

3D Hand Shape and Pose from Images in the Wild

Adnane Boukhayma¹, Rodrigo de Bem^{1,2}, Philip H.S. Torr¹

¹ University of Oxford, UK

² Federal University of Rio Grande, Brazil

Fadnane.boukhayma, rodri go.andradedebem, philip.torr@eng.ox.ac.uk

Abstract

We present in this work the first end-to-end deep learning based method that predicts both 3D hand shape and pose from RGB images in the wild. Our network consists of the concatenation of a deep convolutional encoder, and a fixed model-based decoder. Given an input image, and optionally 2D joint detections obtained from an independent CNN, the encoder predicts a set of hand and view parameters. The decoder has two components: A pre-computed articulated mesh deformation hand model that generates a 3D mesh from the hand parameters, and a re-projection module controlled by the view parameters that projects the generated hand into the image domain. We show that using the shape and pose prior knowledge encoded in the hand model within a deep learning framework yields state-of-the-art performance in 3D pose prediction from images on standard benchmarks, and produces geometrically valid and plausible 3D reconstructions. Additionally, we show that training with weak supervision in the form of 2D joint annotations on datasets of images in the wild, in conjunction with full supervision in the form of 3D joint annotations on limited available datasets allows for good generalization to 3D shape and pose predictions on images in the wild.

1. Introduction

Human hand pose estimation and reconstruction in 3D is a long standing problem in the computer vision and graphics communities that has applications in various domains such as virtual and augmented reality and human-machine interaction [35, 15, 46, 13]. With the abundance of affordable commodity depth cameras, the research literature focused naturally more on estimating 3D hand pose through depth observations (e.g. [62, 66, 10, 36, 61]), and many works also explored this problem in multi-view setups [33, 65, 41, 8, 31, 50]. When it comes to a monocular color

input, the problem becomes inherently ill posed due to the increased depth and scale ambiguities, but that did not prevent several researchers [4, 9, 51, 57, 63, 39] from attempting to solve it in the past albeit with limited results. More recently, the unprecedented success of deep learning on similar tasks motivated new work with encouraging results for 3D hand pose from single images [68, 27, 7, 47, 14]. Nevertheless, this task remains particularly difficult: Unlike clothed human bodies or faces, hands have an almost uniform appearance and lack characteristic local features such as eyes and mouths in faces. Unlike bodies, they can have more complex pose configurations and they can be captured from a much wider range of views. Furthermore when observed in the wild, as in dataset MPII+NZSL [44] (Figure 9), their images usually contain external occlusion, self-occlusion, clutter and blur due to their motion. Besides, hands are often small in size compared to the scene so cropped patches around them have low resolutions.

The main obstacles for 3D hand pose estimation from images with deep learning include: (i) The lack of large datasets annotated with reliable 3D ground-truth and (ii) the incapability of the current 3D annotated datasets to make networks generalize greatly to challenging images in the wild.

The first point is tackled by the literature through training with synthetic images [68], populating datasets by transforming synthetic images into real looking ones [27], or leveraging auxiliary types of data in training like depth [7, 47]. We propose a different and simple yet efficient approach to alleviate both challenges (i) and (ii) by circumventing heavy dependence of 3D data in training: Instead of relying on images paired with 3D joint annotations to learn a prior on hand geometry, we exploit a recently proposed differentiable articulated mesh deformation hand model [40] built with linear blend skinning [18], and we reformulate the prediction problem into a learning-based model fitting, that can be trained using both 3D and 2D joint annotations. Training with 2D annotated images al-

Figure 1: Our pipeline takes as input a hand image and optionally 2D joint heat-maps from an independent CNN. The encoder generates the shape, pose and view parameters. The hand parameters are fed to the hand model that generates a triangulated 3D mesh and its underlying 3D skeleton. The latter are re-projected into the image domain using a weak perspective camera model controlled by the view parameters. This network is trained end-to-end with a combination of weak 2D and full 3D joint supervision. The hand and view parameters are not supervised.

lowers access to larger datasets (e.g. PANOPTIC [44]) with a fair share of annotated images in the wild (e.g. MPII+NZSL [44]) compared to datasets with 3D ground-truth, thus helping improve generalization to this type of challenging data. Given an input image, and optionally 2D joint detections obtained from an independent CNN, a deep convolutional encoder predicts the hand shape and pose parameters and view parameters. The model-based decoder uses the latter to generate a 3D triangulated hand mesh and its underlying skeleton, along with their re-projection in image domain (see Figure 1).

