $\vec{p}$ on the image plane; a projective ambiguity arises when attempting to reverse the imaging process.

Problem is ill-posed

A point $\vec{p}$ on the image plane accumulates information along the ray that connects $\vec{p}$ with the optical center $\vec{c}$. The exact location on the ray where information originated from is lost during projection. This turns estimating 3D entities from a single 2D observations into an under-constrained problem that can only be solved using additional information.

Camera calibration

Mathematically, the transformation between a 3D world point $\vec{P}$ and the resulting point in image coordinates $\vec{p}$ is described by the pinhole camera model

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \vec{p} = \mathbf{K} \cdot \vec{P} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} X_{\text{cam}} \\ Y_{\text{cam}} \\ Z_{\text{cam}} \end{pmatrix} \tag{2.1}$$

using the camera intrinsic $\mathbf{K} \in \mathbb{R}^{3\times3}$ matrix that depends on the focal length f and location of the camera's central point in image space. The ray originating from the camera central point $\vec{c}$, which intersects the image plane perpendicularly at $\left(u_0, v_0\right)$ is called the optical axis. Equation 2.1 takes coordinates in a camera centered 3D coordinate frame and calculates their location in image space. The transformation

$$\begin{pmatrix} X_{\text{cam}} \\ Y_{\text{cam}} \\ Z_{\text{cam}} \end{pmatrix} = \mathbf{R} \cdot \begin{pmatrix} X_{\text{world}} \\ Y_{\text{world}} \\ Z_{\text{world}} \end{pmatrix} + \vec{t} \tag{2.2}$$

describes the rotation and translation between camera and world coordinates. This allows the camera to move freely with respect to a world coordinate frame. For a given camera setting finding $\mathbf{K}$ is referred to as intrinsic calibration and determining $\mathbf{R}$ and $\vec{t}$ is called extrinsic camera calibration. If both are known it is possible to calculate where an arbitrary world point is being projected to in the image frame.

Common assumptions to resolve ambiguity

To calculate the 3D information from their 2D projections additional constraints are needed. Frequently, multiple observations are used, i. e., multiple cameras observe the same 3D point from different locations. Unprojecting the 2D image observations into 3D space yields rays originating from the respective camera centers and leads to a point of minimal distance between the rays. Other information that is commonly used is knowledge about the geometry of the object of interest. In this setting correspondences between 2D observations and 3D object points are found and the objects' pose is calculated. This setting is usually referred to as the *Perspective-n-Point* problem and closed form solutions exist [29, 68]. Note that these approaches assume rigidness of the object, i. e., 3D object points do not move with respect to each other.

Special case articulated objects

In practice many objects of interest do not fulfill this assumption,

e. g., humans, hands or animals can only approximately been seen as part-wise rigid objects with additional parameters describing articulation between parts. This mixture of rigid parts with some well-defined degrees of freedom between them is referred to as articulated objects. In this setting additional parameters describing the motion within the object need to be estimated. The number of DoF is increased by every additional part in the chain, which makes solving all degrees of freedom from a sufficiently large amount of correspondences impractical.

2.2 BROAD CATEGORIZATION OF APPROACHES

Different approaches to pose and shape estimation can be categorized in the following ways.

Dimensionality

A very important separation between approaches is the type of representation being estimated. One family of approaches seeks to estimate 3D structure, which could directly be locations in 3D [90, 123] or parameters of a 3D model [10, 56]. Another one are approaches that aim to predict some quantity in image space, e. g., estimation of points in 2D [14]. While 3D problems can be seen as a reconstruction task, estimating 2D is more similar to discriminative pixel-wise estimation task. This thesis deals with estimating 3D structure and uses 2D tasks only as intermediate representation.

From a reconstruction point of view

The 3D problems can be further split up by a reconstruction point of view. Whereas pose estimation is similar to a sparse reconstruction task, shape estimation is a dense reconstruction task. Historically, reconstructing densely used to be more popular [28, 93], because formulating local optimization objectives between a shape model and observed segmentation masks or depth maps worked well given a good initialization. Using the shape model allows to model effects like interpenetration [41, 129] or reason about the physical plausibility of the estimate [129]. More recently, the sparse reconstruction paradigm became more popular, because with the rise of CNNs powerful and robust keypoint detectors became available [112, 133]. These allow to estimate much more abstract concepts, like joint locations, from a single image. This was impossible before and alleviates the need to use explicit models and provide a known initialization to start tracking from.

Input modality

Another commonly made distinction is the input modality. Most works focus on color images (RGB) or depth maps (D), while less work is done using both modalities (RGBD) and little work is available dealing with more exotic variants like infrared or x-ray images. The most important distinction is if depth information is available or not. Availability of depth measurement eliminates the projective ambiguity and allows to use different types of representations (e. g.,

point clouds) compared to approaches that operate on some sort of gridded 2D sensory data.

Number of cameras

The number of camera views available is another important factor. The minimal case of a single camera is the most common one [51, 88, 95], but it raises the need for approaches that can deal well with occlusions and if no depth is available the scale ambiguity has to be accounted for. If multiple cameras come into play occlusion tends to become less of an issue, but usually applications that fall into this regime are limited to controlled studio settings [112, 115], because temporal synchronization and camera calibration between data sources is necessary. Also less prior knowledge is needed for reconstruction if multiple cameras are available.

Type of approach

Nearest neighbor/ Search-based: Pose estimation is formulated as a retrieval problem, where the closest pose with respect to a large dataset of poses is extracted [5, 101]. This creates a trade-off between accuracy and speed. If the dataset is rather small this will result in a poor pose estimation accuracy. On the other hand a large dataset makes the retrieval problem harder because more comparisons have to be made. Commonly, some sort of descriptor (HoG, skin color, edges, or others) is extracted from the image and matched against the database. Most works in this area either optimize on the matching [48] or descriptor side [18, 109] of this approach.

Model based: There is a model of the 3D entity that is to be represented. In this case estimating pose means finding deformation parameters to align the model with image observations. Core parts of this type of algorithm are: A model, an initial model state, a similarity function between observation and model state, and a optimization procedure to update the model state according to the similarity function. Usually, these kinds of approaches need a good initialization which induced a lot of work on this problem [107, 116, 119]. Furthermore, a large corpus of work deals with different choices on models, which use some combination of geometric primitives, e. g., cylinders, spheres, ellipsoids and cones, to model the articulated object: [66, 93, 116]. The number of primitives is usually kept low, because during optimization iteration through the set of shapes is performed many times, which makes triangle meshes with a large number of faces impractical. These kinds of approaches are also referred to as top-down approaches and are usually optimization driven.

Discriminative: Most recently proposed approaches fall into this category. They use an algorithm to either learn a direct mapping to pose space [126] or to detect known parts of the object within the image, e. g., limbs [94] or joints [133]. In the latter case, pose is assembled from detected parts in a bottom-up fashion [14, 64].

Part II

MAIN PART

The following sections describe the contributions this thesis makes. Each chapter is based on one publication, which is stated at the beginning of each section, as well as the outlining the contributions made by the author. Material presented there is taken from the respective paper and the concluding *Follow-up work* chapters discuss the impact of the paper on their respective research fields.

3

ESTIMATING HAND POSE FROM SINGLE RGB

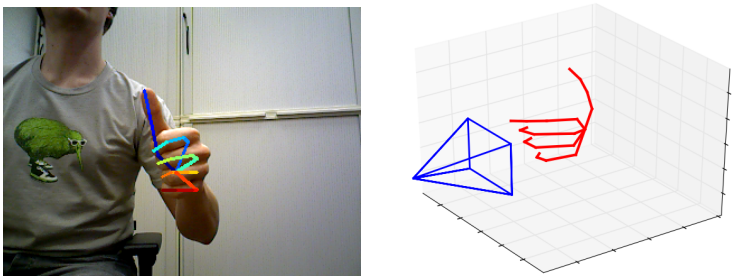

Figure 3.1: **Problem overview.** Given a color image we detect keypoints in 2D (shown overlayed) and learn a prior that allows us to estimate a normalized 3D hand pose.

This chapter describes ideas and experiments that were previously presented in the following work; therefore, copyright lies with © 2017 IEEE.

Learning to Estimate 3D Hand Pose from Single RGB Images
Christian Zimmermann and Thomas Brox
IEEE International Conference on Computer Vision (ICCV), 2017

> *This work introduced neural networks for estimation of 3D hand pose from color images, along with a synthetic dataset for network training.*

The author of this thesis designed the networks' architecture, created the dataset and conducted all experiments. All co-authors contributed to the project discussions as well as writing the publication.

3.1 INTRODUCTION

The hand is the primary operating tool for humans. Therefore, its location, orientation and articulation in space is vital for many potential applications, for instance, object handover in robotics, learning from demonstration, sign language and gesture recognition, and using the hand as an input device for man-machine interaction.

Full 3D hand pose estimation from single images is difficult because of many ambiguities, strong articulation, and heavy self-occlusion, even more so than for the overall human body. Therefore, specific sensing equipment like data gloves or markers are used, which restrict the application to limited scenarios. Also the use of multiple

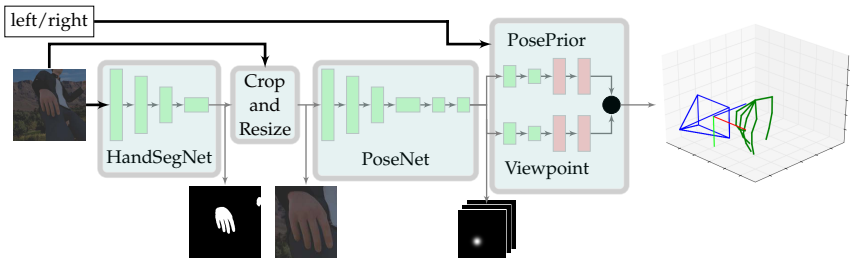

Figure 3.2: **Approach overview.** Our approach consists of three building blocks. First, the hand is localized within the image by a segmentation network (*HandSegNet*). Accordingly to the hand mask, the input image is cropped and serves as input to the *PoseNet*. This localizes a set of hand keypoints represented as score maps $\mathbf{S}$. Subsequently, the *PosePrior* network estimates the most likely 3D structure conditioned on the score maps. This figure serves for illustration of the overall approach and does not reflect the exact architecture of the individual building blocks.

cameras severly limits the application domain. Most contemporary works rely on the depth image from a depth camera. However, depth cameras are not as commonly available as regular color cameras, and they only work reliably in indoor environments.

In this paper, we present an approach to learn full 3D hand pose estimation from single color images without the need for any special equipment. We capitalize on the capability of deep networks to learn sensible priors from data in order to resolve ambiguities. Our overall approach consists of three deep networks that cover important subtasks on the way to the 3D pose; see Figure 3.2. The first network provides a hand segmentation to localize the hand in the image. Based on its output, the second network localizes hand keypoints in the 2D images. The third network finally derives the 3D hand pose from the 2D keypoints, and is the main contribution of this paper. In particular, we introduce a canonical pose representation to make this learning task feasible.

Another difficulty compared to 3D pose estimation at the level of the human body is the restricted availability of data. While human body pose estimation can leverage several motion capture databases, there is hardly any such data for hands. To train a network, a large dataset with ground truth 3D keypoints is needed. Since there is no such dataset with sufficient variability, we created a synthetic dataset with various data augmentation options.

3.2 RELATED WORK

2D Human Pose Estimation. Spurred by the MPII Human Pose benchmark [4] and the advent of Convolutional Neural Networks (CNN) this field made large progress in the last years. The CNN archi-

tecture of Toshev and Szegendy [126] directly regresses 2D cartesian coordinates from color image input. More recent works like Thompson et al. [125] and Wei et al. [133] turned towards regressing score maps. For parts of our work, we employ a comparable network architecture as Wei et al. [133].

3D Human Pose Estimation. We only discuss the most relevant works here and refer to Sarafianos et al. [105] for more information. Like our approach, many works use a two part pipeline [17, 124]. First they detect keypoints in 2D to utilize the discriminative power of current CNN approaches and then attempt to lift the set of 2D detections into 3D space. Different methods for lifting the representation have been proposed: Chen et al. [15] deployed a nearest neighbor matching of a given 2D prediction using a database of 2D to 3D correspondences. Tome et al. [123] created a probabilistic 3D pose model based upon a mixture of probabilistic PCA bases. Pavlakos et al. [96] proposed a volumetric approach that treats pose estimation as per voxel prediction of scores in a coarse-to-fine manner, which gives a natural representation to the data, but is computationally expensive and limited by the GPU memory to fit the voxel grid. Recently, there have been several approaches that apply deep learning to lifting 2D keypoints to 3D pose for human body pose estimation [85, 147]. Mehta et al. [80] uses transfer learning to infer the 3D body pose directly from images with a single network. While these works are all on 3D body pose estimation, we provide the first such work for 3D hand pose estimation, which is substantially harder due to stronger articulation and self-occlusion, as well as less data being available.

Hand Pose Estimation. Athitsos and Sclaroff [5] proposed a single frame based detection approach based on edge maps and Chamfer matching. With the advent of low-cost consumer depth cameras, research focused on hand pose from RGBD data. Oikonomidis et al. [92] proposed a technique based on Particle Swarm Optimization (PSO). Sharp et al. [107] added the possibility for reinitialization. A certain number of candidate poses is created and scored against the observed depth image. Tompson et al. [125] used a CNN for detection of hand keypoints in 2D, which is conditioned on a multi-resolution image pyramid. The pose in 3D is recovered by solving an inverse kinematics optimization problem. Approaches like Zhou et al. [149] or Oberweger et al. [90] train a CNN that directly regresses 3D coordinates given hand cropped depth maps. Whereas Oberweger et al. [90] explored the possibility to encode correlations between keypoint coordinates in a compressing bottleneck, Zhou et al. [149] estimate angles between bones of the kinematic chain instead of Cartesian coordinates. Oberweger et al. [91] presented an approach that replaces the explicit man-made hand model with a CNN that can synthesize a depth map from a given pose estimate. This allows them to successively refine initial pose estimates by min-

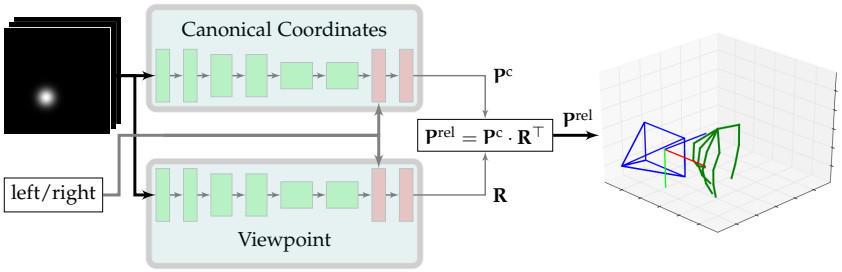

Figure 3.3: **Proposed architecture for the *PosePrior* network.** Two almost symmetric streams estimate canonical coordinates and the viewpoint relative to this coordinate system. Combination of the two predictions yields an estimation for the relative normalized coordinates $\mathbf{P}^{rel}$.

imizing the distance between the observed and the predicted depth image.

There are not yet any approaches that tackle the problem of 3D hand pose estimation from a single color image with a learning based formulation. Previous approaches differ because they rely on depth data [90, 91, 125, 149], they use explicit models to infer pose by matching against a predefined database of poses [5], or they only perform tracking based on an initial pose rather than full pose estimation [92, 107].

3.3 HAND POSE REPRESENTATION

Given a color image $\mathbf{I} \in \mathbb{R}^{N \times M \times 3}$ showing a single hand, we want to infer its 3D pose. We define the hand pose by a set of coordinates $\vec{P}_i = (X_i, Y_i, Z_i)$, which describe the locations of 21 keypoints in 3D space, i.e., $i \in [1, J]$ with $J = 21$.

The problem of inferring 3D coordinates from a single 2D observation is ill-posed. Among other ambiguities, there is a scale ambiguity. Thus, we infer a scale-invariant 3D structure by training a network to estimate normalized coordinates

$$\vec{P}_i^{norm} = \frac{1}{s} \cdot \vec{P}_i, \tag{3.1}$$

where $s = \left\| \vec{P}_{k+1} - \vec{P}_k \right\|_2$ is a sample dependent constant that normalizes the distance between a certain pair of keypoints to unit length. We choose k such that $s = 1$ for the first bone of the index finger.

Moreover, we use relative 3D coordinates to learn a translation invariant representation of hand poses. This is realized by subtracting the location of a defined root keypoint. The relative and normalized 3D coordinates are given by

$$\vec{P}_i^{rel} = \vec{P}_i^{norm} - \vec{P}_r^{norm} \tag{3.2}$$

where r is the root index. In experiments the palm keypoint was the most stable landmark. Thus we use $r = 0$.

3.4 ESTIMATION OF 3D HAND POSE

We estimate three-dimensional normalized coordinates $\mathbf{P}^{\text{rel}}$ from a single input image. An overview of the general approach is provided in Figure 3.2. In the following sections, we provide details on its components.

3.4.1 *Hand segmentation with HandSegNet*

For hand segmentation we deploy a network architecture that is based on and initialized by the person detector of Wei et al. [133]. They cast the problem of 2D person detection as estimating a score map for the center position of the human. The most likely location is used as center for a fixed size crop. Since the hand size drastically changes across images and depends much on the articulation, we rather cast the hand localization as a segmentation problem. Our *HandSegNet* is a smaller version of the network from Wei et al. [133] trained on our hand pose dataset. Details on the network architecture and its training procedure are provided in the supplemental material of [151]. The hand mask provided by *HandSegNet* allows us to crop and normalize the inputs in size, which simplifies the learning task for the *PoseNet*.

3.4.2 *Keypoint score maps with PoseNet*

We formulate localization of 2D keypoints as estimation of 2D score maps $\mathbf{S} = \{\mathbf{S}_1(u, v), \ldots, \mathbf{S}_J(u, v)\}$. We train a network to predict J score maps $\mathbf{S}_i \in \mathbb{R}^{N \times M}$, where each map contains information about the likelihood that a certain keypoint is present at a spatial location.

The network uses an encoder-decoder architecture similar to the Pose Network by Wei et al. [133]. Given the image feature representation produced by the encoder, an initial score map is predicted and is successively refined in resolution. We initialized with the weights from Wei et al. [133], where it applies, and retrained the network for hand keypoint detection. A complete overview over the network architecture is located in the supplemental material of [151].

3.4.3 *3D hand pose with the PosePrior network*

The *PosePrior* network learns to predict relative, normalized 3D coordinates conditioned on potentially incomplete or noisy score maps $\mathbf{S}(u, v)$. To this end, it must learn the manifold of possible hand artic-

ulations and their prior probabilities. Conditioned on the score maps, it will output the most likely 3D configuration given the 2D evidence.

Instead of training the network to predict absolute 3D coordinates, we rather propose to train the network to predict coordinates within a canonical frame and additionally estimate the transformation into the canonical frame. Explicitly enforcing a representation that is invariant to the global orientation of the hand is beneficial to learn a prior, as we show in our experiments in Section 3.5.2.

Given the relative normalized coordinates we propose to use a canonical frame $\mathbf{P}^c$, that relates to $\mathbf{P}^{rel}$ in the following way: An intermediate representation

$$\vec{P}_i^{c*} = \mathbf{R}(\mathbf{P}^{rel}) \cdot \vec{P}_i^{rel} \tag{3.3}$$

with $\mathbf{R}(\mathbf{P}^{rel}) \in \mathbb{R}^{3\times3}$ being a 3D rotation matrix is calculated in a two step procedure. First, one seeks the rotation $\mathbf{R}_{xz}$ around the x- and z-axis such that a certain keypoint $\vec{P}_a^{c*}$ is aligned with the y-axis of the canonical frame:

$$\mathbf{R}_{xz} \cdot \vec{P}_a^{c*} = \lambda \cdot (0, 1, 0)^\top \text{ with } \lambda \geqslant 0. \tag{3.4}$$

Afterwards, a rotation $\mathbf{R}_y$ around the y-axis is calculated such that

$$\mathbf{R}_y \cdot \mathbf{R}_{xz} \cdot \vec{P}_o^{c*} = (\eta, \zeta, 0) \tag{3.5}$$

with $\eta \geqslant 0$ for a specified keypoint index o. The total transformation between canonical and original frame is given by

$$\mathbf{R}(\mathbf{P}^{rel}) = \mathbf{R}_y \cdot \mathbf{R}_{xz}. \tag{3.6}$$

In order to deal appropriately with the symmetry between left and right hands, we flip right hands along the z-axis, which yields the side agnostic representation

$$\vec{P}_i^c = \begin{cases} (X_i^{c*}, Y_i^{c*}, Z_i^{c*})^\top & \text{if its a left hand} \\ (X_i^{c*}, Y_i^{c*}, -Z_i^{c*})^\top & \text{if its a right hand} \end{cases} \tag{3.7}$$

that resembles our proposed canonical coordinate system. Given this canonical frame definition, we train our network to estimate the 3D coordinates within the canonical frame $\mathbf{P}^c$ and separately to estimate the rotation matrix $\mathbf{R}(\mathbf{P}^{rel})$, which we parameterize using axis-angle notation with three parameters. Estimating the transformation $\mathbf{R}$ is equivalent to predicting the viewpoint of a given sample with respect to the canonical frame. Thus, we refer to the problem as *viewpoint estimation*.

The network architecture for the pose prior has two parallel processing streams; see Figure 3.3 and use an almost identical architecture. They first process the stack of J score maps in a series of 6 convolutions with ReLU non-linearities. Information on whether the

image shows a left or right hand is concatenated with the feature representation and processed further by two fully-connected layers. The streams end with a fully-connected layer with linear activation, which yields estimations for viewpoint $\mathbf{R}$ and canonical coordinates $\mathbf{P}^{c}$. Both estimations combined lead to an estimation of $\mathbf{P}^{rel}$.

3.4.4 *Network training*

For training of *HandSegNet* we apply standard softmax cross-entropy loss and L_2 loss for *PoseNet*. The *PosePrior* network uses two loss terms. First a squared L_2 loss for the canonical coordinates

$$\mathcal{L}_c = \left\| \mathbf{P}^{c}_{gt} - \mathbf{P}^{c}_{pred} \right\|_2^2 \tag{3.8}$$

based on the network predictions $\mathbf{P}^{c}_{pred}$ and the ground truth $\mathbf{P}^{c}_{gt}$. Secondly, a squared L_2 loss is imposed on the canonical transformation matrix:

$$\mathcal{L}_r = \left\| \mathbf{R}_{pred} - \mathbf{R}_{gt} \right\|_2^2 . \tag{3.9}$$

The total loss function is the unweighted sum of $\mathcal{L}_c$ and $\mathcal{L}_r$.

We used Tensorflow [1] with the Adam solver [61] for training. Details on the learning procedure are in the supplementary material of [151].

3.4.5 *Available datasets*

There are two available datasets that apply to our problem, as they provide RGB images and 3D pose annotation. The so-called *Stereo Hand Pose Tracking Benchmark* [144] provides both 2D and 3D annotations of 21 keypoints for 18000 stereo pairs with a resolution of 640×480. The dataset shows a single person's left hand in front of 6 different backgrounds and under varying lighting conditions. We divided the dataset into an evaluation set of 3000 images (*S-val*) and a training set with 15000 images (*S-train*).

Dexter [117] is a dataset providing 3111 images showing two operators performing different kinds of manipulations with a cuboid in a restricted indoor setup. The dataset provides color images, depth maps, and annotations for fingertips and cuboid corners. The color images have a spatial resolution of 640×320. Due to the incomplete hand annotation, we use this dataset only for investigating the cross-dataset generalization of our network. We refer to this test set as *Dexter*.

We downsampled both datasets to a resolution of 320×240 to be compatible with our rendered dataset. We transform our results back to coordinates in the original resolution, when we report pixel accuracies in the image domain.

Figure 3.4: Our new dataset provides segmentation maps with 33 classes: three for each finger, palm, person, and background. The 3D kinematic model of the hand provides 21 keypoints per hand: 4 keypoints per finger and one keypoint close to the wrist.

The NYU Hand Pose Dataset by Tompson et al. [125], commonly used for hand pose estimation from depth images, does not apply to a color based approach, because only registered color images are provided. In the supplementary we show more evidence why this dataset cannot be used for our task.

3.4.6 *Rendered hand pose dataset*

The above datasets are not sufficient for training a deep network due to limited variation, number of available samples, and partially incomplete annotation. Therefore, we complement them with a new dataset for training. To avoid the known problem of poor labeling performance by human annotators in three-dimensional data, we utilize freely available 3D models of humans with corresponding animations from Mixamo [83]. Then we used the open source software Blender [26] to render images. The dataset is publicly available.

Our dataset is built upon 20 different characters performing 39 actions. We split the data into a validation set (*R-val*) and a training set (*R-train*), where a character or action can exclusively be in one of the sets but not in the other. Our proposed split results into 16 characters performing 31 actions for training and 4 characters with 8 actions in the validation set.

For each frame we randomly sample a new camera location, which is roughly located in a spherical vicinity around one of the hands. All hand centers lie approximately in a range between 40cm and 65cm from the camera center. Both left and right hands are equally likely and the camera is rotated to ensure that the hand is at least partially visible from the current viewpoint. After the camera location and orientation are fixed, we randomly sample one background image from a pool of 1231 background images downloaded from Flickr [1]. Those

1 http://www.flickr.com

images show different kinds of scenes from cities and landscapes. We manually ensured that they do not contain persons.

To maximize the visual diversity of the dataset, we randomize the following settings for each rendered frame: we apply lighting by 0 to 2 directional light sources and global illumination, such that the color of the sampled background image is roughly matched. Additionally we randomize light positions and intensities. Furthermore, we save our renderings using a lossy JPG compression with the quality factor being randomized from no compression up to 60%. We also randomized the effect of specular reflections on the skin.

In total our dataset provides 41258 images for training and 2728 images for evaluation with a resolution of 320×320 pixels. All samples come with full annotation of a 21 keypoint skeleton model of each hand and additionally 33 segmentation masks are available plus the background. As far as the segmentation masks are concerned there is a class for the human, one for each palm and each finger is composed by 3 segments. Figure 3.4 shows a sample from the dataset. Every finger is represented by 4 keypoints: the tip of the finger, two intermediate keypoints and the end located on the palm. Additionally, there is a keypoint located at the wrist of the model. For each of the hand keypoints, there is information if it is visible or occluded in the image. Also keypoint annotations in the camera pixel coordinate system and in camera centered world coordinates are given. The camera intrinsic matrix and a ground truth depth map are available, too, but were not used in this work.

3.5 EXPERIMENTS

We evaluated all relevant parts of the overall approach: (1) the detection of hand keypoints of the *PoseNet* with and without the hand segmentation network; (2) the 3D hand pose estimation and the learned 3D pose prior. Finally, we applied the hand pose estimation to a sign language recognition benchmark.

3.5.1 *Keypoint detection in 2D*

Table 3.1 shows the performance of *PoseNet* on 2D keypoint estimation. We report the average endpoint error (EPE) in pixels and the area under the curve (AUC) on the percentage of correct keypoints (PCK) for different error thresholds; see Figure 3.6.

We evaluated two cases: one using images, where the hand is cropped with the ground truth oracle (GT), and one using the predictions from *HandSegNet* for cropping (Net). The first case shows the performance of *PoseNet* in isolation, while the second shows the performance of the complete 2D keypoint estimation. The difference between the median and the mean for the latter case show that *Hand-*

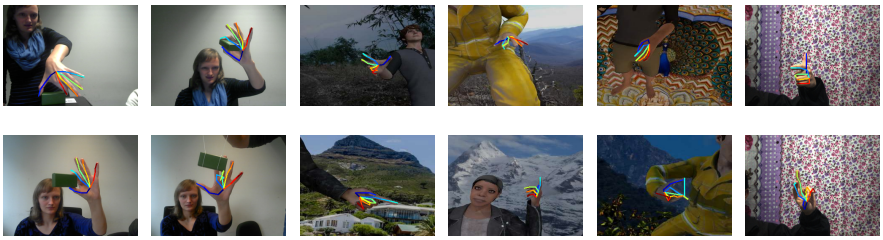

Figure 3.5: **Exemplary 2D keypoint localization results.** The first two columns show samples from *Dexter*, the following three depict *R-val* and the last one are samples from *S-val*.

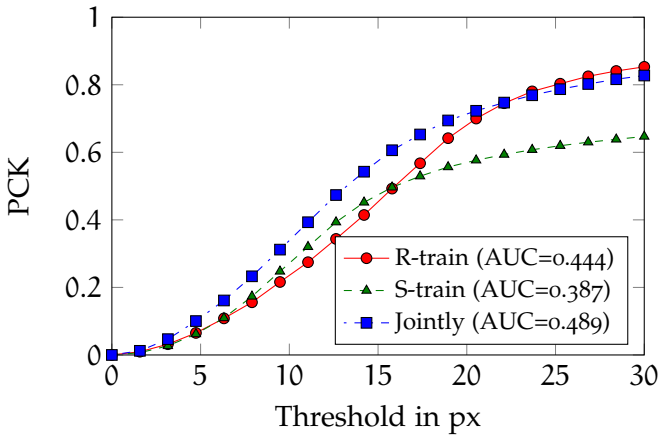

Figure 3.6: **Results on 2D keypoint estimation when using different training sets for *PoseNet*.** Shown is the percentage of correct keypoints (PCK) over a certain threshold in pixels evaluated on *Dexter*. Jointly training on R-train and S-train yields the best results.

SegNet is reliable in most cases but is sometimes not able to segment the hand correctly, which makes the 2D keypoint prediction fail.

The results show that the method works on our synthetic dataset (*R-val*) and the stereo dataset (*S-val*) equally well. The *Dexter* dataset is more difficult because the dataset is different from the training set and because of frequent occlusions of the hand by the handled cube. We did not have samples with occlusion (apart from self-occlusion) in the training set.

In Figure 3.6 we show that training on more diverse data helps cross-dataset generalization. While training only on our synthetic dataset *R-train* yields much better results on *Dexter* than training on the limited stereo dataset *S-train*, training on *R-train* and *S-train* together yields the best results. Figure 3.5 shows some qualitative results of this configuration.

		AUC	EPE median	EPE mean
GT	R-val	0.724	5.001	9.135
	S-val	0.817	5.522	5.013
Net	R-val	0.663	5.833	17.041
	S-val	0.762	5.528	18.581
	Dexter	0.489	13.684	25.160

Table 3.1: **Quantitative results for *PoseNet*.** The top rows (GT) report performance for the *PoseNet* operating on ground truth cropped hand images. The bottom rows (Net) show results when the hand crops are generated using *HandSegNet*. *PoseNet* was trained jointly on *R-train* and *S-train*, whereas *HandSegNet* was only trained on *R-train*. End point errors are reported in pixels with respect to the uncropped image and AUC is calculated over an error range from 0 to 30 pixels.

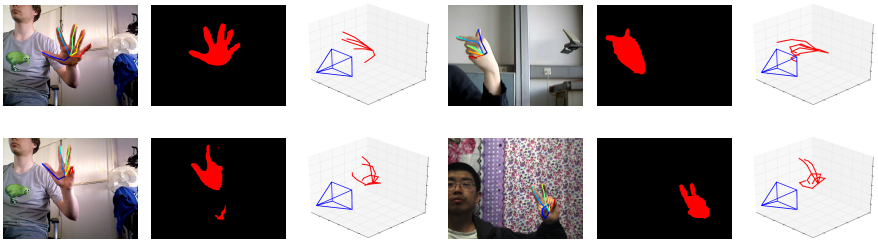

Figure 3.7: **Qualitative examples of our complete system.** Input to the network are color image and the information if its a left or right hand. The network estimates the hand segmentation mask, localizes keypoints in 2D and outputs the most likely 3D pose. The samples on the left hand side are from a dataset we recorded for qualitative evaluation, on the top right hand side is a sample from the sign language dataset and the bottom right sample is taken from *S-val*.

3.5.2 *Lifting the estimation to 3D*

3.5.2.1 *Pose representation*

We evaluated the proposed canonical frame representation for predicting the 3D hand pose from 2D keypoints by comparing it to several alternatives. All variants share a common base architecture that is identical to one stream of the *PosePrior* proposed in Section 3.4.3. They were trained on score maps $\mathbf{S}$ with a spatial resolution of 32 by 32 pixels. To avoid overfitting, we augmented the score maps by applying channelwise dropout with a drop probability of 0.2. This forces the networks to deal with incomplete score maps. Additionally we disturbed the keypoint location with Gaussian noise and randomly translated the keypoints by up to 2.5 pixel. Table 3.2 shows the resulting end point errors per keypoint.

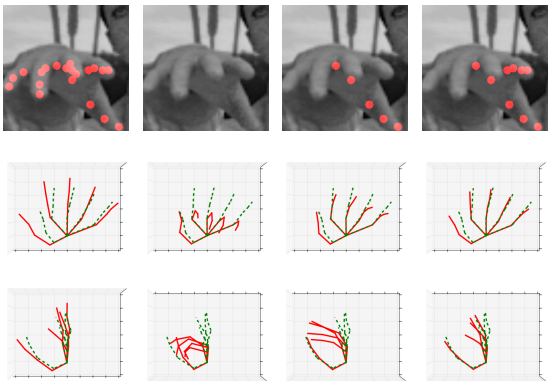

Figure 3.8: **Analysis of the learned prior.** The first row shows the input image as gray scale with the input score map overlayed as red dots. Every column corresponds to a separate forward pass of the network. The second and third row visualize the predicted 3D structure of the network from different viewpoints in canonical coordinates. Ground truth is displayed in dashed green and the network prediction is shown in solid red.

The *Direct* approach tries to lift the 2D keypoints directly to the full 3D coordinates $\mathbf{P}^{\text{rel}}$ without using a canonical frame. This is disadvantageous, because it is difficult for the network to learn separate the global rotation of the hand from the articulation.

The *Bottleneck* approach is inspired by Oberweger et al. [90], who introduced a bottleneck layer before estimating the coordinates. We inserted an additional FC layer before the final FC output layer, parameterize it as in Oberweger et al. with 30 channels and linear activation. The outcome was not better than with the *Direct* approach.

The *Local* approach incorporates the kinematic model of the hand and uses the network to estimate articulation parameters of the model [149]. We generalize by estimating not only the angles but also the bone length. The network is trained to estimate two angles and one length per keypoint, which results in 63 parameters. The angles express rotations in a bone local coordinate system. This approach only works if the hand is always shown from the same direction, but cannot capture the global pose of the hand.

	Direct	Bottleneck	Local	NN	**Prop.**
R-train	20.15	21.07	35.15	0.00	18.54
R-val	20.85	21.91	39.12	26.92	18.84

Table 3.2: **Quantitative results for different Lifting approaches.** Average median end point error per keypoint of the predicted 3D pose is reported given a noisy ground truth 2D pose. Networks were trained on *R-train* and ground truth scale was used at test time to report results in mm.

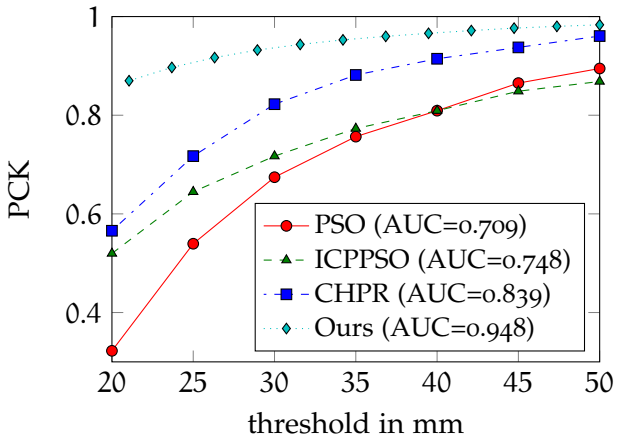

Figure 3.9: **Comparison to literature.** Results for our complete system on *S-val* compared to classical approaches from [144] Shown is the percentage of correct keypoints (PCK) over respective thresholds in mm. *PoseNet* and *PosePrior* are trained on *S-train* and *R-train*, whereas the *HandSegNet* is trained on *R-train*.

Finally, the *NN* approach matches the 2D keypoints to the most similar sample from the training set and retrieves the 3D coordinates from this sample [15]. While this approach trivially works best on the training set, it does not generalize well to new samples.

The generalization of the other approaches is quite good showing similar errors for both the training and the validation set. The proposed approach from Section 3.4.3 worked best and was used for the following experiments.

3.5.2.2 *Analysis of the learned prior*

To examine the 3D prior learned by the network, Figure 3.8 shows the 3D pose prediction from two different viewpoints, for score maps that lack keypoints. The extreme case, with no keypoints provided as input at all, shows the canonical prior learned by the network. As more keypoints are added, the networks adjusts the predicted pose to this additional evidence. This experiment also simulates the situation of occluded 2D keypoints and demonstrates that the learned prior allows the network to still retrieve reasonable poses.

3.5.2.3 *Comparison to literature*

Since there is no work yet on 3D hand pose estimation from RGB images yet, we cannot compare to alternative approaches. To still relate our results coarsely to literature, we compare them to Zhang et al. [144], who provide results in mm for state-of-the-art 3D hand pose tracking on depth data. They run their experiments on the stereo dataset *S-val*, which also contains RGB images. Since in contrast to

Zhang et al. our approach does not use the depth data, it still comes with ambiguities with regard to scale and absolute depth.

Thus, we accessed the absolute position of the root keypoint and the scale of the hand to shift and scale our predicted 3D hand pose, which yields metric world coordinates $\mathbf{P}$ by using (3.1) and (3.2). For this experiment we trained *PosePrior* on score maps predicted by *PoseNet* using the same schedule as for the experiment in Section 3.5.2.2. *PoseNet* is trained separately as described in Section 3.5.1 and then kept fixed. Figure 3.9 shows that our approach largely outperforms the approaches presented in Zhang et al. [144] although we use the depth map only for rescaling and shifting in the end.

Qualitative 3D examples on three different datasets with the complete processing pipeline are shown in Figure 3.7.

3.5.3 *Sign language recognition*

Previous hand pose estimation approaches depending on depth data cannot be applied to most sign language recognition datasets, as they only come with color images. As a last exemplary experiment, we used our hand pose estimation system and trained a classifier for gesture recognition on top of it. The classifier is a fully connected three layer network with ReLU activation functions.

We report results on the so-called *RWTH German Fingerspelling Database* [19]. It contains 35 gestures representing the letters of the alphabet, German umlauts, and the numbers from one to five. The dataset comprises 20 different persons, who did two recordings each for every gesture. Most of the gestures are static except for the ones for the letters J, Z, Ä, Ö, and Ü, which are dynamic. In order to keep this experiment simple, we ran the experiments on the subset restricted to 30 static gestures.

The database contains recordings by two different cameras, but we used only one camera. The short videos sequences have a resolution of 320×240 pixels. We grabbed the middle frame from each video sequence and used those color images and gesture class labels as training data. This dataset has 1160 images, which we separate by signers into a validation set with 232 images and a training set with 928 images. We resize image to 320×320 pixels and trained on randomly sampled 256×256 crops. Because the images were taken from a compressed video stream they exhibit significant compression artifacts previously unseen by our networks. Thus, we labeled 50 images from the training set with hand keypoints, which we use to fine-tune our *PoseNet* upfront. Afterwards the pose estimation part is kept fixed and we solely train the *GestureNet*. Table 3.3 show that our system archives comparable results to Dreuw et al. [19] on the subset of gestures we used for the comparison.

Method	Word error rate
Dreuw et al. [19]	35.7 %
Dreuw on subset [98]	36.56 %
Ours 3D	33.2 %

Table 3.3: **Comparison on Sign Language Recognition.** Word error rates in percent on the RWTH German Fingerspelling Database subset of non dynamic gestures. Results for Dreuw et al. [19] on the subset were taken from [98].

3.6 CONCLUSION

We have presented the first learning based system to estimate 3D hand pose from a single image. We contributed a large synthetic dataset that enabled us to train a network successfully on the task. We have shown that the network learned a 3D pose prior that allows it to predict reasonable 3D hand poses from 2D keypoints in real world images. While the performance of the network is even competitive to approaches that use depth maps, there is still much room for improvements. The performance seems mostly limited by the lack of an annotated large scale dataset with real-world images and diverse pose statistics.

3.7 FOLLOW-UP WORK

Consequently to this publication numerous works appeared that extend the ideas presented.

A lot of work focused on novel approaches that were trained on the dataset presented here. Some examples are the latent 2.5D heatmap representation by Iqbal et al. [51], that achieved state of the art for 3D hand pose estimation in a supervised training setting or Panteleris et al. [95] that combined 2D keypoint detection with explicit optimization using a hand skeleton model. Tekin et al. [120] extended the task from hand pose alone towards estimating interaction of hands and objects being held.

Building on our insights on the usefulness of rendered training data multiple following works presented new datasets in a similar manner. Some examples are extensions towards hand and object interaction [41, 88].

Another line of works tackled the dataset sparsity by exploring ways to combine various data modalities. One prominent work among these is Spurr et al. [113], where encoders are learned that map into a joint representation space for color and depth input images. These are combined with a decoder that can estimate the 3D pose from the joint representation space. This allows to train jointly using both color and depth images and shows improvements over separately training on either resource.

4 ESTIMATING HUMAN POSE FROM SINGLE RGBD

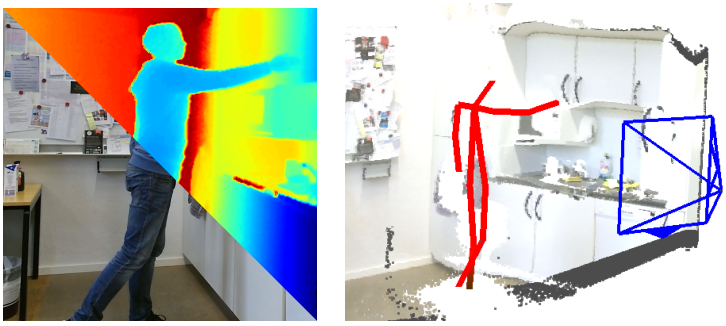

Figure 4.1: **Problem overview.** Given a color image and depth map, our system detects body keypoints in 3D, which are useful for many robotic tasks. Exemplary use in a learning from demonstration setting was described in detail in the respective paper [152] and is shortly summarized in Section 4.6.3.

This chapter describes ideas and experiments that were previously presented in the following work; therefore, copyright lies with © 2018 IEEE.

3D Human Pose Estimation in RGBD Images for Robotic Task Learning
Christian Zimmermann*, Tim Welschehold*, Christian Dornhege, Wolfram Burgard and Thomas Brox *(* equal contribution)*
IEEE International Conf. on Robotics and Automation (ICRA), 2018

> *This work presented an approach for human pose estimation from RGBD input images using a neural network. Using the pose estimation enables robot task learning through learning from human demonstration.*

The author of this thesis designed the network architecture, created the training datasets and conducted experiments related to the pose estimation system. Tim Welschehold developed the action imitation framework and conducted all robotic experiments. All co-authors contributed to the project discussions as well as writing the publication.

4.1 INTRODUCTION

Perception and understanding of the surrounding environment is vital for many robotics tasks. Tasks involving interaction with humans

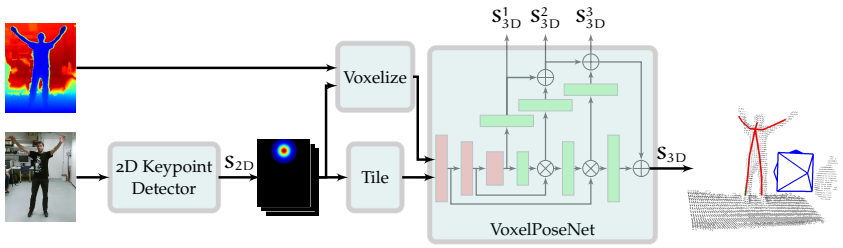

Figure 4.2: **Approach overview.** First, we predict the keypoint locations in the color image. The predicted score maps are tiled along the z-dimension and a person centered occupancy voxel grid is calculated from the depth map. Based on these inputs *VoxelPoseNet* predicts keypoints in 3D. Red and green blocks represent convolutional and deconvolutional operations. Concatenation is denoted by $\otimes$ and $\oplus$ is the elementwise add operation.

heavily rely on prediction of the human location and its articulation in space. These applications involve, e.g., gesture control, hand-over maneuvers, and learning from demonstration.

On the quest of bringing service robots to mass market and into common households, one of the major milestones is their instructability: consumers should be able to teach their personal robots their own custom tasks. Teaching should be intuitive and not require expert knowledge or programming skills. Ideally, the robot should learn from observing its human teacher demonstrating the task at hand. Hence it needs to be able to follow the human motion.

Estimation of human pose is challenging due to variation in appearance, strong articulation and heavy occlusions by themselves or objects. Recent approaches present robust pose estimators in 2D, but for robotic applications full 3D pose estimation in real-world units is indispensable. In this paper, we bridge this gap by lifting 2D predictions into 3D while incorporating information from a depth map. This lifting via a depth map is non-trivial for multiple reasons, for instance, occlusion of the person by an object leads to misleading depths, see Figure 4.7.

In this chapter a learning based approach that predicts full 3D human pose is presented which outperforms existing baseline methods and enables teaching a robot tasks by demonstration.

The approach first predicts human pose in 2D given the color image. A deep network takes the 2D pose and the depth map as input and derives the full 3D pose from this information. Based on this pose estimation system, we demonstrate the feasibility of our action learning from human demonstration approach without the use of artificial markers on the person.

4.2 RELATED WORK

The vast majority of publications in the field of human pose estimation deal with the problem of inferring keypoints in 2D given a color image [14, 133], which is linked to the availability of large scale datasets [4, 70]. Due to the large datasets, networks for keypoint localization in 2D have reached impressive performance, which we integrate into our approach.

Recent techniques learn a prior for human pose that allows prediction of the most likely 3D pose given a single color image [74, 123]. Predictions of most monocular approaches live in a scale and translation normalized frame, which makes them impracticable for many robotic applications. Approaches that can recover full 3D from RGB alone [81] use assumptions to resolve the depth ambiguity. Our approach does not need any assumptions to predict poses in world coordinates.

All approaches that provide predictions in real-world units are based on active depth sensing equipment. Most prominent is the Microsoft Kinect v1 sensor. Shotton et al. [111] describes a discriminative method that is based on random forest classifiers and yields a body part segmentation. This work was followed by numerous approaches that propose using random tree walks [55], a viewpoint invariant representation [40] or local volumetric convolutional networks for local predictions [84]. In contrast to the mentioned techniques, we incorporate depth and color in a joint approach. So far little research went into approaches that incorporate both modalities [11]. We propose a deep learning based approach to combine color and depth. Our approach leverages the discriminative power of keypoint detectors trained on large scale databases for color images and complements them with information from the depth map for lifting to 3D real-world coordinates.

4.3 HUMAN POSE ESTIMATION

We aim to estimate 3D human poses from RGBD input and this procedure is summarized in Figure 4.2.

We aim for estimating the human body keypoints $\mathbf{P} = (\vec{P}_1, \ldots, \vec{P}_J) \in \mathbb{R}^{3\times J}$ for J keypoints in real-world coordinates relative to the Kinect sensor given color image $\mathbf{I} \in \mathbb{R}^{N\times M\times 3}$, depth map $\mathbf{D}' \in \mathbb{R}^{N'\times M'}$ and their calibration. Without loss of generality we define the coordinate system, our predictions live in, to be identical with the color sensors frame.