Our contributions in this paper are as follows: This work is the first to propose end-to-end learning of both 3D hand shape and pose from a single RGB image. We also show for the first time that the prior knowledge of factored hand shape and pose in a pre-computed linear blend skinning [18] hand model [40] combined with a deep-convolutional encoder yields state-of-the-art performance in 3D pose prediction from images, and produces geometrically valid and plausible 3D reconstructions, without the need for post-processing optimizations [27]. We show that this strategy combined with training on 2D annotated datasets of images in the wild produces good generalization in 3D hand reconstruction for challenging images in uncontrolled environments.

We evaluate our work both quantitatively in terms of 3D pose estimation and qualitatively using various public datasets. These evaluation sets account for cases of hand interaction with objects, occlusion and clutter, and contain egocentric view images, third person view images, and images in the wild. Our method obtains state-of-the-art results on standard benchmarks, even compared to methods using additional depth information in training [7, 47], camera intrinsics [27, 34], and post-processing optimization [27]. Our method shows superior qualitative results on a challenging dataset of images in the wild (Figure 9 & supplementary material).

2. Related work

There is a rich literature on 3D hand pose and reconstruction from depth [62, 66, 10, 36, 61, 11, 43, 45, 19, 20, 24, 30, 37, 48, 52, 53, 59, 64], image and depth [26, 32, 49, 28], stereo [33, 65, 41] and multiple images [8, 31, 50]. We focus hereby on research material that solely considers a single color input image.

3D hand pose from a single image

Pre-deep learning There have been attempts to solve 3D hand pose estimation from a monocular color input prior to deep learning with both discriminative and generative approaches [4, 9, 51, 57, 63, 39]. However, most of these methods have limited performance and depend on various requirements such as careful initialization and prior knowledge of the background.

Deep learning The work of [68] was the first to propose 3D hand pose estimation from single images using deep learning. Their method consists of the concatenation of three networks that segment the hand, predict 2D joints, and then predict 3D joints subsequently. The work of [27] shows that the previous method generalizes poorly to real world images since a major part of their training data is synthetic. In turn, they ([27]) propose to use Cycle-GAN [67] to transform synthetic 3D annotated images of hands into real looking ones. The resulting data is used to train a regressor to predict 2D and 3D hand joints. A final optimization step fits a 3D skeleton to the former 2D and 3D predictions using the camera intrinsics. The method in [34] consists in an optimization that fits a hand model to 2D joint detections obtained from a state-of-the-art CNN [44]. We also use a pre-defined hand model [40] but within a pipeline trained end-to-end.

Depth regularization Recent works tackle depth ambiguity in 3D hand pose prediction from images by leveraging depth maps in training. [7] proposes to reduce the

dependency on noisy 3D annotations in real datasets by introducing a network that predicts full depth maps from the 3D joints. This depth regularizer is trained with ground-truth depth data for both real and synthetic training images, while the 3D predictions are only supervised by the reliable synthetic labels. The authors in [47] use multiple variational auto-encoders sharing the same latent space each auto-encoding a separate hand data modality (e.g. images, 2D joints, 3D joints). They show that the auxiliary auto-encoders help regularize the latent space and produce improved cross-modal predictions (e.g. image to 3D joints). [14] shows that predicting an implicit 2.5D heat-map representation yields improved 3D predictions even without explicit full depth-map supervision.

Hand models Many hand models have been proposed in the literature primarily aiming at tracking depth and color data, where the hand is modelled using various techniques such as assembled geometric primitives [32], sum of Gaussians [50], sphere meshes [58] or loop subdivision of a control mesh [20]. In order to better capture the shape of the hand, [32] defines scaling terms to allow bone length to vary, while [54] pre-calibrates the shape to fit the hand of interest. The work in [20] was the first to learn hand shape variation from scans with linear blend skinning [18]. The model proposed recently in [40] and referred to as MANO improves on the latter by learning pose dependent corrective blend shapes [25], thus modelling both hand shape and pose and generating more realistic posed meshes. We use the MANO [40] model in this work.