For the Kinect, the color and depth sensors are located in close proximity, but still the frames resemble two distinct cameras. Our approach needs to collocate information of the two frames. Therefore we transform the depth map into the color frame using the camera

calibration. As a result, our approach operates on the warped depth map $\mathbf{D} \in \mathbb{R}^{N \times M}$. Due to occlusions, differences in resolution and noise, the resulting depth map $\mathbf{D}$ is sparse, but for better visualization a linear interpolation of $\mathbf{D}$ is shown in Figure 4.2.

4.3.1 *Color Keypoint Detector*

The keypoint detector is applied to the color image $\mathbf{I}$, which yields score maps $\mathbf{S}_{2D} \in \mathbb{R}^{N \times M \times J}$ encoding the likelihood of a specific human keypoint being present. The maxima of the score maps $\mathbf{S}_{2D}$ correspond to the predicted keypoint locations $\mathbf{p} = (\vec{p}_0, \ldots, \vec{p}_J) \in \mathbb{R}^{2 \times J}$ in the image plane. Thanks to many datasets with annotated color frames for human pose estimation [4, 70], robust detectors are available. We use the Open Pose Library [14, 112, 133] with fixed weights in this work.

4.3.2 *VoxelPoseNet*

Given the warped depth map $\mathbf{D}$ a voxel occupancy grid $\mathbf{V} \in \mathbb{R}^{K \times K \times K}$ is calculated with $K = 64$. For this purpose the depth map $\mathbf{D}$ is transformed into a point cloud and we calculate an 3D coordinate $\vec{P}_r$, which is the center of $\mathbf{V}$. We calculate $\vec{P}_r$ as back projection of the predicted 2D 'neck' keypoint $\vec{p}_r$ using the median depth d_r extracted from the neighborhood of $\vec{p}_r$ in $\mathbf{D}$:

$$\vec{P}_r = d_r \cdot \mathbf{K}^{-1} \cdot \vec{p}_r. \tag{4.1}$$

Where $\mathbf{K}$ denotes the intrinsic calibration matrix camera and $\vec{p}_r$ is in homogeneous coordinates. We pick the value d_r from the depth map taking into account the closest 3 neighboring valid depth values around $\vec{p}_r$. We calculate $\mathbf{V}$ by setting elements to 1, when there is at least one point of the point cloud lying in the interval represented and zero otherwise. We chose the resolution of the voxel grid to be approximately 3 cm.

VoxelPoseNet gets $\mathbf{V}$ and a volume of tiled score maps $\mathbf{S}_{2D}$ as input and processes them with a series of 3D convolutions. We propose to tile $\mathbf{S}_{2D}$ along the z-axis, which is equivalent to an orthographic projection approximation. *VoxelPoseNet* estimates score volumes $\mathbf{S}_{3D} \in \mathbb{R}^{K \times K \times K \times J}$, which resemble keypoint likelihoods the same way as its 2D counterpart

$$\mathbf{P}_{VPN} = \arg\max_{x,y,z}(\mathbf{S}_{3D}). \tag{4.2}$$

We use the following heuristic to assemble our final prediction: On the one hand $\mathbf{P}_{VPN}$ is predicted by *VoxelPoseNet*. On the other hand we take the z-component of $\mathbf{P}_{VPN}$ and the predicted 2D keypoints

$\mathbf{p}_{2D}$ to calculate another set of world coordinates $\mathbf{P}_{\text{projected}}$. For these coordinates the accuracy in x- and y-direction is not limited by the choice of K anymore. We chose our final prediction $\mathbf{P}$ from $\mathbf{P}_{\text{projected}}$ and $\mathbf{P}_{\text{VPN}}$ based on the 2D networks prediction confidence, which is the score of $\mathbf{S}_{2D}$ at $\mathbf{p}$.

Figure 4.2 shows the network architecture used for *VoxelPoseNet*, which is a encoder decoder architecture inspired by the U-net [103] that uses dense blocks [47] in the encoder. While decoding to the full resolution score map, we incorporate multiple intermediate losses denoted by $\mathbf{s}^{i}_{3D}$, which are discussed in Section 4.4.

4.4 NETWORK TRAINING

We train *VoxelPoseNet* using a sum of squared L_2 losses:

$$L = \sum_{i} \left\| \mathbf{s}^{gt}_{3D} - \mathbf{s}^{i,\,pred}_{3D} \right\|^2_2 \tag{4.3}$$

with a batch size of 2. Datasets used for training are discussed in Section 4.5. The networks are implemented in Tensorflow [1] and we use the ADAM solver [61]. We train for 40000 iterations with an initial learning rate of 10^{-4}, which drops by the factor 0.1 every 10000 iterations. Ground truth score volumes $\mathbf{s}^{gt}_{3D}$ are calculated from the ground truth keypoint location within the voxel $\mathbf{V}$. A Gaussian function is placed at the ground truth location and normalized such that its maximum is equal to 1.

4.5 DATASETS

Currently there are no datasets for the Kinect v2 that provide high-quality skeleton annotation of the person. Due to its long presence, most publicly available sets are recorded with the Kinect v1. These datasets are not suited for our scenario, because of major technical differences between the two models. More recently published datasets, such as Shahroudy et al. [106], transitioned to the new model but used the Kinect SDK's prediction as pseudo ground truth. Using those datasets is prohibitive for exceeding the Kinect SDK's performance.

4.5.1 *Multi View Kinect Dataset (MKV)*

Therefore, for training of our neural network we recorded a new dataset, which comprises 5 actors, 3 locations, and up to 4 viewpoints. There are 2 female and 3 male actors and the locations resemble different indoor setups. Some examples are depicted in Figure 4.3. The poses include various upright and sitting poses as well as walking sequences. Short sequences were recorded simultaneously by multiple

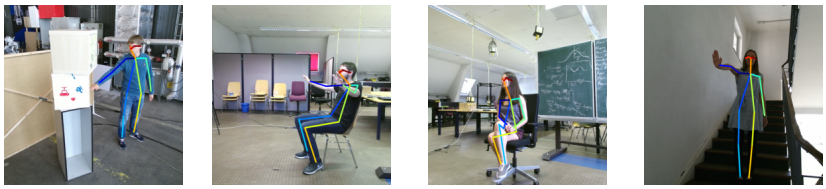

Figure 4.3: **Dataset Samples.** Examples from the *MKV* dataset with ground truth skeleton overlayed. The two leftmost ones are samples from the training set and the other two show the evaluation set.

calibrated Kinect v2 devices with a frame rate of 10 Hz, while recording the skeletal predictions of the Kinect SDK. In a post processing step we applied state-of-the-art Human Keypoint Detectors [14, 112, 133] and used standard triangulation techniques to lift the 2D predictions into 3D. This results in a dataset with 22406 samples. Each sample comprises of color image, depth map, infrared image, the SDK prediction and a ground truth skeleton annotation we get through triangulation. The skeleton annotations comprises of 18 keypoints that follow the Coco definitions [70]. We apply data augmentation techniques and split the set into an evaluation set of 3546 samples (*MVK-e*) and a training set with 18860 (*MVK-t*). We divide the two sets by actors and assign both female actors into the evaluation set, which also leaves one location unique to this set.

4.5.2 *Captury Dataset*

Due to the limited number of cameras in the *MKV* setup and the necessity to avoid occluding too many cameras views at the same time, we are limited in the amount of possible object interaction of the actors. Therefore we present a second dataset that was recorded using a commercial marker-less motion capture system called Captury[1]. It uses 12 cameras to track the actor with 120 Hz and we calibrated a Kinect v2 device with respect to the Captury. The skeleton tracking provides 23 keypoints, from which we use 13 for comparison. We recorded three actors, which performed simple actions like pointing, walking, sitting and interacting with objects like a ball, chair or umbrella. One actor of this setting was already recorded for the *MKV* dataset and therefore constitutes the set used for training. Two previously unseen actors were recorded and form the evaluation set. There are 1535 samples for training (**CAP-t**) and 1505 samples for evaluation (**CAP-e**). The definition of human keypoints between the two datasets is compatible, except for the "head" keypoint, which misses a suitable counterpart in the *MKV* dataset. This keypoint is excluded from evaluation to avoid systematic error in the comparison.

1 http://www.thecaptury.com

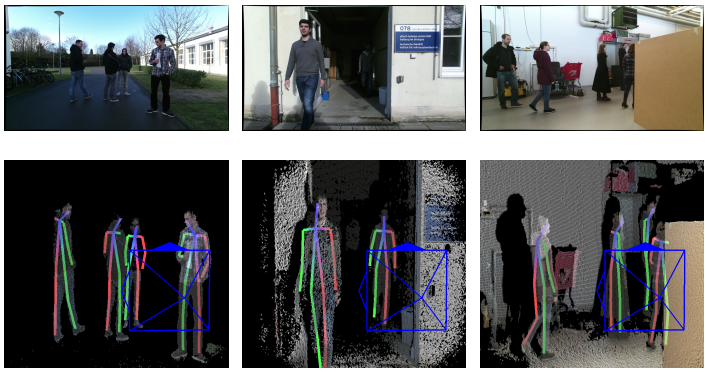

Figure 4.4: **Qualitative results.** Images taken from the *InOutDoor* Dataset [79], which covers a wide variety of environments our approach consistently works well in. Top row shows the color input image and the bottom row contains the respective pose predictions our method yields. Please note that the field of view between depth and color sensor differ, which results in some missing predictions towards the borders of the color image.

4.6 EXPERIMENTS

4.6.1 *Datasets for training*

Table 4.1 shows that the proposed *PoseNet3D* already reaches good results on the evaluation split of both datasets when trained only on *MKV-t*. Training a network only on *CAP-t* leads to inferior performance, which is due to starkly limited variation in the training split of the Captury dataset, which only contains a single actor and scene. Training jointly on both sets performs roughly on par with training exclusively on *MKV-t*. Therefore we use *MKV-t* as default training set for our networks and evaluate on *CAP-e* for following experiments. Furthermore, we confirm generalization of our *MKV-t* trained approach on the *InOutDoor* Dataset [79]. Because the dataset does not contain pose annotations we present qualitative results in Figure 4.4.

Training set	*CAP-e* full	*CAP-e* subset	*MKV-e*
MKV-t	0.627	0.618	0.793
CAP-t	0.603	0.588	0.665
CAP-t & *MKV-t*	0.633	0.625	0.794

Table 4.1: **Generalization between datasets.** Performance measured as area under the curve (AUC) for different training sets of *VoxelPoseNet*. *CAP-t* does not generalize to *MKV-e*, whereas *MKV-t* provides sufficient variation to generalize to *CAP-e*. Training jointly on *CAP-t* and *MKV-t* doesn't improve results much anymore.

	Captury full	Captury subset	Multi Kinect
Kinect SDK	13.5	16.4	8.9
Naive Lifting	14.7	15.2	8.8
Tome et al. [123]	22.7	21.9	15.1
Proposed	**11.2**	**11.6**	**6.1**

Table 4.2: **Approach comparison over various datasets.** Average mean end point error per keypoint of the predicted 3D pose for different approaches in cm. For the Captury dataset we additionally report results on the subset of non-frontal scenes and with object interaction.

4.6.2 *Comparison to literature*

In Table 4.2 we compare our approach with common baseline methods. The first baseline is the Skeleton Tracker integrated in Microsofts Software Development Kit[2] (Kinect SDK). We show that its performance heavily drops on the more challenging subset and therefore argue that it is unsuitable for many robotics applications. Furthermore, Figure 4.6 shows that the Kinect SDK is unable to predict keypoints farther away than a certain distance. The qualitative examples in Figure 4.7 reveal that the SDK is led astray by objects and is unable to distinguish if a person is facing towards or away from the camera, which expresses itself in mixing up left and right side.

The baseline named Naive Lifting uses the same Keypoint detector for color images as our proposed approach and simply picks the corresponding depth value from the depth map. It chooses the depth value as median value of the 3 closest neighbors. The approach shows reasonable performance, but is prone to pick bad depth values from the noisy depth map. Also any kind of occlusion results into an error, which is seen in Figure 4.7.

Tome et al. [123] predicts scale and translation normalized poses. So in order to compare the results to the other approaches we provide the algorithm with ground truth scale and translation. For every prediction we seek scale and translation in order to minimize the reconstruction error between ground truth and prediction. Table 4.2 shows that the approach reaches competitive results, but performs worst in our comparison, which is reasonable given the lack of depth information. In Figure 4.5 the approach stays far behind, which partly lies in the fact that the approach misses to provide predictions in 8.7% of the frames of *CAP-e*, which compares to 12.4% for Kinect SDK and 0% for Naive Lifting and our approach.

VoxelPoseNet outperforms its baseline methods, because it exploits both modalities. On the one hand, color information helps to disambiguate left and right side, which is infeasible from depth alone. On the other hand, the depth map provides valuable information to ex-

2 https://www.microsoft.com/en-us/download/details.aspx?id=44561

actly infer the 3D keypoint. Furthermore, the network learns a prior about possible body part configurations, which makes it possible to infer 3D locations even for completely occluded keypoints (see Figure 4.7).

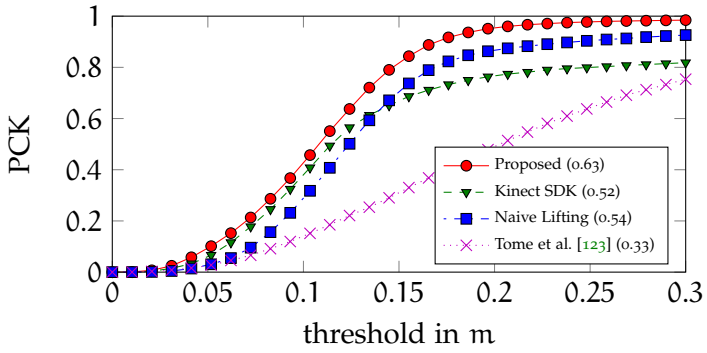

Figure 4.5: **Qualitative comparison between approaches.** Performance of different algorithms on *CAP-e* measured as percentage of correct keypoints (PCK) on the more challenging subset of non-frontal poses and object interaction.

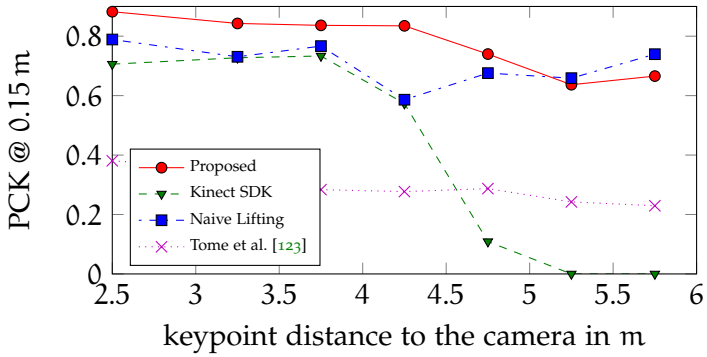

Figure 4.6: **Qualitative comparison over distances.** Percentage of correct keypoints (PCK) over their distance to the camera. Most approaches are only mildly affected by the keypoint distance to the camera, but the Kinect SDK can only provide predictions in a limited range.

4.6.3 *Action Imitation by the Robot*

In our respective paper [152] a recently proposed graph-based approach [135] for learning a mobile manipulation task from human demonstrations was used on data acquired with the approach for 3D human pose estimation presented in this work. The methods were evaluated on the same four tasks as in [135]: one task of opening and moving through a room door and three tasks of opening small furniture pieces. The tasks will be referred to as room door, swivel door, drawer, and sliding door. Each consists of three parts. First a

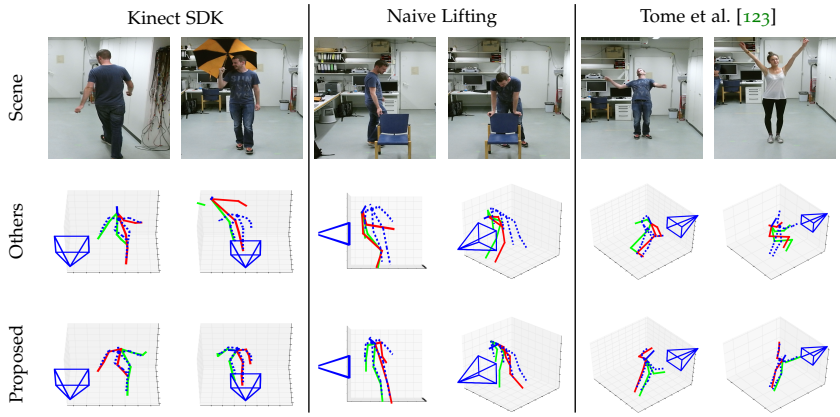

Figure 4.7: **Qualitative Examples.** Typical failure cases of the algorithms evaluated for samples from *CAP-e*. The first row shows the scene and the other two rows depict the ground truth skeleton in dashed blue and the prediction in solid green and red. Green color indicates the persons right side. Predictions of our proposed approach are shown in the last row, whereas the middle row shows predictions by other algorithms. The first two columns correspond to predictions of the Kinect SDK, the next two are by the Naive Lifting approach and the last two by the approach presented by Tome et al. [123]. Typical failures for the SDK are caused by objects and or people that face away from the camera. Naive Lifting fails when any sort of keypoint occlusion is present.

specific part of the object is grasped, i. e., a handle or a knob, then the object is manipulated according to its geometry, and lastly released. The demonstrations were recorded with a Kinect v2 at 10 Hz. As we need to track both, the manipulated object and the human teacher, the actions were recorded from a perspective that show the human from the side or back making pose estimation challenging. For an example of the setup see Figure 4.8.

We used adapted demonstrations from our pose estimation approach to learn action models that our PR2 robot can use to imitate the demonstrated actions in real-world settings. For details about the adaptation and learning process we refer the reader to our paper [152].

With the learned models each action was reproduced five times. For opening the swivel door there was one failure due to localization problems during the grasping. For the drawer and the room door all trials of grasping and manipulating were successful. The sliding door was always grasped successfully but due to the small door knob and the tension resulting from the combined gripper and base motion, the knob was accidentally released during the manipulation process. Five successful trials of opening the sliding door were run by keeping the robot base steady. A visualization of the teaching process and the robot reproducing the action demonstration can be seen in Figure 4.8.

Figure 4.8: **Demonstration and robot execution.** On the left image the teacher demonstrates the task of opening the swivel door. Superimposed on the image the recorded trajectories for hand (orange), torso (green) and manipulated object (blue) are shown which serve as the input for the action learning. The right image shows the robot reproducing the action using a model learned from the teacher demonstration.

4.7 CONCLUSION

A CNN based system was proposed that jointly uses color and depth information in order to predict 3D human pose in real-world units which exceeds the performance of existing methods. Furthermore, two RGBD datasets are proposed, which can be used for future approaches. In Section 4.6.3, the approach for 3D human pose estimation is applied in a task learning application that allows non-expert users to teach tasks to service robots. This is demonstrated in real-world experiments that enabled a PR2 robot to reproduce human-demonstrated tasks without any markers on the human teacher.

4.8 FOLLOW-UP WORK

The presented approach was made available as a ROS node to the community; therefore, numerous works used or extended it: Weschehold et al. [136, 137] used it for human pose estimation. Lindner et al. [71, 72] analyzed its applicability for people detection and Kollmitz et al. [63] performed detection of people conditioned on their mobility aids. Guo et al. [37] used a similar approach to perform gait analysis. Biswas et al. [8] optimized run-time towards real-time capability. Wengefeld et al. [138] specialized in people orientation estimation and optimizes the algorithm for run-time and accuracy for this scenario.

5

ESTIMATING POSE FROM MULTI-VIEW RGB

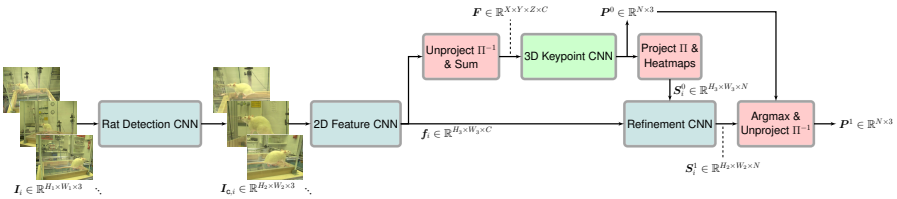

Figure 5.1: **Approach to estimate pose from calibrated multi-view.** Given the camera images first a bounding box detection network is applied. Then image features are extracted from the cropped images and unprojected into a common 3D representation. The 3D representation is used to estimate the initial pose $\mathbf{P}^0$, which is projected into the views for further refinement. Finally, the refined 2D estimations $\tilde{\mathbf{p}}_i$ are used to calculate the final 3D pose $\mathbf{P}^1$.

This chapter describes ideas and experiments that were mostly presented in the following work.

FreiPose: A Deep Learning Framework for Precise Animal Motion Capture in 3D Spaces
Christian Zimmermann*, Artur Schneider*, Mansour Alyahyay, Thomas Brox and Ilka Diester *(* equal contribution)*
Submission is in preparation.
bioRxiv, 2020

> *This work presents a generic approach for pose capture from multi-view RGB input and shows application of the system to pose estimation of laboratory animals. Using the animal pose information it was possible to attribute and quantify the effect of optogenetic stimulation in rodents.*

The author of this thesis implemented FreiPose and analyzed the optogenetic data. A.S. conducted the animal recordings, and performed viral injections for optogenetic stimulation. Contributions of M.A. are not presented in this thesis. A.S., T.B., I.D and the author of this thesis conceived and designed the study and wrote the manuscript.

5.1 INTRODUCTION

Detailed tracking of the animals' movements including single body parts and correlating them to neuronal activity is essential to assign functions to neuronal circuits and their activity. Several new tools

are already available for interpreting video data. However, existing tools are limited by one of two possible factors: (1) Marker-based approaches influence natural movements, are restricted to applicable body sites and rely on the tolerance of the animal. (2) Marker-free analyses have been applied only in 2D so far, thus preventing the true pose reconstruction of freely moving animals covering all three dimensions with their movements. Using multiple cameras for video taping, 2D outputs can be triangulated to yield 3D estimates a posteriori; however, such post-processing suffers from the ambiguities in the initial 2D analysis reducing accuracy and reliability. Though, exactly this tracking accuracy is crucial to subdivide movements into well-defined trajectories and behavioral classes for isolated analysis.

Here, we introduce the new tracking tool FreiPose, which allows reconstructing detailed body postures and single body part movements directly in 3D.

We used FreiPose to quantify the behavioral effect of optogenetic stimulation in motor cortex based on the rat's movements. Importantly, FreiPose even allowed to attribute the stimulation effect to individual body parts of the animal as well as to draw conclusions about the temporal dynamics of the stimulation effect. Altogether, FreiPose enabled marker-free tracking of individual body parts in an unprecedented manner. Thus, FreiPose is particularly suited for studies in which physiological recordings are conducted in freely moving animals and would allow conclusions about the behavioral state of the animal as well as detailed information of individual body parts in trained as well as in spontaneously behaving animals.

5.2 RELATED WORK

Estimating pose of animals from images is a closely coupled research field to human pose estimation, and frequently human pose methods were transferred to animal pose problems with only little modifications, e. g., Mathis et al. [75].