Model-based decoders Several works propose to combine deep convolutional encoders with generative models as decoders for human face [56, 55] and body [17, 60] 3D reconstruction. In many of these works, the decoder is a combination of a parametric model (e.g. linear face model [6], SMPL [25]) and a re-projection/rendering module. While most works fix these decoders, some propose to tune them after a supervised initialization [2, 22, 55]. This is the first work to propose a combination of a CNN encoder with a fixed generative hand model [40] for the problem of 3D hand reconstruction from images.

3. Overview

As illustrated in Figure 1, our pipeline takes as input an image of a hand and optionally 2D joint heat-maps from an independent hand detector. A deep convolutional encoder processes the input and generates a set of hand shape and pose parameters, and a set of view parameters R , t and s . The hand parameters are fed to a differentiable articulated mesh deformation hand model that generates a triangulated 3D mesh and its underlying 3D skeleton. These outputs are then re-projected into the image domain through a weak perspective camera model controlled by the view parameters.

The re-projection module and the hand model together form a model-based decoder whose parameters are fixed and do not require training. The encoder is pre-trained with synthetic examples that we created as elaborated in Section 6. We note that the training of our pipeline is done end-to-end using 2D and 3D joint annotations without supervision on the hand and view parameters, except for a regularization on the hand parameters to ensure their magnitude is small. We detail and explain the functioning of the various parts of the pipeline in the following.

4. Hand model

We use the MANO hand model [40] which is based on the SMPL model for human bodies [25]. It is an articulated mesh deformation model represented with a differentiable function $M(\beta, \theta)$ taking as input two sets of parameters β and θ that control the shape and pose of the generated hand respectively:

$$M(\beta, \theta) = W(T(\beta, \theta), J(\theta), W), \quad (1)$$

where W is a linear blend skinning [18] function applied to a template hand triangulated mesh T rigged with a kinematic tree of $K = 16$ joints. J represents the joint locations and it is learned as a sparse linear regressor from mesh vertices, and W are the blend weights.

In order to reduce the artifacts of linear blend skinning such as overly smooth outputs and mesh collapse around joints, the hand template T is obtained by deforming a mean mesh \bar{T} with both shape and pose corrective blend shapes, S_n and P_n respectively, as follows:

$$T(\beta, \theta) = \bar{T} + \sum_{n=1}^K S_n + \sum_{n=1}^{9K} (R_n(\theta) - R_n(\bar{\theta})) P_n, \quad (2)$$

where $R_n(\theta)$ is the n^{th} element of a vector concatenating rotation matrix coefficients from all joints for pose θ and $\bar{\theta}$ is the rest pose. The model constants $\{\bar{T}, S, P, J, W\}$ are learned using registered hand scans from 31 subjects performing roughly 51 hand poses.

In the SMPL model, the pose vector θ stacks the angle-axis values of the joints. To help the hand model generate physically plausible poses, the authors in [40] reduce this pose representation to a linear embedding by performing Principal Component Analysis on angle-axis values of the joints in the data collected to build the model. The pose vector θ contains the resulting main coefficients from PCA instead of the angle-axis values. 10 coefficients are retained for the pose (\mathbb{R}^{10}), and 10 coefficients are used to represent the shape as well (\mathbb{R}^{10}).

Given input shape and pose parameters, we obtain a hand mesh $M(\beta, \theta)$ of $N = 778$ vertices and 1538 faces, along with the corresponding 3D joints $J(\beta, \theta) = R(J(\beta))$

where R is the global rigid transformation induced by pose θ . As the hand skeleton in MANO does not contain finger tip joints, we append J with 5 vertices from the hand mesh that correspond to these key-points. The final 3D joint output $J(\theta, \phi)$ counts 21 key-points.