Therefore, similar trends can be found in both fields: Historically, marker based approaches like Mimica et al. [82] were used, which can achieve high accuracy when applicable. But this type of approach shows limitations when small animals need to be tracked. Furthermore, it must be prevented that the markers can be removed by the animal, and markers can hinder natural movement. Generative approaches that use a deformable mesh model of the object of interest, like [2, 28, 60], are less common for animal pose estimation, because creating accurate shape models is challenging for animals. On the one hand this is grounded by the fact that creating a 3D model forces the subject to stand still for some period of time (static scene assumption) and on the other hand many types of animals have fur which is intrinsically hard to reconstruct and model. Furthermore, the com-

mon starting pose tracking from a canonical pose requirement (like the commonly used T-pose for human pose estimation) is obviously not enforceable in an animal context. There are only few examples for generative approaches [155, 156] that use offline created parametric shape models. These approaches need foreground cropped images with manual landmark annotation alongside with a good foreground segmentation and run an offline optimization to fit their models to the observations.

This makes marker-less methods especially appealing for animal pose estimation. Lately, approaches were presented that transfer methods from 2D human pose over to animal pose estimation [35, 75, 97], and subsequently solutions were proposed that estimate 3D pose by triangulating points from multiple camera observations [36, 76].

These are fundamentally different from the approach presented here, because they do not estimate a 3D pose directly but it is calculated from 2D point estimations. This includes a hard decision towards one 2D estimate for each keypoint per view and then finding a consistent 3D point to best explain the chosen 2D detections. This makes it inherently hard to account for ambiguities and introduce 3D pose priors.

5.3 METHOD

We combine a bounding box detection network, to extract the region of interest from the full scale images $\mathbf{I}_i \in \mathbb{R}^{H_1 \times W_1 \times 3}$, with a novel pose estimation architecture, see Figure 5.1. The approach resolves ambiguities after integration of information from all views. Due to occlusion, it is typically impossible from a single view to measure the exact location of all body landmarks in that view, yet existing methods attempt to predict the landmark locations in the images, regardless of their visibility. This favors learning priors, to hallucinate the invisible landmarks, over measuring their location diminishing performance in the subsequent 3D lifting step. To circumvent the problem, FreiPose extracts features $\mathbf{f}_i \in \mathbb{R}^{H_3 \times W_3 \times C}$ rather than landmarks from the cropped images $\mathbf{I}_{c,i} \in \mathbb{R}^{H_2 \times W_2 \times 3}$ and deploys a differentiable inverse projection operation Π^{-1}, which maps features into a 3D representation

$$\mathbf{F}_i = \Pi^{-1}(\mathbf{f}_i) \in R^{X \times Y \times Z \times C} \tag{5.1}$$

based on the features $\mathbf{f}_i$ of camera view i. In this notation H and W represent spatial dimensions and C the number of channels for feature representation, which throughout this work are chosen as: $H_2 = W_2 = 224$, $H_3 = W_3 = 28$ and $C = 128$. N denotes the number of keypoints, which is 12 for freely roaming rodents and 14 in the reaching experiment. Input image resolution H_1 and W_1 lies between

600 and 1280 pixels due to varying image resolutions captured by the cameras deployed. The representations across views are merged by averaging across views $\mathbf{F} = 1/N \sum_i (\mathbf{F}_i)$ and deploy a U-Net-like encoder-decoder architecture 3D CNN [23] on the voxelized representation. The 3D network learns to reason on the joint representation and predicts an initial 3D pose $\mathbf{P}^0$ incorporating information from all views.

The pose $\mathbf{P}^0 \in \mathbb{R}^{N\times 3}$ is a matrix representing the location of the N predefined body landmarks at a given time in world coordinates. Subsequently, we use $\vec{P}^0 \in \mathbb{R}^3$ as being a single keypoint sliced from $\mathbf{P}^0$, or $\mathbb{R}^4$ in its homogeneous coordinate form if needed. Similarly, $\vec{p}_i$ denotes a single 2D keypoint in $\mathbb{R}^2$ taken from $\mathbf{p}_i \in \mathbb{R}^{N\times 2}$ of camera view i.

For refinement, the initial 3D pose $\vec{P}^0$ is projected into the camera views

$$\vec{p}_i = \mathbf{K}_i \cdot \underbrace{\mathbf{M}_i \cdot \vec{P}^0}_{=:\vec{P}_i^0} \tag{5.2}$$

using the cameras' intrinsic $\mathbf{K}_i \in \mathbb{R}^{3\times 3}$ and extrinsic matrices $\mathbf{M}_i \in \mathbb{R}^{3\times 4}$, which are obtained via the camera calibration procedure.

Given the initial 2D pose $\vec{p}_i$ and image features $\mathbf{f}_i$ from view i subsequent convolutional layers estimate refined 2D coordinates $\tilde{\vec{p}}_i$. To obtain the final 3D estimate $\vec{P}^1$ the refined 2D landmarks are unprojected into the world using:

$$\vec{P}_i^1 = \vec{P}_i^1(z) \cdot \mathbf{K}_i^{-1} \cdot \tilde{\vec{p}}_i \; . \tag{5.3}$$

$\vec{P}_i^1(z)$ retrieves the third component from the pose in camera coordinates $\vec{P}_i^1$, which corresponds to the respective keypoints' depth in this cameras coordinate frame. Secondly, the scalar prediction confidence c_i is used to calculate the final estimate as a confidence weighted average:

$$\vec{P}^1 = \frac{\sum_i (\vec{P}_i^0 \cdot c_i)}{\sum_i c_i} \; . \tag{5.4}$$

Extensive details on architectural choices and algorithmic hyperparameters are located in the supplemental material of our paper [154] or can simply be taken from the released code.

5.4 EXPERIMENTS

5.4.1 *Skeletal model*

During the freely moving rat experiment we use a 12 keypoint model (Figure 5.2**a**), which includes keypoints along the body axes, faces and paws (Figure 5.2**b**).

a

#	Name
0	Ear Right
1	Eye Right
2	Paw Front Right
3	Paw Back Right Ankle
4	Paw Back Right Tip
5	Ear Left
6	Eye Left
7	Paw Front Left
8	Paw Back Left Ankle
9	Paw Back Left Tip
10	Nose
11	Tail

b

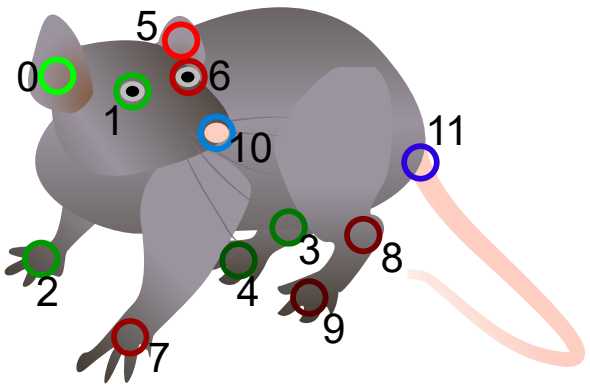

Figure 5.2: **Keypoints defined on the animals' body. a** Names and indices of keypoints. **b** Keypoint locations on the rat body.

5.4.2 *Network architecture and training*

For bounding box detection we used a COCO [70] pretrained MobileNet V2 [46], which was retrained for the task of detecting the foreground objects. In the freely moving rat scenario, it was trained using each view of the 1199 labeled time instances separately, i.e., a total of 9592 samples. We trained it for 150 k iterations using a learning rate of 0.004 and the RMSProp optimizer. As data augmentation operations, we employed random flipping, cropping, scaling, and color space variation.

For pose estimation the network was trained for 60 k using ADAM optimizer [61] with a base learning rate of 10^{-4} and decay by a factor of 0.1 every 30 k steps. To improve convergence we found it helpful to not train the refinement module for the first 30 k steps.

5.4.3 *Motion capture accuracy*

We measured the accuracy (in terms of median 3D error) and reliability (in terms of percentage of samples below a maximum error bound) of FreiPose on video recordings of freely moving rats consisting of 1813 manually labeled samples with 12 distinct body landmarks. The frames were sampled from 12 recording sessions featuring 5 different individuals (3 Long-Evans (hooded) and 2 Sprague Dawley (albino) rats). Some animals were recorded once, some on several days. We split the recordings into training and evaluation, resulting in 1199 training samples and 614 evaluation samples. Each sample contained 7 images recorded from different cameras simultaneously and a single manually annotated 3D pose, which is one 3D location for each

Number of cameras	1	2	3	4	5	6
Camera sets	{1}	{1,5}	{1,5,7}	{1,3,4,5}	{1,3,4,5,7}	{1,2,3,4,5,7}
	{5}	{1,7}	{1,3,5}	{1,3,4,7}	{2,6,4,5,7}	{1,2,3,4,5,6}
	{7}	{1,3}	{4,5,7}	{1,4,5,7}	{1,3,4,5,6}	{1,2,4,5,6,7}

Table 5.1: **Camera subsets for reduced number of views experiment.** Given a number of cameras three different subsets of cameras are selected. Reducing the amount of cameras both leaves less frames for training and increases the difficulty to precisely localize keypoints (Figure 5.3**d**). Evaluation all possible configurations is computationally very expensive so, manually selected subsets are used instead that reflect reasonable camera placements.

keypoint that is obtained from at least two manual 2D annotation in two camera views.

We trained DeepLabCut (DLC) [75], which is a popular tool for 2D landmark tracking, on the same dataset of images and applied standard triangulation methods to compute 3D poses [76]. FreiPose compares favorably in terms of the number of camera views required to reach a certain accuracy (Figure 5.3**d**), data efficiency (lower median error with the same number of labeled samples, Figure 5.3**e**), accuracy (median error of 4.54 mm vs. 7.81 mm for the full sample setting, Figure 5.3**e**), and reliability (percentage of landmarks with an error smaller than 7.5 mm is 82.8% vs. 48.1%, Figure 5.4**b**).

The experimental motion capture results of Figure 5.3 were obtained by splitting the base dataset of 1199 training and 614 evaluation frames providing 7 cameras into different subsets. To analyze the role the available number of cameras plays (Figure 5.3**d**), for a given number of cameras the experiment is run in 3 trials chosing different sets of cameras as listed in Table 5.1. For example, if only 2 cameras are used we pick the following camera pairs: $\{\{1,5\},\{1,7\},\{1,3\}\}$. Each number uniquely identifies a camera (Figure 5.3**a**) and the pairs chosen correspond to the cases 'long side + short side', 'long side + bottom' and 'long side + long side'. Each of the resulting datasets still covers 1199 time instances, but only $1199 \cdot 2 = 2398$ individual frames compared to $1199 \cdot 7 = 8393$ in the all camera setting. The same procedure is applied to the evaluation set. Table 5.1 lists the selected subsets of cameras used for experiments in Figure 5.3**d**. Please note, that testing all possible permutations is computationally very expensive, why we resort to testing manually chosen subsets representing meaningful cases, i. e., chose cameras how one would if only a limited number of cameras is available.

To simulate sparsity of labeled samples (Figure 5.3**e**) we use all cameras, but randomly select a subset of 10%, 20%, 30%, 60%, 80% or all time instances. For example, in the 20% case there are $1199 \cdot 0.2 =$

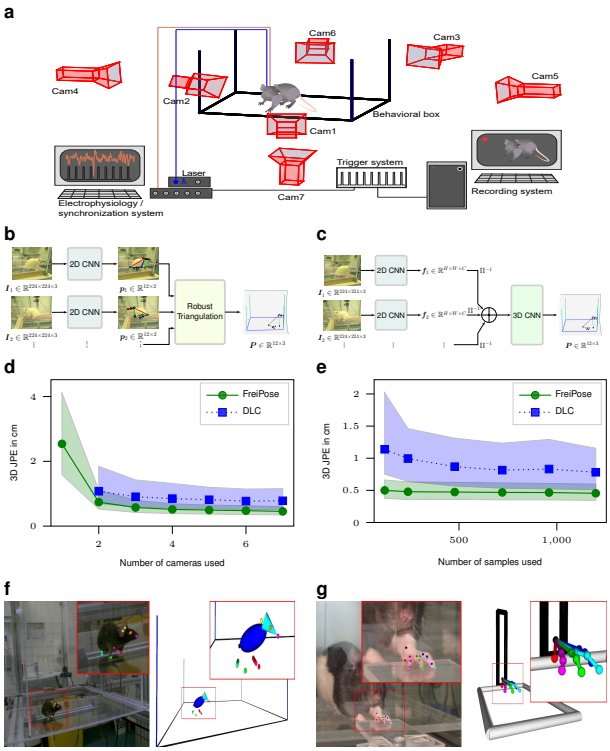

Figure 5.3: **Overview over the proposed motion capture framework and its evaluation.** **a** Motion capture setup with required hardware elements. Orange line - connection for electrophysiology, blue line - fiber optics for optogenetic stimulation. **b** State-of-the-art motion capture methods predict independent 2D poses for each view independently and subsequently calculate a 3D prediction, which requires resolving ambiguities in each view separately. **c** FreiPose accumulates evidence from all views leading to a holistic 3D prediction. Ambiguities are only resolved after information from all views is available. **d**, **e** Median of the Joint Position Error in cm for different number of cameras and number of samples compared to the DeepLabCut (DLC) et al. [75] based single view network. Shaded areas refer to the 30% and 70% percentiles. FreiPose is more data efficient and performs better regardless of the number of cameras. Largest differences can be observed for highly articulated landmarks, e. g.the front paws (see Figure 5.4). **f**, **g** FreiPose can be easily adapted towards other tasks, e. g. pose estimation of mice or paw estimation during a pellet reaching task (see Figure 5.5.)

239 time instances of 7 cameras in the training set, which results in an effective number of $239 \cdot 7 = 1673$ camera frames used. This dataset is used for training both methods, FreiPose and DLC, and evaluation is performed with respect to the complete evaluation set. Each level of sparsity is sampled 3 times for a more robust estimation.

Building on the notation introduced in Section 5.3 the median error is calculated as follows: Let $\hat{\vec{P}} \in \mathbb{R}^3$ denote the predicted keypoint

coordinate of one keypoint and $\vec{P} \in \mathbb{R}^3$ represent the label then the reported metric is defined as

$$\text{JPE} = \left\| \vec{P} - \hat{\vec{P}} \right\|_2 \tag{5.5}$$

and represents the **J**oint **P**osition **E**rror (JPE). For Figure 5.3**d**, **e** the median, 30% and 70% percentiles of the JPE are calculated over all trials, evaluation frames and keypoints.

Detailed Results. We complement the experiments with Figure 5.4, which provides JPE results on a per keypoint level in the full dataset and full camera setting. Figure 5.4**a** shows the JPE as box plots for both approaches. Largest errors are present for the highly articulated paw and tail keypoints. Compared to FreiPose the error and variance of DLC is much larger for these keypoints. Figure 5.4**b** is obtained by calculating the percentage of predictions that do not exceed a certain error threshold, which shows that FreiPose can detect keypoints much more reliably than DLC. Within an 5 mm error threshold FreiPose can detect 57.3% keypoints compared to 28.7%, which DLC can detect.

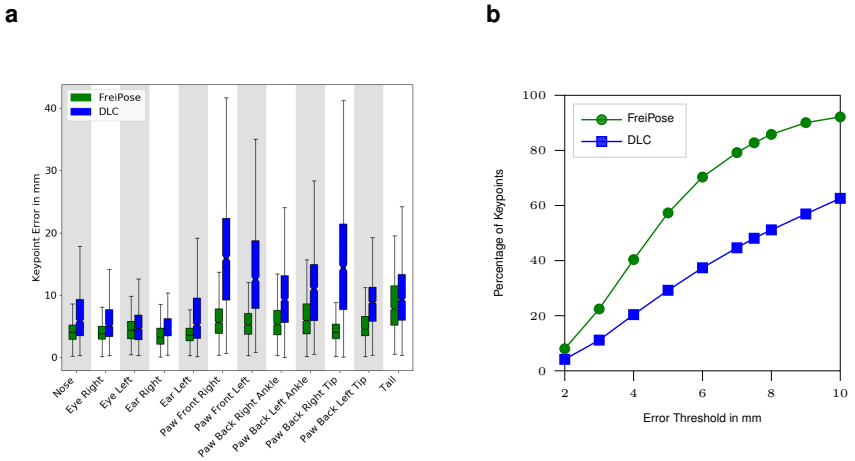

Figure 5.4: **Keypoint errors and percentage of correct keypoints for freely moving rats.** **a** Box plot of the keypoint prediction error per keypoint for FreiPose (green) and DLC (blue), where whiskers indicate 1.5 × IQR. Largest improvements between FreiPose and DLC are observed for highly articulated keypoints, e. g., paws. **b** Percentage of correct keypoints for a given threshold. For any given error tolerance FreiPose retrieves more keypoints correctly than DLC.

Qualitative Examples. Comparison of DLC and FreiPose method on a qualitative basis shows that the DLC based single view estimation plus post-hoc triangulation is prone to erroneous predictions from individual views (Figure 5.5). DLC was run with a RANSAC based triangulation method to take outlier measurements into account. Keypoint predictions with a confidence below 0.1 were discarded. The

triangulation method is part of the released code within the FreiPose Github repository. Despite these modifications, DLC's predictions were not reliably correct.

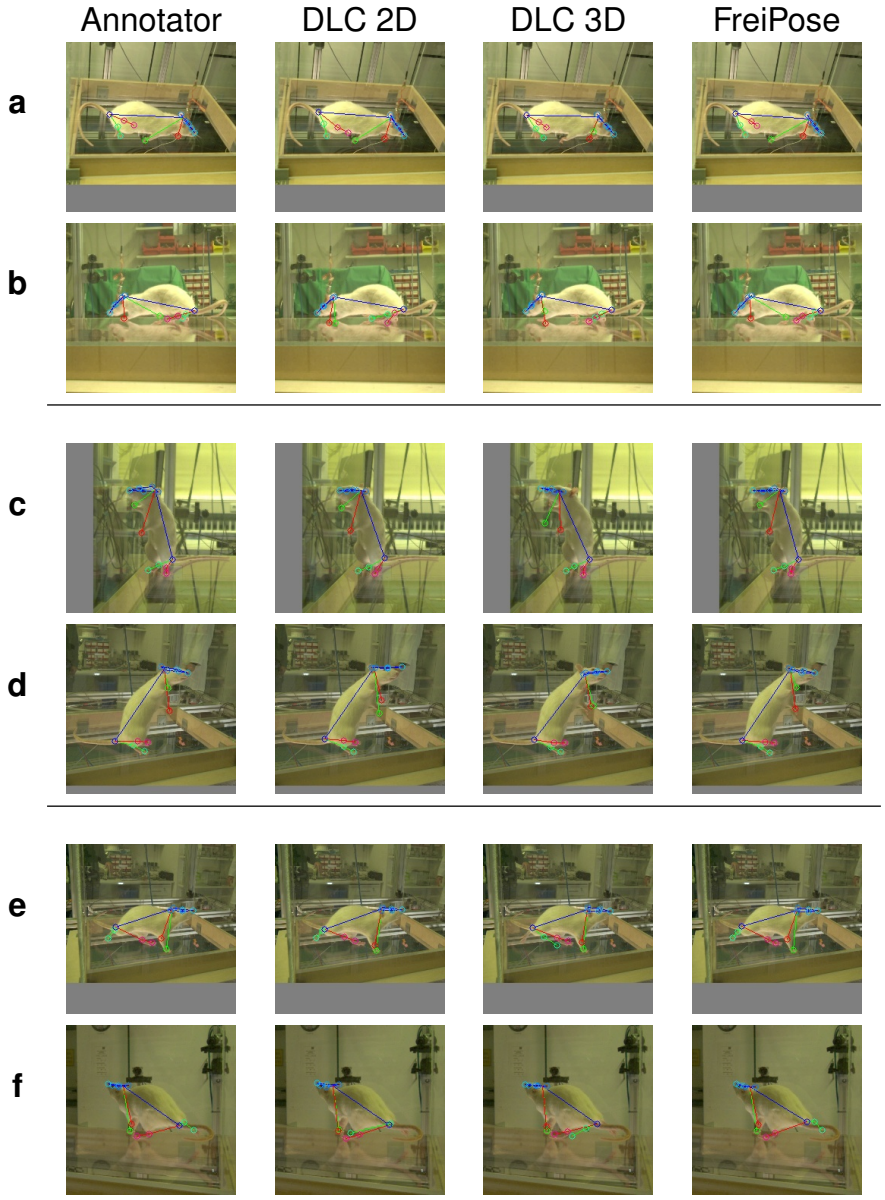

Figure 5.5: **Qualitative comparison between FreiPose and DLC.** Rows **a**, **b**, respectively **c**, **d** and **e**, **f**, show images recorded at the same time but from different cameras. DLC is able to correctly estimate poses in rows **a**, **c**, **e** but during triangulation from 2D predictions to 3D points the less accurate predictions from **b**, **d** and **f** have a deteriorating influence on the final results even though robust triangulation techniques were used. On the other side FreiPoses' predictions look visually similar to the annotations created by a human annotator.

Name	Number of factors
Keypoints in cartesian rat local frame	$12 \cdot 3 = 36$
Keypoints velocity in cartesian rat local frame	$12 \cdot 3 = 36$
Keypoints distance to rat local frame origin	12
Keypoints distance velocity to rat local frame origin	12
STFT of Keypoint distance to rat local frame origin using 33 frequencies	$12 \cdot 33 = 396$
Keypoint distance to ground plane	12
STFT of Keypoint distance to ground plane using 33 frequencies	$12 \cdot 33 = 396$
Body elevation: $\sphericalangle(g, [11, \{0, 5\}])$	1
Head elevation: $\sphericalangle(g, [10, \{0, 5\}])$	1
Head roll: $\sphericalangle(g, [0, 5])$	1
Head nick: $\sphericalangle([11, \{0, 5\}], [\{0, 5\}, 2])$	1
Paw right: $\sphericalangle([11, \{0, 5\}], [\{0, 5\}, 2])$	1
Paw left: $\sphericalangle([11, \{0, 5\}], [\{0, 5\}, 7])$	1
Angle velocity of the aforementioned angles	6
Total	912

Table 5.2: **Tested behavioral variables.** $\sphericalangle(.)$ measures the angle between two three dimensional vectors and $[a, b]$ defines an vector that goes from point a to point b. The notation $\{a, b\}$ calculates the average of the points a and b. 3D points are denoted as keypoint indices, that find their textual counterpart in Figure 5.2. The rat local coordinate frame is defined with its origin at $\{0, 5\}$, its z axis being aligned with the up pointing normal of the ground plane and its y axis rotated towards $[\{o, 5\}, 11]$. STFT - short-time Fourier transform.

5.4.4 *Quantifying the effect of optogenetic stimulation.*

Manipulations in motor cortex are likely to impact movements. This effect can be quantified via coarse measurements, e. g., rotational behavior, speed, mobile time, mobile episodes, and distance traveled [34, 73]. Recently, effects on single body parts have also been started to be investigated via video analysis [20]. To systematically investigate the effect of optogenetic stimulation in freely moving rats, we recorded the movements of 3 animals in 4 sessions, 2 sessions with 30 Hz laser burst frequency and 2 sessions with 10 Hz with a stimulation duration of 5 s, 10% duty cycle. During each recording session we stimulated each animal 5 to 7 times, with a minimum interstimulus time interval of 45 sec. We retrained FreiPose based on 136 samples from these recordings and systematically defined 908 behavioral variables for every time step from the predicted poses. Behavioral variables included transformation of the pose into a rat-aligned Cartesian coordinate frame, the distance of landmarks with respect to the ground floor as well as their velocities and Fourier transforms. Additionally, we calculated angles between body limbs with respect to each other and the direction of gravity (e. g., angle between head and body axis, see Table 5.2 for the full list of variables).

To reveal changes in behavior, we followed an *attribution-by-classification* paradigm, i.e.; given the behavioral variables at a time step t we trained a linear SVM model ($C = 0.0025$) to classify every time step into stimulated (i. e., *positive*) or not stimulated (i. e., *negative*). We trained separate classifiers for each animal, used one recording for training and left one for evaluation. The resulting classifiers achieved a balanced average accuracy of 59.1% to 73.1% on their evaluation sets. The average classifier response was able to follow the temporal dynamics of the stimulation effect and revealed an increasing effect over the course of stimulation (Figure 5.6**a**).

To attribute the effect of stimulation to individual body parts, we trained classifiers on a single variable level. A separate classifier was trained for each animal, burst frequency, and behavioral variable (Figure 5.6**b** and Figure 5.7). The rhythmic 10 Hz movements of the *Right Front Paw* was a strong indicator for the 10 Hz stimulation. The height of this paw in the rats' body reference frame was a highly correlated behavioral variable for the application of the 30 Hz laser stimulation. The pronounced effect on the *Right Front Paw* was in line with the expected outcome for the stimulation of the left motor cortex [45]. More importantly, the classifier exclusively trained on a single sequence of *Animal3* was able to generalize to another sequence of the same animal as well as to recordings of *Animal1* and *Animal2* indicating that FreiPose performs robustly across sessions and animals. Thus, FreiPose allows a detailed comparison of stimulation effects across animals without the need to retrain for individual cases (Figure 5.7).

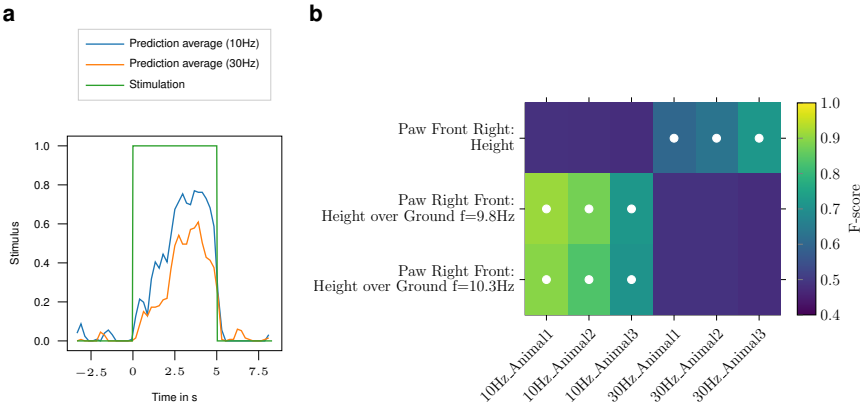

Figure 5.6: **Automatic evaluation of the optogenetic stimulation effect via FreiPose. a** Automatic detection of the effect during optogenetic stimulation with temporal resolution. The classifier predicted whether a time step was stimulated '1' or not '0' based o the behavioral variables calculated from FreiPose. Shown is the averaged predicted stimulus of the classifier model within temporally aligned windows across stimulation trials on a recording which was withheld for evaluation. The stimulation spans from 0 sec to 5 sec and the prediction scores trend indicates an increasingly visible effect over time. **b** Attribution of the stimulation effect to individual body parts. We trained an ensemble of classifiers to distinguish (not) stimulated frames given only a single behavior variable as input to each classifier. Analyzing the resulting classifiers allowed to distinguish important factors from less important ones. Shown is the *F-score* of the respective classifier and white dots indicate significance below a *p-value* of 0.001 (Bonferroni adjusted) supported by the *chi2* test between predicted and actual classes. The classifiers shown here were exclusively trained on *Animal3*, but generalize across animals. Other configurations are shown in Figure 5.7.

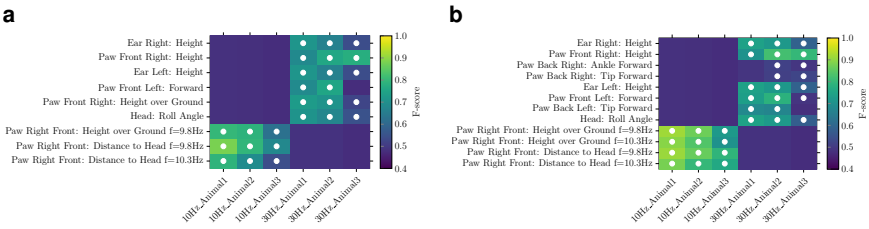

Figure 5.7: **Permutations of the optogenetic stimulation experiment.** Similar findings during automatic attribution of the simulation effect when training was performed on Animal1 (**a**) or Animal2 (**b**). Shown is the *F-score* of the respective classifier and white dots indicate significance below a *p-value* of 0.001 (Bonferroni adjusted) supported by the *chi2* test between predicted and actual classes.

5.5 CONCLUSION

Here we present the marker-free, deep learning based motion capture tool FreiPose for holistic 3D tracking of individual body parts

and pose reconstruction of freely moving animals. Instead of triangulating 2D pose estimates, FreiPose directly reconstructs body poses and movement trajectories in 3D resulting in unprecedented precision. Analyzing the problem holistically by fusing information from all views into a joint 3D predictions allows us to surpass the state of the art in pose estimation of freely moving rats.

5.6 FOLLOW-UP WORK

Estimating keypoints of articulated objects is still a very active field of research. Most works still follow estimating keypoints from 2D and triangulating to 3D paradigm [108, 142, 148], and potentially combine it with dense surface depth estimation [142, 148] to fit shape models to the surface.

Where applicable also still marker-based systems are applied, e. g., Kearneyet al. [59] work on pose estimation for dogs and use dedicated motion suits in conjunction with a commercial capture system to extract 3D ground-truth skeletons.

Concurrent to our work is the approach by Iskakov et al. [52], who proposed a very similar holistic 3D estimation approach. They show strong results and achieve state of the art in human pose estimation from multiple views in a supervised training setting.

Great impact is to be expected by releasing our approach embedded as a toolkit to the research community. It is suitable to investigate sorts of biological questions. One early example is Eriksson et al. [21] where FreiPose is used to track paw trajectories and quantify paw movement.

6

ESTIMATING SHAPE FROM MULTI-VIEW RGB

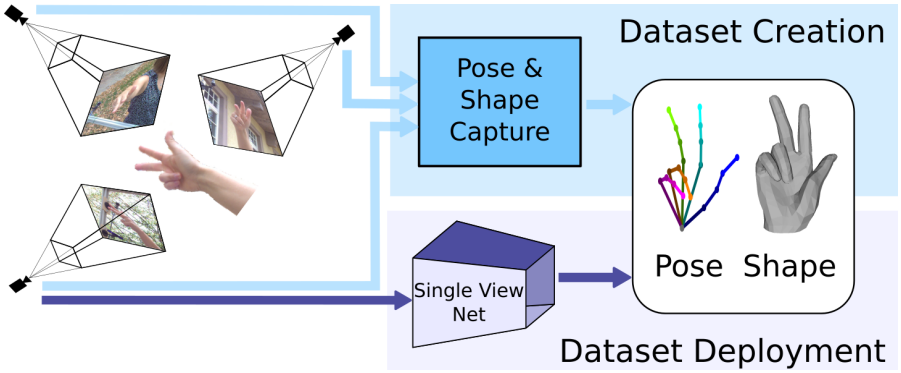

Figure 6.1: **Overview of creation and deployment of the dataset.** We create a hand dataset via a novel iterative procedure that utilizes multiple views and sparse annotation followed by verification. This results in a large scale real-world dataset with pose and shape labels, which can be used to train single-view networks that have superior cross-dataset generalization performance on pose and shape estimation.

This chapter describes ideas and experiments that were previously presented in the following work; therefore, copyright lies with © 2019 IEEE.

FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images
Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus and Thomas Brox
IEEE/CVF International Conference on Computer Vision (ICCV), 2019

> *This work presents a semi-automated human-in-the-loop approach, which includes hand fitting optimization to infer both the 3D pose and shape for multi-view samples. The approach is used to create a large-scale dataset, which shows superior cross-dataset generalization.*

The author of this thesis recorded the dataset, developed the fitting optimization and conducted all experiments. All co-authors contributed to the project discussions as well as writing the publication.

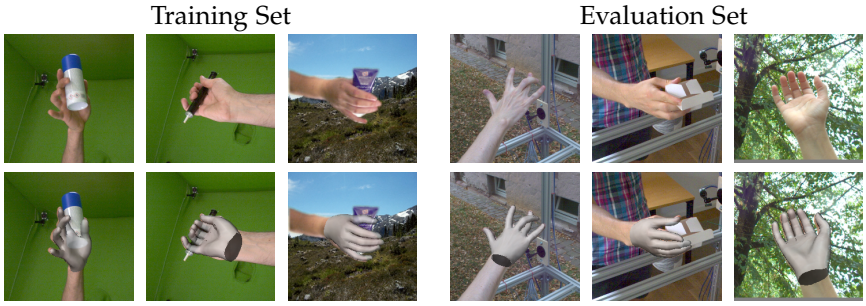

Figure 6.2: **Qualitative examples from the FreiHAND dataset.** Examples from our proposed dataset showing images (top row) and hand shape annotations (bottom row). The training set contains composited images from green screen recordings, whereas the evaluation set contains images recorded indoors and outdoors. The dataset features several subjects as well as object interactions.

6.1 INTRODUCTION

3D hand pose and shape estimation from a single RGB image has a variety of applications in gesture recognition, robotics, and AR. Various deep learning methods have approached this problem, but the quality of their results depends on the availability of training data. Such data is created either by rendering synthetic datasets [10, 31, 87, 88, 151] or by capturing real datasets under controlled settings typically with little variation [32, 112, 129]. Both approaches have limitations, discussed in our related work section.

Synthetic datasets use deformable hand models with texture information and render this model under varying pose configurations. As with all rendered datasets, it is difficult to model the wide set of characteristics of real images, such as varying illumination, camera lens distortion, motion blur, depth of field and debayering. Even more importantly, rendering of hands requires samples from the true distribution of feasible and realistic hand poses. In contrast to human pose, such distributional data does not exist to the same extent. Consequently, synthetic datasets are either limited in the variety of poses or sample many unrealistic poses.

Capturing a dataset of real human hands requires annotation in a post-processing stage. In single images, manual annotation is difficult and cannot be easily crowd sourced due to occlusions and ambiguities. Moreover, collecting and annotating a large scale dataset is a respectable effort.

In this paper, we analyze how these limitations affect the ability of single-view hand pose estimation to generalize across datasets and to *in-the-wild* real application scenarios. We find that datasets show excellent performance on the respective evaluation split, but have rather poor performance on other datasets, i.e., we see a classical dataset bias.

As a remedy to the dataset bias problem, we created a new large-scale dataset by increasing variation between samples. We collect a real-world dataset and develop a methodology that allows us to automate large parts of the labeling procedure, while manually ensuring very high-fidelity annotations of 3D pose and 3D hand shape. One of the key aspects is that we record synchronized images from multiple views, an idea already used previously in [7, 112]. The multiple views remove many ambiguities and ease both the manual annotation and automated fitting. The second key aspect of our approach is a semi-automated *human-in-the-loop* labeling procedure with a strong bootstrapping component. Starting from a sparse set of 2D keypoint annotations (e.g., finger tip annotations) and semi-automatically generated segmentation masks, we propose a hand fitting method that fits a deformable hand model [102] to a set of multi-view input. This fitting yields both 3D hand pose and shape annotation for each view. We then train a multi-view 3D hand pose estimation network using these annotations. This network predicts the 3D hand pose for unlabeled samples in our dataset along with a confidence measure. By verifying confident predictions and annotating least-confident samples in an iterative procedure, we acquire 11592 annotations with moderate manual effort by a human annotator.

The dataset spans 32 different people and features fully articulated hand shapes, a high variation in hand poses and also includes interaction with objects. Part of the dataset, which we mark as training set, is captured against a green screen. Thus, samples can easily be composed with varying background images. The test set consists of recordings in different indoor and outdoor environments; see Figure 6.2 for sample images and the corresponding annotation.

Training on this dataset clearly improves cross-dataset generalization compared to training on existing datasets. Moreover, we are able to train a network for full 3D hand shape estimation from a single RGB image. For this task, there is not yet any publicly available data, neither for training nor for benchmarking. Our dataset is available on our project page and therefore can serve both as training and benchmarking dataset for future research in this field.

6.2 RELATED WORK

Since datasets are crucial for the success of 3D hand pose and shape estimation, there has been much effort on acquiring such data.

In the context of hand shape estimation, the majority of methods fall into the category of model-based techniques. These approaches were developed in a strictly controlled environment and utilize either depth data directly [121, 122, 128] or use multi-view stereo methods for reconstruction [7]. More related to our work are approaches that fit statistical human shape models to observations [9, 67] from *in-the-*

train \ eval	STB	RHD	GAN	PAN	LSMV	FPA	HO-3D	**Ours**	Average Rank
STB [144]	**0.783**	0.179	0.067	0.141	0.072	0.061	0.138	0.138	6.0
RHD [151]	0.362	**0.767**	0.184	0.463	0.544	0.101	0.450	0.508	2.9
GAN [87]	0.110	0.103	**0.765**	0.092	0.206	0.180	0.087	0.183	5.4
PAN [53]	0.459	0.316	0.136	**0.870**	0.320	0.184	0.351	0.407	3.0
LSMV [32]	0.086	0.209	0.152	0.189	**0.717**	0.129	0.251	0.276	4.1
FPA [30]	0.119	0.095	0.084	0.120	0.118	**0.777**	0.106	0.163	6.0
HO-3D [38]	0.154	0.130	0.091	0.111	0.149	0.073	-	0.169	6.1
Ours	0.473	0.518	0.217	0.562	0.537	0.128	**0.557**	**0.678**	2.2

Table 6.1: **Quantitative evaluation of cross-dataset generalization.** This table shows cross-dataset generalization measured as area under the curve (AUC) of percentage of correct keypoints following [151]. Each row represents the training set used and each column the evaluation set. The last column shows the average rank each training set achieved across the different evaluation sets. The top-three ranking training sets for each evaluation set are marked as follows: **first**, second or third. Note that the evaluation set of *HO-3D* was not available at time of submission, therefore one table entry is missing and the other entries within the respective column report numbers calculated on the training set.

wild color images as input. Such methods require semi-automatic methods to acquire annotations such as keypoints or segmentation masks for each input image to guide the fitting process.

Historically, acquisition methods often incorporated markers onto the hand that allow for an easy way to estimate pose. Common choices are infrared markers [44], color coded gloves [132], or electrical sensing equipment [150]. This alters hand appearance and, hence, makes the data less valuable for training discriminative methods.

Annotations can also be provided manually on hand images [87, 117, 144]. However, the annotation is limited to visible regions of the hand. Thus, either the subject is required to retain from complex hand poses that result in severe self-occlusions, or only a subset of hand joints can be annotated.

To avoid occlusions and annotate data at larger scale, Simon et al. [112] leveraged a multi-view recording setup. They proposed an iterative bootstrapping approach to detect hand keypoints in each view and triangulate them to generate 3D point hypotheses. While the spirit of our data collection strategy is similar, we directly incorporate the multi-view information into a neural network for predicting 3D keypoints and our dataset consists of both pose and shape annotations.

Since capturing real data comes with an expensive annotation setup and process, more methods rather deployed synthetic datasets recently [87, 151].

6.3 ANALYSIS OF EXISTING DATASETS

We thoroughly analyze state-of-the-art datasets used for 3D hand pose estimation from single RGB images by testing their ability to generalize to unseen data. We identify seven state-of-the-art datasets that provide samples in the form of an RGB image and the accompanying 3D keypoint information as shown in Table 6.2.

6.3.1 *Considered Datasets*

Stereo Tracking Benchmark (**STB**) [144] dataset is one of the first and most commonly used datasets to report performance of 3D keypoint estimation from a single RGB image. The annotations are acquired manually limiting the setup to hand poses where most regions of the hands are visible. Thus, the dataset shows a unique subject posing in a frontal pose with different background scenarios and without objects.

The **Panoptic** (**PAN**) dataset [53] was created using a dense multi-view capture setup consisting of 10 RGB-D sensors, 480 VGA and 31 HD cameras. It shows humans performing different tasks and interacting with each other. There are 83 sequences publicy available and 12 of them have hand annotation. We select *171204_pose3* to serve as evaluation set and use the remaining 11 sequences from the *range motion, haggling and tools* categories for training.

Garcia et al. [30] proposed the **First-person hand action benchmark** (**FPA**), a large dataset that is recorded from an egocentric perspective and annotated using magnetic sensors attached to the finger tips of the subjects. Wires run along the fingers of the subject altering the appearance of the hands significantly. 6 DOF sensor measurements are utilized in an inverse kinematics optimization of a given hand model to acquire the full hand pose annotations.

Using the commercial Leap Motion device [86] for keypoint annotation, Gomez et al. [32] proposed the **Large-scale Multiview 3D Hand Pose Dataset** (**LSMV**). Annotations given by the device are transformed into 4 calibrated cameras that are approximately time synchronized. Due to the limitations of the sensor device, this dataset does not show any hand-object interactions.

The **Rendered Hand Pose Dataset** (**RHD**) proposed by Zimmermann et al. [151] is a synthetic dataset rendered from 20 characters performing 31 different actions in front of a random background image without hand object interaction.

Building on the SynthHands [87] dataset Mueller et al. [88] presented the **GANerated** (**GAN**) dataset. SynthHands was created by retargeting measured human hand articulation to a rigged meshed model in a mixed reality approach. This allowed for hand object interaction to some extend, because the subject could see the rendered

dataset	num. frames	num. subjects	real	obj-ects	shape	labels
STB [144]	15 k / 3 k	1	✓	✗	✗	manual
PAN [53]	641 k / 34 k	> 10	✓	✓	✗	MVBS [112]
FPA [30]	52 k / 53 k	6	✓	✓	✗	marker
LSMV [32]	117 k / 31 k	21	✓	✗	✗	leapmotion
RHD [151]	41 k / 2.7 k	20	✗	✗	✗	synthetic
GAN [88]	266 k / 66 k	-	✗	✓	✗	synthetic
HO-3D [38]	11 k / -	3	✓	✓	✓	automatic [38]
Ours	33 k / 4 k	32	✓	✓	✓	hybrid

Table 6.2: **State-of-the-art datasets for the task of 3D keypoint estimation from a single color image used in our analysis.** We report dataset size in number of frames, number of subjects, if it is real or rendered data, regarding hand object interaction, if shape annotation is provided and which method was used for label generation.

scene in real time and pose the hand accordingly. In the following *GANerated* hand dataset, a CycleGAN approach is used to bridge the synthetic to real domain shift.

Recently, Hampali et al. [38] proposed an algorithm for dataset creation deploying an elaborate optimization scheme incorporating temporal and physical consistencies, as well as silhouette and depth information. The resulting dataset is referred to as **HO-3D**.

6.3.2 *Evaluation Setup*

We trained a state-of-the-art network architecture [50] that takes as input an RGB image and predicts 3D keypoints on the training split of each of the datasets and report its performance on the evaluation split of all other datasets. For each dataset, we either use the standard training/evaluation split reported by the authors or create an 80%/20% split otherwise.

The single-view network takes an RGB image $\mathbf{I}$ as input and infers 3D hand pose $\mathbf{P} = \{\vec{P}_k\}$ with each $\vec{P}_k \in \mathbb{R}^3$, representing a predefined landmark or keypoint situated on the kinematic skeleton of a human hand. Due to scale ambiguity, the problem to estimate real-world 3D keypoint coordinates in a camera centered coordinate frame is ill-posed. Hence, we adopt the problem formulation of [50] to estimate coordinates in a root relative and scale normalized fashion:

$$\vec{P}_k = s \cdot \hat{\vec{P}}_k = s \cdot \begin{pmatrix} \hat{X}_k \\ \hat{Y}_k \\ \hat{Z}_k \end{pmatrix} = \begin{pmatrix} \hat{X}_k \\ \hat{Y}_k \\ \hat{Z}_k^{\text{rel}} + \hat{Z}^{\text{root}} \end{pmatrix}, \tag{6.1}$$

where the normalization factor s is chosen as the length of one reference bone in the hand skeleton, $\hat{Z}^{\text{root}}$ is the root depth and $\hat{Z}_k^{\text{rel}}$ the

relative depth of keypoint k. We define the resulting 2.5D representation as:

$$\hat{\tilde{\mathbf{P}}}_{\text{rel}_k} = \left(\hat{X}_k,\ \hat{Y}_k,\ \hat{Z}_k^{\text{rel}}\right)^T. \tag{6.2}$$

Given scale constraints and 2D projections of the points in a calibrated camera, 3D hand pose $\mathbf{P}$ can be recovered from $\hat{\mathbf{P}}_{\text{rel}}$. For details about this procedure we refer to [50].

We train the single-view network using the same hyperparameter choices as Iqbal et al. [50]. However, we use only a single stage and reduce the number of channels in the network layers, which leads to a significant speedup in terms of training time at only a marginal decrease in accuracy. We apply standard choices of data augmentation including color, scale and translation augmentation as well as rotation around the optical axis. We apply this augmentation to each of the datasets.

6.3.3 *Results*

It is expected that the network performs the best on the dataset it was trained on, yet it should also provide reasonable predictions for unseen data when being trained on a dataset with sufficient variation (e.g., hand pose, viewpoint, shape, existence of objects, etc.).

Table 6.1 shows for each existing training dataset the network is able to generalize to the respective evaluation split and reaches the best results there. On the other hand, performance drops substantially when the network is tested on other datasets.

Both *GAN* and *FPA* datasets appear to be especially hard to generalize indicating that their data distribution is significantly different from the other datasets. For *FPA* this stems from the appearance change due to the markers used for annotation purposes. The altered appearance gives the network trained on this dataset strong cues to solve the task that are not present for other datasets at evaluation time. Thus, the network trained on *FPA* performs poorly when tested on other datasets. Based on visual inspection of the *GAN* dataset, we hypothesize that subtle changes like missing hand texture and different color distribution are the main reasons for generalization problems. We also observe that while the network trained on *STB* does not perform well on remaining datasets, the networks trained on other datasets show reasonable performance on the evaluation split of *STB*. We conclude that a good performance on *STB* is not a reliable measure for how a method generalizes to unseen data.

Based on the performance of each network, we compute a cumulative ranking score for each dataset that we report in the last column of Table 6.1. To calculate the cumulative rank we assign ranks for each

column of the table separately according to the performance the respective training sets achieve. The cumulative rank is then calculated as average over all evaluation sets, i. e., rows of the table. Based on these observations, we conclude that there is a need for a new benchmarking dataset that can provide superior generalization capability.

We present the FreiHAND Dataset to archieve this goal. It consists of real images, provides sufficient viewpoint and hand pose variation, and shows samples both with and without object interactions. Consequently, the single-view network trained on this dataset achieves a substantial improvement in terms of ranking for cross-dataset generalization. We next describe how we acquired and annotated this dataset.

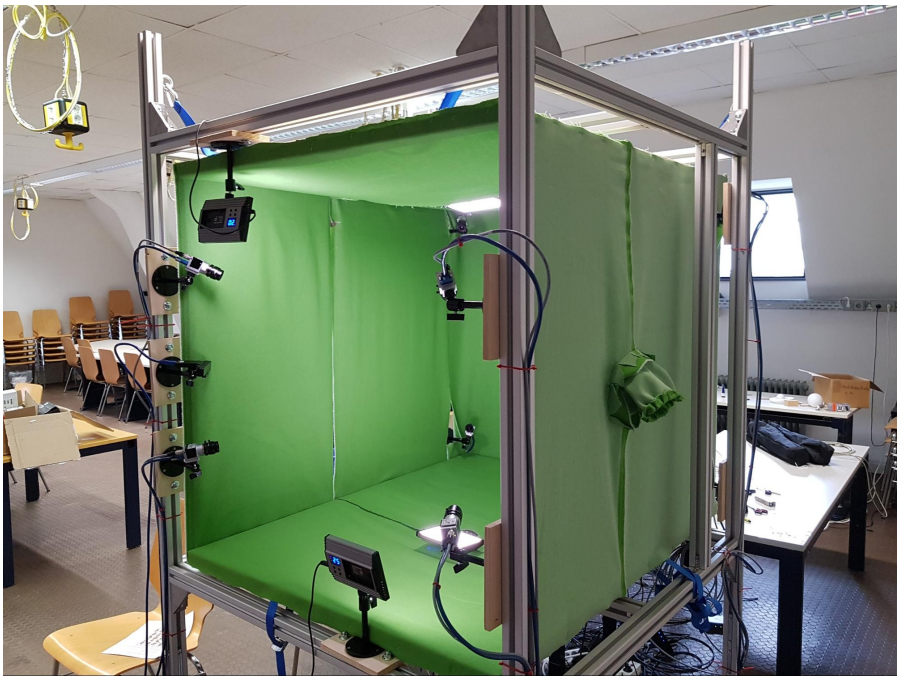

Figure 6.3: **The Recording setup used.** It contains 8 calibrated and temporally synchronized RGB cameras located at the corners of a cube. A green screen background can be mounted into the the setup, enabling easier background subtraction.