5. Camera model

In order to re-project the 3D hand mesh vertices $M(\theta, \phi)$ and 3D joints $J(\theta, \phi)$ into the 2D image plane, we use the weak perspective model. This approximation allows us to train with annotated images even in the absence of camera intrinsics, which is the case of images in the wild obtained from Youtube videos for instance (e.g. dataset MPII+NZSL). Given a global rotation matrix $R \in \text{SO}(3)$, a translation $t \in \mathbb{R}^2$ and a scaling $s \in \mathbb{R}^+$, the projection writes:

$$\hat{x} = s \cdot (RJ(\theta, \phi)) + t, \quad (3)$$

$$\hat{y} = s \cdot (RM(\theta, \phi)) + t, \quad (4)$$

where $\hat{\cdot}$ is the orthographic projection.

6. Encoder

Given an input hand image, the goal of the encoder is to predict the corresponding hand pose and shape parameters $\{\theta, \phi\}$ and camera parameters $\{R, t, s\}$. We use the ResNet-50 network [12] and we adjust the final fully connected layer to output a vector $v = \{R, t, s, \theta, \phi\} \in \mathbb{R}^{26}$. We note that global rotation R is encoded with axis-angle values and is hence represented with 3 parameters. We also experiment with feeding 2D hand joint heat-maps obtained with a state of the art method [44] as additional channel input to the hand RGB image.

Figure 2: Examples from our synthetic dataset created to pre-train the encoder.

Network pre-training We pre-train the encoder to ensure that the camera and hand parameters converge towards acceptable values. For this purpose, we create a synthetic dataset of paired hand images with their ground-truth camera and hand parameters using the same generative model that we use as a decoder. Hand geometries are obtained by sampling poses $[-2, 2]^{10}$ and shapes

$[-0.03, 0.03]^{10}$ then applying rotations R , translations t and scalings s . Although the work of [40] does not model hand appearance, the authors provide the scans used to build the geometry model with their registered counterparts. The original scans come with 3D coordinates and RGB values for each vertex. We create example hand appearances using the registered scan topology: To each vertex in a registered mesh, we assign the RGB value of the closest vertex in the original corresponding scan, and we interpolate these values inside faces. The textured hands are finally rendered on top of random background images. Figure 2 shows examples from the resulting dataset.

7. Training objective

We combine multiple losses to train our pipeline: A 2D joint re-projection loss L_{2D} , a 3D joint loss L_{3D} , a hand mask loss L_{mask} and a model parameter regularization loss L_{reg} .

$$L = L_{2D} + \lambda_{3D} L_{3D} + \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{reg}} L_{\text{reg}}, \quad (5)$$

where $\lambda_{3D} = 10^2$, $\lambda_{\text{mask}} = 10^2$ and $\lambda_{\text{reg}} = 10^1$ are weighting factors.

2D joint re-projection loss This loss ensures that the re-projected hand joints in the image plane coincide with the ground-truth 2D hand joint annotations:

$$L_{2D} = \|\hat{x} - x\|_1, \quad (6)$$

where x is a vector containing the ground-truth 2D hand joint coordinates. We use the L_1 loss to account for inaccuracies in hand annotations in our training datasets.

3D joint loss When ground-truth 3D hand joint annotations are available (e.g STEREO dataset), this loss minimises the distance between the latter and the 3D hand joints generated by the hand model:

$$L_{3D} = \|RJ(\theta, \phi) - x_{3D}\|_2^2, \quad (7)$$

where x_{3D} is a vector containing the ground-truth 3D hand joint coordinates.

Hand mask loss We introduce this novel loss to help speed up the convergence of our training and refine hand shape predictions. This loss penalizes re-projected hand vertices that lie outside of the hand region in a binary mask, which is pre-computed prior to training:

$$L_{\text{mask}} = 1 - \frac{1}{N} \sum_i H(\hat{y}_i), \quad (8)$$

where H is an occlusion-aware hand mask, i.e $H(u) = 1$ if pixel u is inside the hand region even if the hand is occluded in the image, and $H(u) = 0$ otherwise. Notice that

(a) (b) (c)

Figure 3: GrabCut [42] hand segmentation initialized with 2D joint annotation. (a) Input image, (b) foreground, background and undecided regions from 2D joints, (c) final segmentation.

these masks cannot be obtained with hand skin segmentation methods (e.g. [23, 5]) as they are sensitive to occlusions.