6.4 FREIHAND DATASET

The dataset was captured with the multi-view setup shown in Figure 6.3. The setup is portable enabling both indoor and outdoor capture. We capture hand poses from 32 subjects of different genders and ethnic backgrounds. Each subject is asked to perform actions with and without objects. To capture hand-object interactions, subjects are given a number of everyday household items that allow for reasonable one-handed manipulation and are asked to demonstrate different grasping techniques.

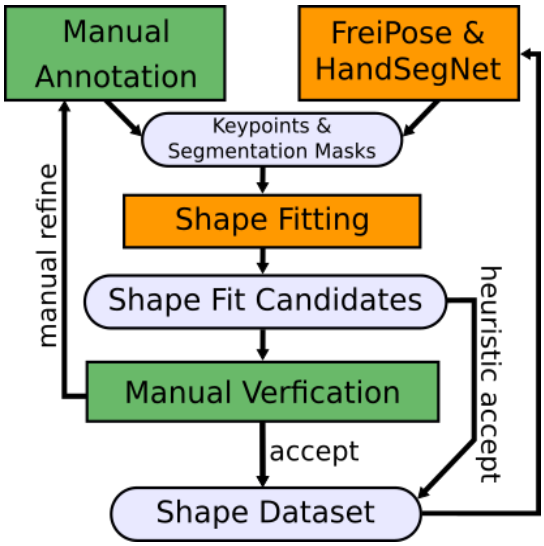

Figure 6.4: **Workflow overview.** The dataset labeling starts from manual annotation followed by the shape fitting process described in Section 6.4.1, which yields candidate shape fits for our data samples. Sample fits are manually verified allowing them to be accepted, rejected or queued for further annotation. Alternatively a heuristic can accept samples without human interaction. The initial dataset allows for training the networks involved, which for subsequent iterations of the procedure, can predict information needed for fitting. The labeling process can be bootstrapped, allowing more accepted samples to accumulate in the dataset.

To preserve the realistic appearance of hands, no markers are used during the capture. Instead we resort to post-processing methods that generate 3D labels. Manual acquisition of 3D annotations is obviously unfeasible. An alternative strategy is to acquire 2D keypoint annotations for each input view and utilize the multi-view camera setup to lift such annotations to 3D similar to Simon et al. [112].

We found after initial experiments that current 2D hand pose estimation methods perform poorly, especially in case of challenging hand poses with self- and object occlusions. Manually annotating all 2D keypoints for each view is prohibitively expensive for large-scale data collection. Annotating all 21 keypoints across multiple-views with a specialized tool takes about 15 minutes for each multi-view set. Furthermore, keypoint annotation alone is not sufficient to obtain shape information.

We address this problem with a novel bootstrapping procedure (see Figure 6.4) composed of a set of automatic methods that utilize sparse 2D annotations. Since our data is captured against a green screen, the foreground can be extracted automatically. Refinement is needed only to co-align the segmentation mask with the hand model's wrist. In addition, a sparse set of six 2D keypoints (finger tips and wrist) is manually annotated. These annotations are relatively cheap to acquire at a reasonably high quality. For example, manually correcting

a segmentation mask takes on average 12 seconds, whereas annotating a keypoint takes around 2 seconds. Utilizing this information we fit a deformable hand model to multi-view images using a novel fitting process described in Section 6.4.1. This yields candidates for both 3D hand pose and shape labels. These candidates are then manually verified, before being added to a set of labels.

Given an initial set of labels, we train our proposed network, *FreiPose*, that takes as inputs multi-view images and predicts 3D keypoint locations along with a confidence score, described in Section 6.4.2. Keypoint predictions can be used in lieu of manually annotated keypoints as input for the fitting process. This bootstrapping procedure is iterated. The least-confident samples are manually annotated (Section 6.4.3). With this *human-in-the-loop* process, we quickly obtain a large scale annotated dataset. Next we describe each stage of this procedure in detail.

6.4.1 *Hand Model Fitting with Sparse Annotations*

Our goal is to fit a deformable hand shape model to observations from multiple views acquired at the same time. We build on the statistical *MANO* model, proposed by Romero et al. [102], which is parameterized by $\theta \in \mathbb{R}^{61}$. The model parameters $\theta = (\alpha, \beta, \gamma)^T$ include shape $\alpha \in \mathbb{R}^{10}$, articulation $\beta \in \mathbb{R}^{45}$ as well as global translation and orientation $\gamma \in \mathbb{R}^{6}$. Using keypoint and segmentation information we optimize a multi-term loss,

$$\mathcal{L} = \mathcal{L}_{\text{kp}}^{\text{2D}} + \mathcal{L}_{\text{kp}}^{\text{3D}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{pose}}, \tag{6.3}$$

to estimate the model parameters $\tilde{\theta}$, where the tilde indicates variables that are being optimized. We describe each of the terms in (6.3) next.

2D Keypoint Loss $\mathcal{L}_{\text{kp}}^{\text{2D}}$: The loss is the sum of distances between the 2D projection Π^i of the models' 3D keypoints $\vec{\tilde{P}}_k \in \mathbb{R}^3$ to the 2D annotations $\vec{q}_k^{\,i}$ over views i and visible keypoints $k \in V_i$:

$$\mathcal{L}_{\text{kp}}^{\text{2D}} = w_{\text{kp}}^{\text{2D}} \cdot \sum_i \sum_{k \in V_i} \left\| \vec{q}_k^{\,i} - \Pi^i(\vec{\tilde{P}}_k) \right\|_2 . \tag{6.4}$$

3D keypoint Loss $\mathcal{L}_{\text{kp}}^{\text{3D}}$: This loss is defined in a similar manner as (6.4), but over 3D keypoints. Here, $\vec{P}_k$ denotes the 3D keypoint annotations, whenever such annotations are available (i. e., if predicted by *FreiPose*),

$$\mathcal{L}_{\text{kp}}^{\text{3D}} = w_{\text{kp}}^{\text{3D}} \cdot \sum_{i \in V} \left\| \vec{P}_k - \vec{\tilde{P}}_k \right\|_2 . \tag{6.5}$$

Segmentation Loss $\mathcal{L}_{\text{seg}}$: For shape optimization we use a sum of l_2 losses between the model dependent mask $\tilde{\mathbf{M}}^i$ and the manual annotation $\mathbf{M}^i$ over views i:

$$\mathcal{L}_{\text{seg}} = w_{\text{seg}} \cdot \sum_i (\left\|\mathbf{M}^i - \tilde{\mathbf{M}}^i\right\|_2 + \left\|\text{EDT}(\mathbf{M}^i) \cdot \tilde{\mathbf{M}}^i\right\|_2). \tag{6.6}$$

Additionally, we apply a silhouette term based on the Euclidean Distance Transform (EDT). Specifically, we apply a symmetric EDT to $\mathbf{M}^i$, which contains the distance to the closest boundary pixel at every location.

Shape Prior $\mathcal{L}_{\text{shape}}$: For shape regularization we employ

$$\mathcal{L}_{\text{shape}} = w_{\text{shape}} \cdot \left\|\tilde{\boldsymbol{\beta}}\right\|_2, \tag{6.7}$$

which enforces the predicted shape to stay close to the mean shape of *MANO*.

Pose Prior $\mathcal{L}_{\text{pose}}$: The pose prior has two terms. The first term applies a regularization on the PCA coefficients a_j used to represent the pose $\tilde{\boldsymbol{\alpha}}$ in terms of PCA basis vectors $\mathbf{c}_j$ (i.e., $\tilde{\boldsymbol{\alpha}} = \sum_j \tilde{a}_j \cdot \mathbf{c}_j$). This regularization enforces predicted poses to stay close to likely poses with respect to the PCA pose space of *MANO*. The second term regularizes the distance of the current pose $\tilde{\boldsymbol{\alpha}}$, to the N nearest neighbors of a hand pose dataset acquired from [30]:

$$\mathcal{L}_{\text{pose}} = w_{\text{pose}} \cdot \sum_j \left\|\tilde{a}_j\right\|_2 + w_{\text{nn}} \cdot \sum_{n \in N} \left\|\boldsymbol{\alpha}^n - \tilde{\boldsymbol{\alpha}}\right\|_2. \tag{6.8}$$

We implement the fitting process in Tensorflow [1] and use *MANO* to implement a differentiable mapping from $\tilde{\theta}$ to 3D model keypoints $\tilde{\mathbf{p}}_k$ and 3D model vertex locations $\tilde{\mathbf{V}} \in \mathbb{R}^{778\times 3}$. We adopt the Neural Renderer [58] to render the segmentation masks $\tilde{\mathbf{M}}^i$ from the hand model vertices $\tilde{\mathbf{V}}$ and use the ADAM optimizer [61] to minimize:

$$\theta = \arg\min_{\tilde{\theta}}(\mathcal{L}(\tilde{\subseteq})) \tag{6.9}$$

6.4.2 *Multiview 3D Keypoint Estimation*

To automate the fitting process, we seek to estimate 3D keypoints automatically. We propose to use *FreiPose*, as presented in Chapter 5, shown in Figure 6.5 that aggregates information from all eight camera images $\mathbf{I}_i$ and predicts a single hand pose $\mathbf{P} = \{\vec{P}_1, \ldots, \vec{P}_J\}$.

6.4.3 *Iterative Refinement*

In order to generate annotations at large scale, we propose an iterative, *human-in-the-loop* procedure which is visualized in Figure 6.4.

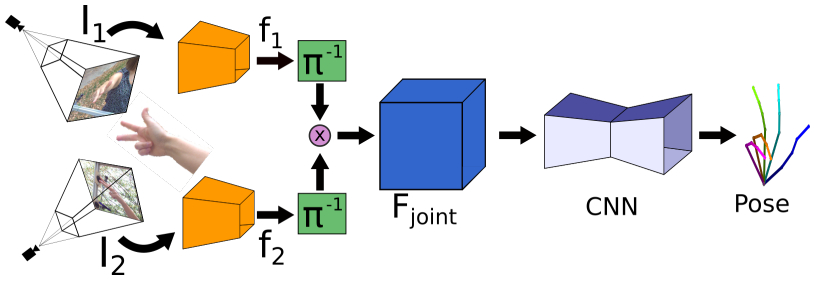

Figure 6.5: **Approach for prediction of keypoints in a multi-view setting.** *FreiPose* predicts a single hand pose $\mathbf{P}$ using images of all 8 views (for simplicity only 2 are shown). Each image is processed separately by a 2D CNN that is shared across views. This yields 2D feature maps $\mathbf{f}_i$. These are individually reprojected into a common coordinate frame using the known camera calibration to obtain $\mathbf{F}_i = \Pi^{-1}(\mathbf{f}_i)$. The $\mathbf{F}_i$ are aggregated over all views and finally a 3D CNN localizes the 3D keypoints within a voxel representation.

Method	mesh error ↓	F@5mm ↑	F@15mm ↑
Mean shape	1.63	0.340	0.839
MANO Fit	1.44	0.416	0.880
MANO CNN	**1.07**	**0.529**	**0.935**
Boukhayma et al. [10]	1.30	0.435	0.898
Hasson et al. [41]	1.32	0.436	0.908

Table 6.3: **Quantitative evaluation of single view shape estimation.** This table shows shape prediction performance on the evaluation split of FreiHAND after alignment. We report two measures: The mean mesh error and the F-score at two different distance thresholds.

For initial bootstrapping we use a set of manual annotations to generate the initial dataset $\mathcal{D}_0$. In iteration i we use dataset $\mathcal{D}_i$, a set of images and the corresponding *MANO* fits, to train *FreiPose* and *HandSegNet* [151]. *FreiPose* makes 3D keypoint predictions along with confidence scores for the remaining unlabeled data and *HandSegNet* predicts hand segmentation masks. Using these predictions, we perform the hand shape fitting process of Section 6.4.1. Subsequently, we perform *verification* that either accepts, rejects or partially annotates some of these data samples.

Heuristic Verification. We define a heuristic consisting of three criteria to identify data samples with good *MANO* fits. First, we require the mean *FreiPose* confidence score to be above 0.8 and all individual keypoint confidences to be at least 0.6, which enforces a minimum level of certainty on the 3D keypoint prediction. Second, we define a minimum threshold for the intersection over union (IoU) between predicted segmentation mask and the mask derived from the *MANO* fitting result. We set this threshold to be 0.7 on average across all views while also rejecting samples that have more than 2

views with an IoU below 0.5. Third, we require the mean Euclidean distance between predicted 3D keypoints and the keypoints of the fitted MANO to be at most 0.5 cm where no individual keypoint has a Euclidean distance greater than 1 cm. We accept only samples that satisfy all three criteria and add these to the set $\mathcal{D}_i^h$.

Manual Verification and Annotation. The remaining unaccepted samples are sorted based on the confidence score of *FreiPose* and we select samples from the 50^{th} percentile upwards. We enforce a minimal temporal distance between samples selected to ensure diversity as well as choosing samples for which the current pose estimates are sufficiently different to a flat hand shape as measured by the Euclidean distance in the pose parameters. We ask the annotators to evaluate the quality of the *MANO* fits for these samples. Any sample that is verified as a good fit is added to the set $\mathcal{D}_i^m$. For remaining samples, the annotator has the option of either discarding the sample or provide additional annotations (e.g., annotating mislabeled finger tips) to help improve the fit. These additionally annotated samples are added to the set $\mathcal{D}_i^l$.

Joining the samples from all streams yields a larger labeled dataset

$$\mathcal{D}_{i+1} = \mathcal{D}_i + \mathcal{D}_i^h + \mathcal{D}_i^m + \mathcal{D}_i^l \tag{6.10}$$

which allows us to retrain both *HandSegNet* and *FreiPose*. We repeated this process 4 times to obtain our final dataset.

6.5 EXPERIMENTS

6.5.1 *Cross-Dataset Generalization of FreiHAND*

To evaluate the cross-dataset generalization capability of our dataset and to compare to the results of Table 6.1, we define the following training and evaluation split: there are samples with and without green screen and we chose to use all green screen recordings for training and the remainder for evaluation. Training and evaluation splits contain data from 24 and 11 subjects, respectively, with only 3 subjects shared across splits. The evaluation split is captured in 2 different indoor and 1 outdoor location. We augmented the training set by leveraging the green screen for easy and effective background subtraction and creating composite images using new backgrounds. To avoid green color bleeding at the hand boundaries we applied the image harmonization method of Tsai et al. [127] and the deep image colorization approach by Zhang et al. [145] separately to our data. Both the automatic and sampling variant of [145] were used. With the original samples this quadruples the training set size from 33 k unique to 132 k augmented samples. Examples of resulting images are shown in Figure 6.2.

Given the training and evaluation split, we train the single view 3D pose estimation network on our data and test it across different datasets. As shown in Table 6.1, the network achieves strong accuracy across all datasets and ranks first in terms of cross-dataset generalization.

6.5.2 *3D Shape Estimation*

Having both pose and shape annotations, our acquired dataset can be used for training shape estimation models in a fully supervised way. In addition, it serves as the first real dataset that can be utilized for evaluating shape estimation methods. Building on the approach of Kanazawa et al. [56], we train a network that takes as input a single RGB image and predicts the *MANO* parameters $\tilde{\theta}$ using the following loss:

$$\mathcal{L} = w_{3D} \left\| \mathbf{P}_k - \tilde{\mathbf{P}}_k \right\|_2 + w_{2D} \left\| \Pi(\mathbf{P}_k) - \Pi(\tilde{\mathbf{P}}) \right\|_2 + w_p \left\| \theta - \tilde{\theta} \right\|_2 . \tag{6.11}$$

We deploy l_2 losses for 2D and 3D keypoints as well as the model parameters and chose the weighting to $w_{3D} = 1000$, $w_{2D} = 10$ and $w_p = 1$.

We also provide two baseline methods, constant mean shape prediction, without accounting for articulation changes, and fits of the *MANO* model to the 3D keypoints predicted by our single-view network.

For comparison, we use two scores. The *mesh error* measures the average Euclidean distance between corresponding mesh vertices in the ground truth and the predicted hand shape. We also evaluate the F-score [62] which, given a distance threshold, defines the harmonic mean between recall and precision between two sets of points [62]. In our evaluation, we use two distances: F@5mm and F@15mm to report the accuracy both at fine and coarse scale. In order to decouple shape evaluation from global rotation and translation, we first align the predicted meshes using Procrustes alignment. Results are summarized in Table 6.3. Estimating *MANO* parameters directly with a CNN performs better across all measures than the baseline methods. The evaluation reveals that the difference in F-score is more pronounced in the high accuracy regime. Qualitative results of our network predictions are provided in Figure 6.6.

6.5.3 *Evaluation of Iterative Labeling*

In the first step of iterative labeling process, we set $w_{kp}^{2D} = 100$ and $w_{kp}^{2D} = 0$ (since no 3D keypoint annotations are available), $w_{seg} = 10.0$, $w_{shape} = 100.0$, $w_{nn} = 10.0$, and $w_{pose} = 0.1$. (For subsequent

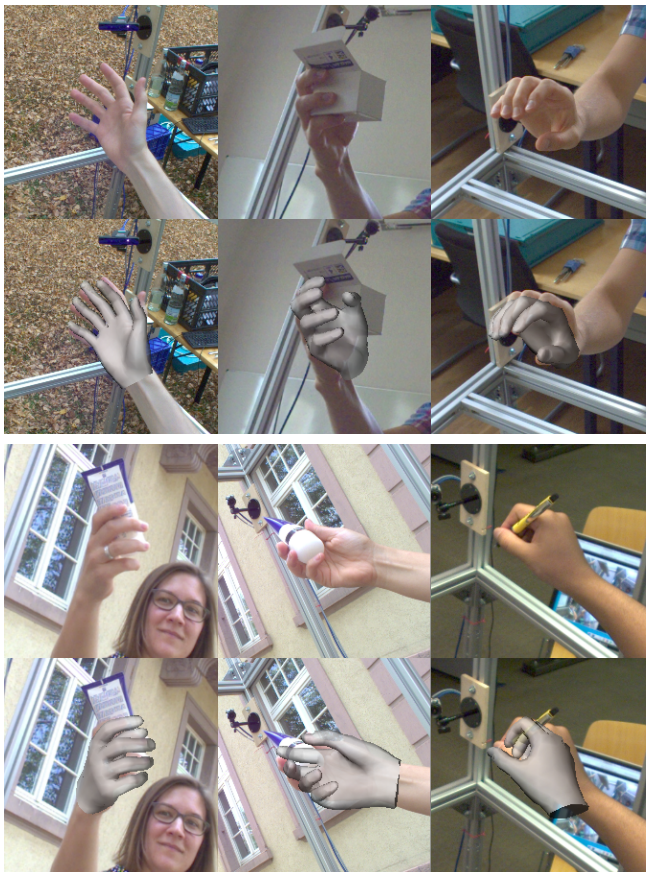

Figure 6.6: **Quantitative results for shape estimation from a single view.** Given a single RGB image as input (top rows), qualitative results of predicted hand shapes (bottom rows) are shown. Please note that we don't apply any alignment of the predictions with respect to the ground truth.

iterations we set $w_{\text{kp}}^{\text{2D}} = 50$ and $w_{\text{kp}}^{\text{3D}} = 1000$.) Given the fitting results, we train *FreiPose* and test it on the remaining dataset. After the first verification step, 302 samples are accepted. Validating a sample takes about 5 seconds and we find that the global pose is captured correctly in most cases, but in order to obtain high quality ground truth, even fits with minor inaccuracies are discarded.

We use the additional accepted samples to retrain *FreiPose* and *HandSegNet* and iterate the process. At the end of the first iteration we are able to increase the dataset to 993 samples, 140 of which are automatically accepted by heuristic, and the remainder from verifying 1000 samples. In the second iteration the total dataset size increases to 1449, 289 of which are automatically accepted and the remainder stems from verifying 500 samples. In subsequent iterations the complete dataset size is increased to 2609 and 4565 samples, where heuristic accept yields 347 and 210 samples respectively. This is the dataset we use for the cross-dataset generalization (see Table 6.1) and shape estimation (see Table 6.3) experiments.

Dataset	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$
#samples	302	993	1449	2609	4565
RHD	0.244	0.453	0.493	0.511	0.518
PAN	0.347	0.521	0.521	0.539	0.562

Table 6.4: **FreiHAND dataset over iterations.** Bootstrapping convergence is evaluated by reporting cross-dataset generalization to *RHD* and *PAN*. The measure of performance is AUC, which shows monotonous improvement throughout.

We evaluate the effectiveness of the iterative labeling process by training a single view 3D keypoint estimation network on different iterations of our dataset. For this purpose, we chose two evaluation datasets that reached a good average rank in Table 6.1. Table 6.4 reports the results and shows a steady increase for both iterations as our dataset grows.

6.6 CONCLUSION

We present FreiHAND, a large RGB dataset with hand pose and shape labels of real images. We capture this dataset using a novel iterative procedure. The dataset allows us improve generalization performance for the task of 3D hand pose estimation from a single image, as well as supervised learning of monocular hand shape estimation.

To facilitate research on hand shape estimation, we extended our dataset towards a challenging benchmark that takes the community a big step towards evaluation under realistic *in-the-wild* conditions.

6.7 FOLLOW-UP WORK

Due to its good generalization properties, the proposed dataset *FreiHAND* was used in numerous works either for training or evaluation of approaches: Yang et al. [139] trained their approach using the proposed dataset to learn hand shape estimation of image sequences. Kulon et al. [65] worked in a weakly-supervised setting where hand shapes were fitted to crowd-sourced videos and our dataset was used for evaluation. In Spurr et al. [114] 3D hand pose estimation was learned from only 2D annotations and additional biomechanical constraints using *FreiHAND* for training and evaluation.

In the meantime similar datasets were presented using comparable techniques: Hampali et al. [39] used a single RGBD sensor and jointly optimized hand and object poses in a sequence-wise manner. In [142, 148] similar multi-view setups are built, but many more cameras are used, which allows for dense surface reconstruction.

Part III

EPILOGUE

7 CONCLUSION

There are many tasks humans can solve with their vision system at ease, and estimating the pose of a known object is one at which they excel, but why are we so good at it?

It is believed that the reason is probably in how humans combine the information from their vision system with their prior knowledge. The effective use of priors describing structure and the ability to reason about plausibility are powerful tools that we are all using without even realizing it. We know what a hand looks like in general and how it can move anatomically. There is some understanding what feasible grasps look like and how these are conditioned on the object being manipulated, while taking into account functionality of the grasp.

The goal of this thesis is to bring perception algorithms closer to the human level of performance. In its different projects similar strategies like the use of prior knowledge and integrating it into the structure of algorithms were successfully pursued.

Results from Chapter 3 showed that hand pose estimation from a color image is possible and that introducing additional structure into the task, i. e., separating viewpoint and articulation, eases the learning problem significantly. It also shows that using synthetic data for this task is beneficial.