We obtain an approximation of these masks (Figure 3) for training images using the GrabCut [42] algorithm, by initializing the foreground, background and probable foreground/background regions using the 2D hand joint annotations: As illustrated in Figure 3b, we create an initial foreground by drawing lines of 1 pixel width connecting joints according to the hand skeleton hierarchy. Pixels inside triangles formed by joints that belong anatomically to the hand surface are appended to the foreground as well. The undecided area is defined as the region within 70 pixels at most from the foreground, and the remaining pixels are assigned to the initial background.

Regularization loss This loss acts on the hand model parameters at the encoder output by reducing their magnitude for physically plausible hand reconstructions and reduced mesh distortions:

$$L_{\text{reg}} = \frac{2}{2} + \frac{2}{2}, \quad (9)$$

where $\lambda = 10^4$ is a weighting factor.

8. Evaluation

We evaluate our method’s 3D pose estimates quantitatively and its 3D reconstructions qualitatively on several datasets and with respect to state-of-the-art methods. Without access to camera intrinsics, and trained merely with 2D and 3D joint annotations, our method outperforms deep learning based competing methods, including those using additional depth information in training or camera intrinsics in evaluation. We show particularly superior 3D reconstructions on images in the wild that present challenging situations such as blur, low resolution, occlusion, extremely varying viewpoints and hand pose configurations.

Similar to [44], input images are assumed to be crops of fixed size around the hand. To achieve this, we use a hand key-point detector [44] to find the tightest rectangular box of edge size l containing the hand. Images are then cropped with a square patch of size $2.2l$ centred at the same 2D location as the previously detected box. The resulting cropped images are subsequently resized to have a width and height of 320. As done in [44], we use the right hand model and images of left hands are flipped horizontally.

Finally, we train our pipeline (Figure 1) using the Adam solver [21] with a learning rate of 10^{-4} and weight decay of 10^{-5} .

Datasets Our training set is made of dataset PANOPTIC [44] that counts 14847 images, the training set of MPII+NZSL [44] that counts 1912 images following the split in [44], and the training set of STEREO [65] that counts 15000 images following the split in [68]. This amounts to 31729 training images, 15000 (STEREO) with 3D joint annotations, and the remaining 16729 (PANOPTIC & MPII+NZSL) with 2D joint annotations only.

The PANOPTIC dataset [44] contains hands in various poses observed from multiple views in the Panoptic studio [16]. The MPII+NZSL dataset [44] is a combination of manually annotated images from The MPII Human Pose dataset [3] containing images from YouTube videos, and images from the New Zealand Sign Language (NZSL) Exercises of the Victoria University of Wellington [38]. The STEREO dataset [65] shows an actor’s hand in third person view counting with the fingers and moving the hand randomly.

For evaluation, we use the DEXTER+OBJECT dataset [49] which shows interactions of an actor’s hand with a cuboid object from a third person view. To evaluate robustness to occlusions and clutter, we use the EGODEXTER dataset [28] that displays a hand from an egocentric view interacting with various objects. We finally use the testing set of MPII+NZSL [44] to assess performance in the presence of blur, low resolution, varying viewpoints and hand pose configurations, among other characteristics of datasets of images in the wild.

Metrics To quantitatively evaluate 3D hand pose estimations, we report the percentage of correct points in 3D (3D PCK) and the average 3D Euclidean distance between the estimated 3D joints and the ground-truth when the latter is available, where distances are expressed in millimeters (mm). When only ground-truth 2D joint annotations are available (dataset MPII+NZSL), we report 2D PCK and the average 2D Euclidean distance between the estimated 2D re-projected joints and the ground-truth, where distances are expressed in pixels (px).

Comparison to competing methods We compare our results on the STEREO dataset to state-of-the-art methods in terms of 3D PCK in Figures 4 and 5, and we show 3D joint

Figure 4: 3D PCK for STEREO.

Figure 6: 3D PCK for DEXTER+OBJECT.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit	Spurr et al.	Zimm. et al.
3D distance	33.16	25.53	25.93	41.18	40.20	34.75

Table 2: Average 3D joint distance (mm) to ground-truth for DEXTER+OBJECT.

Figure 5: 3D PCK for STEREO.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit
3D distance	9.76	10.18	10.46	23.21

Table 1: Average 3D joint distance (mm) to ground-truth for STEREO.

errors in Table 1. Figure 4 shows deep learning based methods (Cai et al. [7], Iqbal et al. [14], Spurr et al. [47], Mueller et al. [27], Zimm. et al [68]) and Figure 5 shows methods that do not rely on deep learning (Panteleris et al. [34], PSO, ICPPSO, CHPR [65]). For this experiment, we add a key-point at the center of the hand palm in the MANO model [40] as an interpolation of several mesh vertices to match the annotation of the STEREO dataset. We reproduce the evaluation protocol initially introduced in [68] by training on 10 sequences and testing on the remaining 2 and aligning predictions to the ground-truth hand root joint. Additionally, for a fair comparison to works [7, 47, 14], we crop the hand images for this experiment such that the final image size is 150% the size of the hand. Using RGB image input only, we obtain state-of-the results even though some of the competing methods use depth data in training ([7, 14]) in addition to images, while others ([27]) post-process their output with an optimization that fits their hand skeleton to their 3D and 2D joint predictions, and which uses the camera intrinsics as an additional input.

Figure 6 shows the performance of our method under occlusions and clutter with 3D PCK on the DEXTER+OBJECT dataset, and Table 2 shows 3D joint errors. Additionally,

Figure 7: 3D PCK for EGODEXTER.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit	Spurr et al.	Zimm. et al.
3D distance	51.87	45.58	45.33	56.59	56.92	52.77

Table 3: Average 3D joint distance (mm) to ground-truth for EGODEXTER.

Figure 7 shows our results on a hand in ego-centric view and in interaction with various objects in terms of 3D PCK on the EGODEXTER dataset, and Table 3 shows 3D joint errors. Our method outperforms the competition in these settings as illustrated in the Figures. We note that we show relative 3D pose estimates for all methods except [14] where the authors report absolute values.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit	Zimm. et al.
2D distance	23.04	18.95	20.65	22.36	59.40

Table 4: Average re-projected 2D joint distance (px) to ground-truth for MPII+NZSL

We expect our method to perform particularly well on datasets of images in the wild, as our training set contains this type of data and accounts for hands in low resolution, blurry, occluded and in challenging views and pose configu-

Figure 8: 2D PCK for MPII+NZSL.

rations. In fact, we compare our results to [68] on the testing set of MPII+NZSL dataset in Figure 8 and Table 4 through 2D PCK and 2D joint error respectively. We outperform [68] with a substantial margin as the Figure shows. The superiority of our method on this dataset is visually confirmed in Figure 9.

Comparison to 2D fitting In the case where 2D joint detections are used as input, an alternative way of solving 3D hand pose estimation is to perform a 2D fitting between the re-projected hand model joints and the key-points detected on the image, in a similar fashion to the work proposed by [34]. Our implementation of this strategy consists in minimizing the following objective function with respect to the weak perspective camera parameters $\{R, t, s\}$ and the hand shape and pose parameters $\{\theta, \phi\}$:

$$E(R, t, s, \theta, \phi) = \sum_i p_i (s (R J_i(\theta, \phi)) + t - x_i)^2 + \frac{\lambda}{2} \|s\|^2 + \frac{\lambda}{2} \|\theta\|^2, \quad (10)$$

where p_i is the i^{th} hand joint estimate confidence provided by the detector CNN [44]. Similarly to the loss in Equation 9, regularization in the second line of Equation 10 is important to ensure plausible 3D hand reconstructions. We perform this optimization using Powell’s Dogleg method [29] within the Chumpy [1] framework.

We compare this method (2D fit) to our proposed approach on datasets STEREO, DEXTER+OBJECT and EGODEXTER with 3D PCK in Figures 5, 6 and 7 and 3D joint error in Tables 1, 2 and 3 respectively, and also on dataset MPII+NZSL with 2D PCK in Figure 8 and 2D joint error in Table 4. Results show that our approach outperforms the 2D fitting based strategy for all datasets. We observe that while the optimization catches up slightly with our method in 2D (MPII+NZSL), its performance drops considerably in 3D. Our method benefits clearly from solving the fitting problem in a learning framework and leverages visual cues in predicting the 3D hand position and configuration, while the 2D fitting relies merely on the 2D joint detection information. We also outperform the 2D fitting based method in [34] that uses a similar hand model to [32]

and a perspective projection model on dataset STEREO in Figure 5.