This thesis also brought robots a little bit closer towards the human level of learning: namely understanding demonstrations of new tasks from vision alone, see Chapter 4. Observing a teacher perform the task in question multiple times proved to be sufficient for imitation of the action demonstrated.

Understanding animal motion from a single camera turned out to be infeasible due to frequent occlusions. Thus, in Chapter 5 multiple cameras were deployed. The approach developed integrates camera geometry in a clear and concise way. This imposes a strong structure onto the algorithm, helping it to learn faster and more precisely than approaches that do not incorporate this information. The resulting algorithm was used for animal motion capture and allowed quantifying behavior during biological experiments.

Adding more structure was also leveraged in Chapter 6, where the strong prior of a parametric hand shape model was used to fit the hand shape from noisy predictions of hand joints and information about the hand silhouette. This allowed creating a real-world dataset with superior generalization properties and learning shape estimation from a single view.

In summary, this thesis introduces multiple new approaches for marker-less pose and shape prediction tasks, which present small improvements over previous works. In the course of their development datasets were created and made available to the community enabling future research in this area.

8

OUTLOOK

While this thesis shows encouraging steps towards better perception systems, there is still plenty of room for future improvement. In my view there are some directions that future research should focus on.

Weaker supervision. In the supervised setting training on a larger dataset usually leads to better results. This phenomenon is commonly explained by the fact that neural networks have a very large number of parameters that are optimized and more training samples help finding a better, i. e., more general solution. Unfortunately, creating labeled data is costly and therefore naturally limited, which makes exploring methods that do not need full supervision increasingly attractive. A possible future direction is to leverage consistencies between different data sources to learn. For example one could use multiple observations across different camera views and learn from the constraint that they show the same pose, but seen from different angles. Another possibility could leverage the agreement of two simultaneously recorded data modalities, for example the color and depth image. These concepts provide a much weaker supervision signal than explicit labels, but if performed at scale it could help tremendously in pushing algorithms to work robustly and with few failures. Initial steps into this direction were already taken [12, 99, 100, 131], but here is still much room for further improvement left.

Occluded Keypoints. Current approaches are usually trained to learn a mapping from image to pose space from annotated image pose pair samples. There is no notion if a keypoint is visible in an image or not, if there is an annotation it has to be predicted similarly to every other point. This creates an imbalance between points that are visible in the image and those that are occluded. This is because for visible points information can be extracted from the image. Whereas, the missing information for occluded points must be compensated for with learned priors, which can be highly specific to the respective dataset and therefore hinder generalization. Future work should address this issue and find a concise way to handle this problem. Possible approaches could incorporate a measure of uncertainty or to allow multiple hypothesis for occluded keypoints. The first attempt in this direction was taken by Ye et al. [141], but they only show performance below state-of-the-art on depth datasets that do not contain object interaction.

Tracking. In the past, approaches used a model that was tracked over multiple frames. Each frame being initialized from the last one and the initial estimate being handled separately. Currently, most research focuses on approaches that allow making predictions from only a single frame. While these might be scientifically more interesting, many applications for pose estimation allow leveraging temporal smoothness in motion over the course of a video stream. There is plenty of room in practical applications to mix both approaches and combine the discriminative power of CNNs to extract information from single frames and complement it with a model enforcing temporal smoothness. The work by Müller et al. [89] shows promising results using such an hybrid approach, but they only show results using a single RGBD camera and can not yet handle the presence of objects.

Interaction. Another direction for further research is to make the task harder by introducing additional degrees of freedom. One possibility is to focus on the interaction of two tightly interacting hands. Another possibility is to jointly estimate pose of hands and interacting objects. The setting of two hands is very difficult because to estimate pose it is not only necessary to disambiguate fingers from each other, but also to assign fingers to their correct hand. This results into a much more difficult learning task, which is set in a regime where acquiring training data is more difficult. Motion for synthetic datasets is hard to animate when two hands are interacting and human annotation is also much more difficult to obtain. Introducing objects into the task suffers from similar problems. Annotated data is harder to obtain and the hand is occluded much more frequently when interacting closely with the object. Also, a new degree of variation emerges: how to represent different objects in a unified way that allows pose estimation to generalize to unseen objects. This will make creating training datasets much more difficult and require new algorithms to be developed that account for physical plausibility and functionality of a grasp. Hasson et al. [41] presents an initial attempt to deal with a priori known objects. Müller et al. [89] and Tzionas et al. [129] can track two interacting hands from RGBD input.

BIBLIOGRAPHY

[1] Martín Abadi et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." In: *CoRR* abs/1603.04467 (2016). arXiv: 1603.04467. URL: http://arxiv.org/abs/1603.04467 (cit. on pp. 21, 35, 67).

[2] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. "Performance capture from sparse multi-view video." In: *ACM Trans. Graph.* 27.3 (2008), p. 98. DOI: 10.1145/1360612.1360697. URL: https://doi.org/10.1145/1360612.1360697 (cit. on p. 44).

[3] Irene Albrecht, Jörg Haber, and Hans-Peter Seidel. "Construction and animation of anatomically based human hand models." In: *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, San Diego, CA, USA, July 26-27, 2003*. Ed. by Rick Parent, Karan Singh, David E. Breen, and Ming C. Lin. The Eurographics Association, 2003, pp. 98–109. DOI: 10.2312/SCA03/098-109. URL: https://doi.org/10.2312/SCA03/098-109.

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 3686–3693. DOI: 10.1109/CVPR.2014.471. URL: https://doi.org/10.1109/CVPR.2014.471 (cit. on pp. 16, 33, 34).

[5] Vassilis Athitsos and Stan Sclaroff. "Estimating 3D Hand Pose from a Cluttered Image." In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*. IEEE Computer Society, 2003, pp. 432–442. DOI: 10.1109/CVPR.2003.1211500. URL: https://doi.org/10.1109/CVPR.2003.1211500 (cit. on pp. 12, 17, 18).

[6] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. "Weakly-Supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6121–6131.

[7] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. "Motion Capture of Hands in Action Using Discriminative Salient Points." In: *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, Oc-*

tober 7-13, 2012, Proceedings, Part VI. Ed. by Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Vol. 7577. Lecture Notes in Computer Science. Springer, 2012, pp. 640–653. DOI: 10.1007/978-3-642-33783-3_46. URL: https://doi.org/10.1007/978-3-642-33783-3_46 (cit. on p. 59).

[8] Abhijat Biswas, Henny Admoni, and Aaron Steinfeld. "Fast on-board 3D torso pose recovery and forecasting." In: *Proceedings of the International Conference on Robotics and Automation (ICRA'19)*. 2019, pp. 20–24 (cit. on p. 41).

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. "Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image." In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9909. Lecture Notes in Computer Science. Springer, 2016, pp. 561–578. DOI: 10.1007/978-3-319-46454-1_34. URL: https://doi.org/10.1007/978-3-319-46454-1_34 (cit. on p. 59).

[10] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. "3D Hand Shape and Pose From Images in the Wild." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 10843–10852. DOI: 10.1109/CVPR.2019.01110. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Boukhayma_3D_Hand_Shape_and_Pose_From_Images_in_the_Wild_CVPR_2019_paper.html (cit. on pp. 11, 58, 68).

[11] Koen Buys, Cedric Cagniart, Anatoly Baksheev, Tinne De Laet, Joris De Schutter, and Caroline Pantofaru. "An adaptable system for RGB-D based human body detection and pose estimation." In: *J. Vis. Commun. Image Represent.* 25.1 (2014), pp. 39–52. DOI: 10.1016/j.jvcir.2013.03.011. URL: https://doi.org/10.1016/j.jvcir.2013.03.011 (cit. on p. 33).

[12] Y. Cai, L. Ge, J. Cai, N. Magnenat-Thalmann, and J. Yuan. "3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB Images." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1 (cit. on p. 77).

[13] Sylvain Calinon, Zhibin Li, Tohid Alizadeh, Nikos G. Tsagarakis, and Darwin G. Caldwell. "Statistical dynamical systems for skills acquisition in humanoids." In: *12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), Osaka, Japan, November 29 - Dec. 1, 2012*. IEEE, 2012, pp. 323–329. DOI:

10.1109/HUMANOIDS.2012.6651539. URL: https://doi.org/10.1109/HUMANOIDS.2012.6651539.

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1302–1310. DOI: 10.1109/CVPR.2017.143. URL: https://doi.org/10.1109/CVPR.2017.143 (cit. on pp. 3, 11, 12, 33, 34, 36).

[15] Ching-Hang Chen and Deva Ramanan. "3D Human Pose Estimation = 2D Pose Estimation + Matching." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5759–5767. DOI: 10.1109/CVPR.2017.610. URL: https://doi.org/10.1109/CVPR.2017.610 (cit. on pp. 17, 27).

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.4 (2018), pp. 834–848. DOI: 10.1109/TPAMI.2017.2699184. URL: https://doi.org/10.1109/TPAMI.2017.2699184 (cit. on p. 3).

[17] Xianjie Chen and Alan L. Yuille. "Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations." In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger. 2014, pp. 1736–1744. URL: http://papers.nips.cc/paper/5291-articulated-pose-estimation-by-a-graphical-model-with-image-dependent-pairwise-relations (cit. on p. 17).

[18] Paul Doliotis, Vassilis Athitsos, Dimitrios I. Kosmopoulos, and Stavros J. Perantonis. "Hand Shape and 3D Pose Estimation Using Depth Data from a Single Cluttered Frame." In: *Advances in Visual Computing - 8th International Symposium, ISVC 2012, Rethymnon, Crete, Greece, July 16-18, 2012, Revised Selected Papers, Part I*. Ed. by George Bebis et al. Vol. 7431. Lecture Notes in Computer Science. Springer, 2012, pp. 148–158. DOI: 10.1007/978-3-642-33179-4_15. URL: https://doi.org/10.1007/978-3-642-33179-4_15 (cit. on p. 12).

[19] Philippe Dreuw, Thomas Deselaers, Daniel Keysers, and Hermann Ney. "Modeling Image Variability in Appearance-Based Gesture Recognition." In: *ECCV Workshop on Statistical Methods*

in Multi-Image and Video Processing. Graz, Austria, May 2006, pp. 7–18 (cit. on pp. 28, 29).

[20] Teppei Ebina et al. "Arm movements induced by noninvasive optogenetic stimulation of the motor cortex in the common marmoset." In: *Proceedings of the National Academy of Sciences* 116.45 (2019), pp. 22844–22850. DOI: 10.1073/pnas.1903445116 (cit. on p. 53).

[21] David Eriksson, Mona Heiland, Artur Schneider, and Ilka Diester. "Cortical activity at different time scales: high-pass filtering separates motor planning and execution." In: *bioRxiv* (2020). DOI: 10.1101/857300. eprint: https://www.biorxiv.org/content/early/2020/06/06/857300.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/06/857300 (cit. on p. 55).

[22] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. "Vision-based hand pose estimation: A review." In: *Comput. Vis. Image Underst.* 108.1-2 (2007), pp. 52–73. DOI: 10.1016/j.cviu.2006.10.012. URL: https://doi.org/10.1016/j.cviu.2006.10.012.

[23] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. "U-Net: deep learning for cell counting, detection, and morphometry." In: *Nature Methods* 16.1 (2019), p. 67 (cit. on p. 46).

[24] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. "Pictorial Structures for Object Recognition." In: *Int. J. Comput. Vis.* 61.1 (2005), pp. 55–79. DOI: 10.1023/B:VISI.0000042934.15159.49. URL: https://doi.org/10.1023/B:VISI.0000042934.15159.49.

[25] Martin A. Fischler and Robert A. Elschlager. "The Representation and Matching of Pictorial Structures." In: *IEEE Trans. Computers* 22.1 (1973), pp. 67–92. DOI: 10.1109/T-C.1973.223602. URL: https://doi.org/10.1109/T-C.1973.223602.

[26] Blender Foundation. *Free and Open 3D Creation Software*. URL: http://www.blender.org (cit. on p. 22).

[27] Oren Freifeld, Alexander Weiss, Silvia Zuffi, and Michael J. Black. "Contour people: A parameterized model of 2D articulated human shape." In: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. IEEE Computer Society, 2010, pp. 639–646. DOI: 10.1109/CVPR.2010.5540154. URL: https://doi.org/10.1109/CVPR.2010.5540154.

[28] Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. "Motion capture using joint skeleton tracking and surface estimation." In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 1746–1753. DOI: 10.1109/CVPR.2009.5206755. URL: https://doi.org/10.1109/CVPR.2009.5206755 (cit. on pp. 11, 44).

[29] Xiao-Shan Gao, Xiaorong Hou, Jianliang Tang, and Hang-Fei Cheng. "Complete Solution Classification for the Perspective-Three-Point Problem." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 25.8 (2003), pp. 930–943. DOI: 10.1109/TPAMI.2003.1217599. URL: https://doi.org/10.1109/TPAMI.2003.1217599 (cit. on p. 10).

[30] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. "First-Person Hand Action Benchmark With RGB-D Videos and 3D Hand Pose Annotations." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 409–419. DOI: 10.1109/CVPR.2018.00050. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Garcia-Hernando_First-Person_Hand_Action_CVPR_2018_paper.html (cit. on pp. 60, 61, 62, 67).

[31] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. "3D Hand Shape and Pose Estimation From a Single RGB Image." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 10833–10842. DOI: 10.1109/CVPR.2019.01109. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Ge_3D_Hand_Shape_and_Pose_Estimation_From_a_Single_RGB_CVPR_2019_paper.html (cit. on p. 58).

[32] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. "Large-scale multiview 3D hand pose dataset." In: *Image Vis. Comput.* 81 (2019), pp. 25–33. DOI: 10.1016/j.imavis.2018.12.001. URL: https://doi.org/10.1016/j.imavis.2018.12.001 (cit. on pp. 58, 60, 61, 62).

[33] Claire C Gordon, Cynthia L Blackwell, Bruce Bradtmiller, Joseph L Parham, Patricia Barrientos, Stephen P Paquette, Brian D Corner, Jeremy M Carson, Joseph C Venezia, Belva M Rockwell, et al. *2012 Anthropometric survey of US Army personnel: Methods and summary statistics*. Tech. rep. Army Natick Soldier Research Development and Engineering Center MA, 2014.

[34] Viviana Gradinaru, Murtaza Mogri, Kimberly R. Thompson, Jaimie M. Henderson, and Karl Deisseroth. "Optical deconstruction of parkinsonian neural circuitry." In: *Science* 324.5925 (2009), pp. 354–359. DOI: 10.1126/science.1167093 (cit. on p. 53).

[35] Jacob M. Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R. Costelloe, and Iain D. Couzin. "DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning." In: *eLife* 8 (2019). DOI: 10.7554/eLife.47994 (cit. on p. 45).

[36] Semih Günel, Helge Rhodin, Daniel Morales, João Campagnolo, Pavan Ramdya, and Pascal Fua. "DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult Drosophila." In: *eLife* 8 (2019). DOI: 10.7554/eLife.48571 (cit. on p. 45).

[37] Yao Guo, Fani Deligianni, Xiao Gu, and Guang-Zhong Yang. "3-D Canonical Pose Estimation and Abnormal Gait Recognition With a Single RGB-D Camera." In: *IEEE Robotics Autom. Lett.* 4.4 (2019), pp. 3617–3624. DOI: 10.1109/LRA.2019.2928775. URL: https://doi.org/10.1109/LRA.2019.2928775 (cit. on p. 41).

[38] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. "HO-3D: A Multi-User, Multi-Object Dataset for Joint 3D Hand-Object Pose Estimation." In: *CoRR* abs/1907.01481 (2019). arXiv: 1907.01481. URL: http://arxiv.org/abs/1907.01481 (cit. on pp. 60, 62).

[39] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. "HOnnotate: A Method for 3D Annotation of Hand and Object Poses." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3196–3206 (cit. on p. 72).

[40] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Fei-Fei Li. "Towards Viewpoint Invariant 3D Human Pose Estimation." In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9905. Lecture Notes in Computer Science. Springer, 2016, pp. 160–177. DOI: 10.1007/978-3-319-46448-0_10. URL: https://doi.org/10.1007/978-3-319-46448-0_10 (cit. on p. 33).

[41] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. "Learning Joint Reconstruction of Hands and Manipulated Objects." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer

Vision Foundation / IEEE, 2019, pp. 11807–11816. DOI: 10.1109/CVPR.2019.01208. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Hasson_Learning_Joint_Reconstruction_of_Hands_and_Manipulated_Objects_CVPR_2019_paper.html (cit. on pp. 11, 29, 68, 78).

[42] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. "Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 571–580.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. URL: https://doi.org/10.1109/CVPR.2016.90 (cit. on p. 3).

[44] Gerrit Hillebrand, Martin Bauer, Kurt Achatz, Gudrun Klinker, and Am Oferl. "Inverse kinematic infrared optical finger tracking." In: *Proceedings of the 9th International Conference on Humans and Computers (HC 2006), Aizu, Japan*. Citeseer. 2006, pp. 6–9 (cit. on p. 60).

[45] Riichiro Hira, Shin-Ichiro Terada, Masashi Kondo, and Masanori Matsuzaki. "Distinct Functional Modules for Discrete and Rhythmic Forelimb Movements in the Mouse Motor Cortex." In: *The Journal of Neuroscience* 35.39 (2015), pp. 13311–13322. DOI: 10.1523/JNEUROSCI.2731-15.2015 (cit. on p. 53).

[46] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." In: *CoRR* abs/1704.04861 (2017). arXiv: 1704.04861. URL: http://arxiv.org/abs/1704.04861 (cit. on p. 47).

[47] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243. URL: https://doi.org/10.1109/CVPR.2017.243 (cit. on p. 35).

[48] James Steven Supancic III, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. "Depth-Based Hand Pose Estimation: Data, Methods, and Challenges." In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1868–1876.

DOI: 10.1109/ICCV.2015.217. URL: https://doi.org/10.1109/ICCV.2015.217 (cit. on p. 12).

[49] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1647–1655. DOI: 10.1109/CVPR.2017.179. URL: https://doi.org/10.1109/CVPR.2017.179 (cit. on p. 3).

[50] Umar Iqbal. "Articulated Human Pose Estimation in Unconstrained Images and Videos." PhD thesis. University of Bonn, Germany, 2018. URL: http://hss.ulb.uni-bonn.de/2018/5292/5292.htm (cit. on pp. 62, 63).

[51] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. "Hand Pose Estimation via Latent 2.5D Heatmap Regression." In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11215. Lecture Notes in Computer Science. Springer, 2018, pp. 125–143. DOI: 10.1007/978-3-030-01252-6_8. URL: https://doi.org/10.1007/978-3-030-01252-6_8 (cit. on pp. 12, 29).

[52] Karim Iskakov, Egor Burkov, Victor S. Lempitsky, and Yury Malkov. "Learnable Triangulation of Human Pose." In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 7717–7726. DOI: 10.1109/ICCV.2019.00781. URL: https://doi.org/10.1109/ICCV.2019.00781 (cit. on p. 55).

[53] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. "Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 8320–8329. DOI: 10.1109/CVPR.2018.00868. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Joo_Total_Capture_A_CVPR_2018_paper.html (cit. on pp. 60, 61, 62).

[54] Hanbyul Joo et al. "Panoptic Studio: A Massively Multiview System for Social Interaction Capture." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.1 (2019), pp. 190–204. DOI: 10.1109/TPAMI.2017.2782743. URL: https://doi.org/10.1109/TPAMI.2017.2782743.

[55] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. "Random tree walk toward instantaneous 3D human pose estimation." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE

Computer Society, 2015, pp. 2467–2474. DOI: 10.1109/CVPR.2015.7298861. URL: https://doi.org/10.1109/CVPR.2015.7298861 (cit. on p. 33).

[56] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. "End-to-End Recovery of Human Shape and Pose." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 7122–7131. DOI: 10.1109/CVPR.2018.00744. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Kanazawa_End-to-End_Recovery_of_CVPR_2018_paper.html (cit. on pp. 11, 70).

[57] Abhishek Kar, Christian Häne, and Jitendra Malik. "Learning a Multi-View Stereo Machine." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 365–376. URL: http://papers.nips.cc/paper/6640-learning-a-multi-view-stereo-machine.

[58] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. "Neural 3D Mesh Renderer." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3907–3916. DOI: 10.1109/CVPR.2018.00411. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Kato_Neural_3D_Mesh_CVPR_2018_paper.html (cit. on p. 67).

[59] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. "RGBD-Dog: Predicting Canine Pose from RGBD Sensors." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8336–8345 (cit. on p. 55).

[60] Roland Kehl and Luc Van Gool. "Markerless tracking of complex human motions from multiple views." In: *Comput. Vis. Image Underst.* 104.2-3 (2006), pp. 190–209. DOI: 10.1016/j.cviu.2006.07.010. URL: https://doi.org/10.1016/j.cviu.2006.07.010 (cit. on p. 44).

[61] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980 (cit. on pp. 21, 35, 47, 67).

[62] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. "Tanks and temples: benchmarking large-scale scene reconstruction." In: *ACM Trans. Graph.* 36.4 (2017), 78:1–78:13. DOI: 10.1145/3072959.3073599. URL: https://doi.org/10.1145/3072959.3073599 (cit. on p. 70).

[63] Marina Kollmitz, Andreas Eitel, Andres Vasquez, and Wolfram Burgard. "Deep 3D perception of people and their mobility aids." In: *Robotics Auton. Syst.* 114 (2019), pp. 29–40. DOI: 10.1016/j.robot.2019.01.011. URL: https://doi.org/10.1016/j.robot.2019.01.011 (cit. on p. 41).

[64] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. "PifPaf: Composite Fields for Human Pose Estimation." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 11977–11986. DOI: 10.1109/CVPR.2019.01225. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Kreiss_PifPaf_Composite_Fields_for_Human_Pose_Estimation_CVPR_2019_paper.html (cit. on p. 12).

[65] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. "Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4990–5000 (cit. on p. 72).

[66] Martin de La Gorce, David J. Fleet, and Nikos Paragios. "Model-Based 3D Hand Pose Estimation from Monocular Video." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.9 (2011), pp. 1793–1805. DOI: 10.1109/TPAMI.2011.33. URL: https://doi.org/10.1109/TPAMI.2011.33 (cit. on p. 12).

[67] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. "Unite the People: Closing the Loop Between 3D and 2D Human Representations." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 4704–4713. DOI: 10.1109/CVPR.2017.500. URL: https://doi.org/10.1109/CVPR.2017.500 (cit. on p. 59).

[68] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. "EPnP: An Accurate $O(n)$ Solution to the PnP Problem." In: *Int. J. Comput. Vis.* 81.2 (2009), pp. 155–166. DOI: 10.1007/s11263-008-0152-6. URL: https://doi.org/10.1007/s11263-008-0152-6 (cit. on p. 10).

[69] John Y. Lin, Ying Wu, and Thomas S. Huang. "Modeling the Constraints of Human Hand Motion." In: *Workshop on Human Motion, HUMO 2000, Austin, Texas, USA, December 7-8, 2000, Proceedings*. IEEE Computer Society, 2000, pp. 121–126. DOI: 10.1109/HUMO.2000.897381. URL: https://doi.org/10.1109/HUMO.2000.897381 (cit. on p. 6).

[70] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context." In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Ed. by David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars. Vol. 8693. Lecture Notes in Computer Science. Springer, 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48. URL: https://doi.org/10.1007/978-3-319-10602-1_48 (cit. on pp. 33, 34, 36, 47).

[71] Timm Linder, Kilian Y Pfeiffer, Narunas Vaskevicius, Robert Schirmer, and Kai O Arras. "Accurate detection and 3D localization of humans using a novel YOLO-based RGB-D fusion approach and synthetic training data." In: () (cit. on p. 41).

[72] Timm Linder, Dennis Griesser, Narunas Vaskevicius, and Kai O Arras. "Towards accurate 3D person detection and localization from RGB-D in cluttered environments." In: *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Workshop on Robotics for Logistics in Warehouses and Environments Shared with Humans*. 2018 (cit. on p. 41).

[73] Luiz Alexandre Viana Magno et al. "Optogenetic Stimulation of the M2 Cortex Reverts Motor Dysfunction in a Mouse Model of Parkinson's Disease." In: *The Journal of Neuroscience* 39.17 (2019), pp. 3234–3248. DOI: 10.1523/JNEUROSCI.2277-18.2019 (cit. on p. 53).

[74] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. "A Simple Yet Effective Baseline for 3d Human Pose Estimation." In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2659–2668. DOI: 10.1109/ICCV.2017.288. URL: https://doi.org/10.1109/ICCV.2017.288 (cit. on p. 33).

[75] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning." In: *Nature Neuroscience* 21.9 (2018), pp. 1281–1289. DOI: 10.1038/s41593-018-0209-y (cit. on pp. xv, 44, 45, 48, 49).

[76] Mackenzie Weygandt Mathis and Alexander Mathis. "Deep learning tools for the measurement of animal behavior in neuroscience." In: *Current Opinion in Neurobiology* 60 (2019), pp. 1–11. DOI: 10.1016/j.conb.2019.10.008 (cit. on pp. 45, 48).

[77] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. "What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation?" In: *Int. J. Comput. Vis.* 126.9 (2018), pp. 942–960. DOI: 10.1007/s11263-018-1082-6. URL: https://doi.org/10.1007/s11263-018-1082-6 (cit. on p. 6).

[78] Trevor M McLain. "The use of factor analysis in the development of hand sizes for glove design." In: (2010).

[79] Oier Mees, Andreas Eitel, and Wolfram Burgard. "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments." In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*. IEEE, 2016, pp. 151–156. DOI: 10.1109/IROS.2016.7759048. URL: https://doi.org/10.1109/IROS.2016.7759048 (cit. on p. 37).

[80] Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. "Monocular 3D Human Pose Estimation Using Transfer Learning and Improved CNN Supervision." In: *CoRR* abs/1611.09813 (2016). arXiv: 1611.09813. URL: http://arxiv.org/abs/1611.09813 (cit. on p. 17).

[81] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. "VNect: real-time 3D human pose estimation with a single RGB camera." In: *ACM Trans. Graph.* 36.4 (2017), 44:1–44:14. DOI: 10.1145/3072959.3073596. URL: https://doi.org/10.1145/3072959.3073596 (cit. on p. 33).

[82] Bartul Mimica, Benjamin A. Dunn, Tuce Tombaz, V. P. T. N. C. Srikanth Bojja, and Jonathan R. Whitlock. "Efficient cortical coding of 3D posture in freely behaving rats." In: *Science* 362.6414 (2018), pp. 584–589. DOI: 10.1126/science.aau2013 (cit. on p. 44).

[83] Mixamo. URL: http://www.mixamo.com (cit. on p. 22).

[84] Gyeongsik Moon, Ju Yong Chang, Yumin Suh, and Kyoung Mu Lee. "Holistic Planimetric prediction to Local Volumetric prediction for 3D Human Pose Estimation." In: *CoRR* abs/1706.04758 (2017). arXiv: 1706.04758. URL: http://arxiv.org/abs/1706.04758 (cit. on p. 33).

[85] Francesc Moreno-Noguer. "3D Human Pose Estimation from a Single Image via Distance Matrix Regression." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1561–1570. DOI: 10.1109/CVPR.2017.170. URL: https://doi.org/10.1109/CVPR.2017.170 (cit. on p. 17).

[86] Leap Motion. *https://www.leapmotion.com*. URL: https://www.leapmotion.com (cit. on p. 61).

[87] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. "Real-Time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor." In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 1163–1172. DOI: 10.1109/ICCV.2017.131. URL: https://doi.org/10.1109/ICCV.2017.131 (cit. on pp. 58, 60, 61).

[88] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. "GANerated Hands for Real-Time 3D Hand Tracking From Monocular RGB." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 49–59. DOI: 10.1109/CVPR.2018.00013. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Mueller_GANerated_Hands_for_CVPR_2018_paper.html (cit. on pp. 12, 29, 58, 61, 62).

[89] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. "Real-time pose and shape reconstruction of two interacting hands with a single depth camera." In: *ACM Trans. Graph.* 38.4 (2019), 49:1–49:13. DOI: 10.1145/3306346.3322958. URL: https://doi.org/10.1145/3306346.3322958 (cit. on p. 78).

[90] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. "Hands Deep in Deep Learning for Hand Pose Estimation." In: *CoRR* abs/1502.06807 (2015). arXiv: 1502.06807. URL: http://arxiv.org/abs/1502.06807 (cit. on pp. 11, 17, 18, 26).

[91] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. "Training a Feedback Loop for Hand Pose Estimation." In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 3316–3324. DOI: 10.1109/ICCV.2015.379. URL: https://doi.org/10.1109/ICCV.2015.379 (cit. on pp. 17, 18).

[92] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. "Efficient model-based 3D tracking of hand articulations using Kinect." In: *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*. Ed. by Jesse Hoey, Stephen J. McKenna, and Emanuele Trucco. BMVA Press, 2011, pp. 1–11. DOI: 10.5244/C.25.101. URL: https://doi.org/10.5244/C.25.101 (cit. on pp. 17, 18).

[93] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. "Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints." In: *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*. Ed. by Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool. IEEE Computer Society, 2011, pp. 2088–2095. DOI: 10.1109/ICCV.2011.6126483. URL: https://doi.org/10.1109/ICCV.2011.6126483 (cit. on pp. 11, 12).

[94] Gabriel L. Oliveira, Abhinav Valada, Claas Bollen, Wolfram Burgard, and Thomas Brox. "Deep learning for human part discovery in images." In: *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*. Ed. by Danica Kragic, Antonio Bicchi, and Alessandro De Luca. IEEE, 2016, pp. 1634–1641. DOI: 10.1109/ICRA.2016.7487304. URL: https://doi.org/10.1109/ICRA.2016.7487304 (cit. on p. 12).

[95] Paschalis Panteleris, Iason Oikonomidis, and Antonis A. Argyros. "Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the Wild." In: *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. IEEE Computer Society, 2018, pp. 436–445. DOI: 10.1109/WACV.2018.00054. URL: https://doi.org/10.1109/WACV.2018.00054 (cit. on pp. 12, 29).

[96] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1263–1272. DOI: 10.1109/CVPR.2017.139. URL: https://doi.org/10.1109/CVPR.2017.139 (cit. on p. 17).

[97] Talmo D Pereira, Diego E Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S-H Wang, Mala Murthy, and Joshua W Shaevitz. "Fast animal pose estimation using deep neural networks." In: *Nature Methods* (2019), p. 117 (cit. on p. 45).

[98] *RWTH German Fingerspelling Database*. http://www-i6.informatik.rwth-aachen.de/~dreuw/fingerspelling.php. Accessed: 01.03.2017 (cit. on p. 29).

[99] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. "Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation." In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11214. Lecture Notes in Computer Science. Springer, 2018, pp. 765–782. DOI: 10.1007/978-3-030-01249-6_46. URL: https://doi.org/10.1007/978-3-030-01249-6_46 (cit. on p. 77).

[100] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. "Learning Monocular 3D Human Pose Estimation From Multi-View Images." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 8437–8446. DOI: 10.1109/CVPR.2018.00880. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Rhodin_Learning_Monocular_3D_CVPR_2018_paper.html (cit. on p. 77).

[101] Javier Romero, Hedvig Kjellström, and Danica Kragic. "Hands in action: real-time 3D reconstruction of hands in interaction with objects." In: *IEEE International Conference on Robotics and Automation, ICRA 2010, Anchorage, Alaska, USA, 3-7 May 2010*. IEEE, 2010, pp. 458–463. DOI: 10.1109/ROBOT.2010.5509753. URL: https://doi.org/10.1109/ROBOT.2010.5509753 (cit. on p. 12).

[102] Javier Romero, Dimitrios Tzionas, and Michael J. Black. "Embodied hands: modeling and capturing hands and bodies together." In: *ACM Trans. Graph.* 36.6 (2017), 245:1–245:17. DOI: 10.1145/3130800.3130883. URL: https://doi.org/10.1145/3130800.3130883 (cit. on pp. 59, 66).

[103] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi. Vol. 9351. Lecture Notes in Computer Science. Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28. URL: https://doi.org/10.1007/978-3-319-24574-4_28 (cit. on p. 35).

[104] Benjamin Sapp, Rizwan Chaudhry, Xiaodong Yu, Gautam Singh, Ian Perera, Francis Ferraro, Evelyne Tzoukermann, Jana Kosecka, and Jan Neumann. "Recognizing manipulation actions in arts and crafts shows using domain-specific visual and textual cues." In: *IEEE International Conference on Computer Vision Workshops,*

ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011. IEEE Computer Society, 2011, pp. 1554–1561. DOI: 10.1109/ICCVW.2011.6130435. URL: https://doi.org/10.1109/ICCVW.2011.6130435.

[105] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. "3D Human pose estimation: A review of the literature and analysis of covariates." In: *Comput. Vis. Image Underst.* 152 (2016), pp. 1–20. DOI: 10.1016/j.cviu.2016.09.002. URL: https://doi.org/10.1016/j.cviu.2016.09.002 (cit. on p. 17).

[106] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 1010–1019. DOI: 10.1109/CVPR.2016.115. URL: https://doi.org/10.1109/CVPR.2016.115 (cit. on p. 35).

[107] Toby Sharp et al. "Accurate, Robust, and Flexible Real-time Hand Tracking." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*. Ed. by Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo. ACM, 2015, pp. 3633–3642. DOI: 10.1145/2702123.2702179. URL: https://doi.org/10.1145/2702123.2702179 (cit. on pp. 12, 17, 18).

[108] Swathi Sheshadri, Benjamin Dann, Timo Hueser, and Hansjörg Scherberger. "3D reconstruction toolbox for behavior tracked with multiple cameras." In: *J. Open Source Softw.* 5.45 (2020), p. 1849. DOI: 10.21105/joss.01849. URL: https://doi.org/10.21105/joss.01849 (cit. on p. 55).

[109] Nobutaka Shimada, Yoshiaki Shirai, Yoshinori Kuno, and Jun Miura. "Hand Gesture Estimation and Model Refinement Using Monocular Camera - Ambiguity Limitation by Inequality Constraints." In: *3rd International Conference on Face & Gesture Recognition (FG '98), April 14-16, 1998, Nara, Japan*. IEEE Computer Society, 1998, pp. 268–273 (cit. on p. 12).

[110] Jamie Shotton et al. "Efficient Human Pose Estimation from Single Depth Images." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.12 (2013), pp. 2821–2840. DOI: 10.1109/TPAMI.2012.241. URL: https://doi.org/10.1109/TPAMI.2012.241.

[111] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew W. Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. "Real-time human pose recognition in parts from single depth images." In: *Commun. ACM* 56.1 (2013), pp. 116–124. DOI: 10.1145/2398356.2398381. URL: http://doi.acm.org/10.1145/2398356.2398381 (cit. on p. 33).

[112] Tomas Simon, Hanbyul Joo, Iain A. Matthews, and Yaser Sheikh. "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 4645–4653. DOI: 10.1109/CVPR.2017.494. URL: https://doi.org/10.1109/CVPR.2017.494 (cit. on pp. 11, 12, 34, 36, 58, 59, 60, 62, 65).

[113] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. "Cross-Modal Deep Variational Hand Pose Estimation." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 89–98. DOI: 10.1109/CVPR.2018.00017. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Spurr_Cross-Modal_Deep_Variational_CVPR_2018_paper.html (cit. on p. 29).

[114] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. "Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints." In: *CoRR* abs/2003.09282 (2020). arXiv: 2003.09282. URL: https://arxiv.org/abs/2003.09282 (cit. on p. 72).

[115] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. "Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data." In: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 2456–2463. DOI: 10.1109/ICCV.2013.305. URL: https://doi.org/10.1109/ICCV.2013.305 (cit. on p. 12).

[116] Srinath Sridhar, Helge Rhodin, Hans-Peter Seidel, Antti Oulasvirta, and Christian Theobalt. "Real-Time Hand Tracking Using a Sum of Anisotropic Gaussians Model." In: *2nd International Conference on 3D Vision, 3DV 2014, Tokyo, Japan, December 8-11, 2014, Volume 1*. IEEE Computer Society, 2014, pp. 319–326. DOI: 10.1109/3DV.2014.37. URL: https://doi.org/10.1109/3DV.2014.37 (cit. on p. 12).

[117] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. "Real-Time Joint Tracking of a Hand Manipulating an Object from RGB-D Input." In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9906. Lecture Notes in Computer Science. Springer, 2016, pp. 294–310. DOI: 10.1007/978-3-319-46475-6_19. URL: https://doi.org/10.1007/978-3-319-46475-6_19 (cit. on pp. 21, 60).

[118] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew W. Fitzgibbon. "The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation." In: *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. IEEE Computer Society, 2012, pp. 103–110. DOI: 10.1109/CVPR.2012.6247664. URL: https://doi.org/10.1109/CVPR.2012.6247664.

[119] Jonathan Taylor et al. "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences." In: *ACM Trans. Graph.* 35.4 (2016), 143:1–143:12. DOI: 10.1145/2897824.2925965. URL: https://doi.org/10.1145/2897824.2925965 (cit. on p. 12).

[120] Bugra Tekin, Federica Bogo, and Marc Pollefeys. "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4511–4520. DOI: 10.1109/CVPR.2019.00464. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Tekin_HO_Unified_Egocentric_Recognition_of_3D_Hand-Object_Poses_and_Interactions_CVPR_2019_paper.html (cit. on p. 29).

[121] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. "Sphere-meshes for real-time hand modeling and tracking." In: *ACM Trans. Graph.* 35.6 (2016), 222:1–222:11. URL: http://dl.acm.org/citation.cfm?id=2980226 (cit. on p. 59).

[122] Anastasia Tkach, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew W. Fitzgibbon. "Online generative model personalization for hand tracking." In: *ACM Trans. Graph.* 36.6 (2017), 243:1–243:11. DOI: 10.1145/3130800.3130830. URL: https://doi.org/10.1145/3130800.3130830 (cit. on p. 59).

[123] Denis Tomè, Chris Russell, and Lourdes Agapito. "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5689–5698. DOI: 10.1109/CVPR.2017.603. URL: https://doi.org/10.1109/CVPR.2017.603 (cit. on pp. 11, 17, 33, 38, 39, 40).

[124] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation." In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger. 2014, pp. 1799–1807. URL: http://papers.

nips.cc/paper/5573-joint-training-of-a-convolutional-network-and-a-graphical-model-for-human-pose-estimation (cit. on p. 17).

[125] Jonathan Tompson, Murphy Stein, Yann LeCun, and Ken Perlin. "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks." In: *ACM Trans. Graph.* 33.5 (2014), 169:1–169:10. DOI: 10.1145/2629500. URL: https://doi.org/10.1145/2629500 (cit. on pp. 17, 18, 22).

[126] Alexander Toshev and Christian Szegedy. "DeepPose: Human Pose Estimation via Deep Neural Networks." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1653–1660. DOI: 10.1109/CVPR.2014.214. URL: https://doi.org/10.1109/CVPR.2014.214 (cit. on pp. 12, 17).

[127] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. "Deep Image Harmonization." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2799–2807. DOI: 10.1109/CVPR.2017.299. URL: https://doi.org/10.1109/CVPR.2017.299 (cit. on p. 69).

[128] Dimitrios Tzionas, Abhilash Srikantha, Pablo Aponte, and Juergen Gall. "Capturing Hand Motion with an RGB-D Sensor, Fusing a Generative Model with Salient Points." In: *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings*. Ed. by Xiaoyi Jiang, Joachim Hornegger, and Reinhard Koch. Vol. 8753. Lecture Notes in Computer Science. Springer, 2014, pp. 277–289. DOI: 10.1007/978-3-319-11752-2_22. URL: https://doi.org/10.1007/978-3-319-11752-2_22 (cit. on p. 59).

[129] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. "Capturing Hands in Action Using Discriminative Salient Points and Physics Simulation." In: *Int. J. Comput. Vis.* 118.2 (2016), pp. 172–193. DOI: 10.1007/s11263-016-0895-4. URL: https://doi.org/10.1007/s11263-016-0895-4 (cit. on pp. 11, 58, 78).

[130] Benjamin Ummenhofer. "Introduction to dense reconstruction from multiple images." PhD thesis. University of Freiburg, Freiburg im Breisgau, Germany, 2018. URL: https://freidok.uni-freiburg.de/data/17313.

[131] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. "Self-Supervised 3D Hand Pose Estimation Through Training by Fitting." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*.

Computer Vision Foundation / IEEE, 2019, pp. 10853–10862. DOI: 10.1109/CVPR.2019.01111. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Wan_Self-Supervised_3D_Hand_Pose_Estimation_Through_Training_by_Fitting_CVPR_2019_paper.html (cit. on p. 77).

[132] Robert Y. Wang and Jovan Popovic. "Real-time hand-tracking with a color glove." In: *ACM Trans. Graph.* 28.3 (2009), p. 63. DOI: 10.1145/1531326.1531369. URL: https://doi.org/10.1145/1531326.1531369 (cit. on p. 60).

[133] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. "Convolutional Pose Machines." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4724–4732. DOI: 10.1109/CVPR.2016.511. URL: https://doi.org/10.1109/CVPR.2016.511 (cit. on pp. 11, 12, 17, 19, 33, 34, 36).

[134] Tim Welschehold, Christian Dornhege, and Wolfram Burgard. "Learning manipulation actions from human demonstrations." In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*. IEEE, 2016, pp. 3772–3777. DOI: 10.1109/IROS.2016.7759555. URL: https://doi.org/10.1109/IROS.2016.7759555.

[135] Tim Welschehold, Christian Dornhege, and Wolfram Burgard. "Learning mobile manipulation actions from human demonstrations." In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*. IEEE, 2017, pp. 3196–3201. DOI: 10.1109/IROS.2017.8206152. URL: https://doi.org/10.1109/IROS.2017.8206152 (cit. on p. 39).

[136] Tim Welschehold, Christian Dornhege, Fabian Paus, Tamim Asfour, and Wolfram Burgard. "Coupling Mobile Base and End-Effector Motion in Task Space." In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1-5, 2018*. IEEE, 2018, pp. 1–9. DOI: 10.1109/IROS.2018.8593534. URL: https://doi.org/10.1109/IROS.2018.8593534 (cit. on p. 41).

[137] Tim Welschehold, Nichola Abdo, Christian Dornhege, and Wolfram Burgard. "Combined Task and Action Learning from Human Demonstrations for Mobile Manipulation Applications." In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*. IEEE, 2019, pp. 4317–4324. DOI: 10.1109/IROS40897.2019.8968091. URL: https://doi.org/10.1109/IROS40897.2019.8968091 (cit. on p. 41).

[138] Tim Wengefeld, Benjamin Lewandowski, Daniel Seichter, Lennard Pfennig, and Horst-Michael Gross. "Real-time Person Orientation Estimation using Colored Pointclouds." In: *2019 European Conference on Mobile Robots, ECMR 2019, Prague, Czech Republic, September 4-6, 2019*. IEEE, 2019, pp. 1–7. DOI: 10.1109/ECMR.2019.8870914. URL: https://doi.org/10.1109/ECMR.2019.8870914 (cit. on p. 41).

[139] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. "SeqHAND: RGB-Sequence-Based 3D Hand Pose and Shape Estimation." In: *CoRR* abs/2007.05168 (2020). arXiv: 2007.05168. URL: https://arxiv.org/abs/2007.05168 (cit. on p. 72).

[140] Yi Yang and Deva Ramanan. "Articulated pose estimation with flexible mixtures-of-parts." In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, 2011, pp. 1385–1392. DOI: 10.1109/CVPR.2011.5995741. URL: https://doi.org/10.1109/CVPR.2011.5995741.

[141] Qi Ye and Tae-Kyun Kim. "Occlusion-Aware Hand Pose Estimation Using Hierarchical Mixture Density Network." In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11214. Lecture Notes in Computer Science. Springer, 2018, pp. 817–834. DOI: 10.1007/978-3-030-01249-6_49. URL: https://doi.org/10.1007/978-3-030-01249-6_49 (cit. on p. 77).

[142] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. "HUMBI: A Large Multiview Dataset of Human Body Expressions." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2990–3000 (cit. on pp. 55, 72).

[143] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=Sy8gdB9xx (cit. on p. 6).

[144] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. "A hand pose tracking benchmark from stereo matching." In: *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*. IEEE, 2017, pp. 982–986. DOI: 10.1109/ICIP.2017.8296428. URL: https://doi.org/10.1109/ICIP.2017.8296428 (cit. on pp. 21, 27, 28, 60, 61, 62).

[145] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. "Real-time user-guided image colorization with learned deep priors." In: *ACM Trans. Graph.* 36.4 (2017), 119:1–119:11. DOI: 10.1145/3072959.3073703. URL: https://doi.org/10.1145/3072959.3073703 (cit. on p. 69).

[146] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. "Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets without Superior Knowledge." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6528–6537.

[147] Ruiqi Zhao, Yan Wang, and Aleix M. Martínez. "A Simple, Fast and Highly-Accurate Algorithm to Recover 3D Shape from 2D Landmarks on a Single Image." In: *CoRR* abs/1609.09058 (2016). arXiv: 1609.09058. URL: http://arxiv.org/abs/1609.09058 (cit. on p. 17).

[148] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. "Hand-3d-Studio: A New Multi-View System for 3d Hand Reconstruction." In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 2478–2482 (cit. on pp. 55, 72).

[149] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. "Model-Based Deep Hand Pose Estimation." In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. Ed. by Subbarao Kambhampati. IJCAI/AAAI Press, 2016, pp. 2421–2427. URL: http://www.ijcai.org/Abstract/16/345 (cit. on pp. 17, 18, 26).

[150] Thomas G Zimmerman, Jaron Lanier, Chuck Blanchard, Steve Bryson, and Young Harvill. "A hand gesture interface device." In: *ACM SIGCHI Bulletin*. ACM. 1987, pp. 189–192 (cit. on p. 60).

[151] Christian Zimmermann and Thomas Brox. "Learning to Estimate 3D Hand Pose from Single RGB Images." In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 4913–4921. DOI: 10.1109/ICCV.2017.525. URL: https://doi.org/10.1109/ICCV.2017.525 (cit. on pp. 19, 21, 58, 60, 61, 62, 68).

[152] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. "3D Human Pose Estimation in RGBD Images for Robotic Task Learning." In: *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*. IEEE, 2018, pp. 1986–1992. DOI: 10.1109/ICRA.2018.8462833. URL: https://doi.org/10.1109/ICRA.2018.8462833 (cit. on pp. 31, 39, 40).

[153] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max J. Argus, and Thomas Brox. "FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images." In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 813–822. DOI: 10.1109/ICCV.2019.00090. URL: https://doi.org/10.1109/ICCV.2019.00090.

[154] Christian Zimmermann, Artur Schneider, Mansour Alyahyay, Thomas Brox, and Ilka Diester. "FreiPose: A Deep Learning Framework for Precise Animal Motion Capture in 3D Spaces." In: *bioRxiv* (2020) (cit. on p. 46).

[155] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. "Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape From Images." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3955–3963. DOI: 10.1109/CVPR.2018.00416. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Zuffi_Lions_and_Tigers_CVPR_2018_paper.html (cit. on p. 45).

[156] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. "3D Menagerie: Modeling the 3D Shape and Pose of Animals." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5524–5532. DOI: 10.1109/CVPR.2017.586. URL: https://doi.org/10.1109/CVPR.2017.586 (cit. on p. 45).