Ablation study We evaluate the difference between using images only (Ours RGB), using 2D joint heat-maps obtained from a state-of-the-art hand detector [44] only (Ours 2D), and finally using both together as input (Ours RGB+2D). We carry comparisons on datasets STEREO, DEXTER+OBJECT and EGODEXTER with 3D PCK in Figures 5, 6 and 7 and 3D joint error in Tables 1, 2 and 3 respectively, and also on dataset MPII+NZSL with 2D PCK in Figure 8 and 2D joint error in Table 4. On dataset STEREO, training on images alone yields the best performance, while training with a combination of images and 2D joint heat-maps is generally the most suitable approach for the other datasets that we tested on.

Qualitative Figure 9 shows our 3D hand reconstructions on the challenging testing set of MPII+NZSL. As shown in this Figure, the input data (9a) displays images of hands that are sometimes blurry, low resolved, occluded, viewed from varying viewpoints and in varying pose configurations. We show our 3D mesh overlaid on the input image (9b) and in alternative views (9c, 9d). We also compare our hand skeleton (9e) to the 2D and 3D pose predictions of [68] (9f, 9g) and the 3D predictions of [47] (9h). Our method obtains visually plausible results while the methods in [68] and [47] fail to predict good 3D pose estimates for many cases in the MPII+NZSL dataset. We show more examples in the supplementary material.

9. Conclusion

We presented a method to predict 3D hand pose and shape from a single RGB image. We combine a deep convolutional encoder with a generative hand model as decoder and train the resulting network end-to-end with 2D and 3D hand joint annotated images. The encoder predicts hand parameters that are inputted to the hand model, and view parameters that are used to re-project the generated 3D hand into the image domain. We generate state-of-the-art results on 3D pose benchmarks and show compelling 3D reconstruction on a challenging set of images in the wild. This method could benefit greatly from a hand appearance model by leveraging a photometric loss in training as proposed in [56, 55] for faces. One possible extension to this work could be to allow some components of the MANO [40] model such as the corrective blend shapes S and P (Equation 2) to be fine-tuned in training for improved performance.

Acknowledgement

This work was supported by the ERC grant ERC-2012-AdG 321162-HELIOS, the EPSRC grant See-bibyte EP/M013774/1 and the EPSRC/MURI grant EP/N019474/1.

(a) Input (b) Our mesh (c) Back view (d) Side view (e) Our skeleton (f) [68]2D (g) [68]3D (h) [47]

Figure 9: Our 3D hand reconstruction on examples from the challenging testing set of MPH+NZSL compared to the 3D hand pose predictions of [68] and [47].

References

- [1] <http://chumpy.org>. 7
- [2] V. F. Abrevaya, S. Wuhler, and E. Boyer. Multilinear autoencoder for 3d face model learning. In *WACV*, 2018. 3
- [3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5
- [4] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *CVPR*, 2003. 1, 2
- [5] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 5
- [6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Conference on Computer graphics and interactive techniques*, 1999. 3
- [7] Y. Cai, L. Ge, J. Cai, and J. Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, 2018. 1, 2, 6
- [8] T. E. de Campos and D. W. Murray. Regression-based hand pose estimation from multiple cameras. In *CVPR*, 2006. 1, 2
- [9] M. de La Gorce, D. J. Fleet, and N. Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence*, 2011. 1, 2
- [10] L. Ge, Y. Cai, J. Weng, and J. Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *CVPR*, 2018. 1, 2
- [11] L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *CVPR*, 2016. 2
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 4
- [13] W. Hürst and C. Van Wezel. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 2013. 1
- [14] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018. 1, 3, 6
- [15] Y. Jang, S.-T. Noh, H. J. Chang, T.-K. Kim, and W. Woo. 3d finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint. *IEEE Transactions on Visualization and Computer Graphics*, 2015. 1
- [16] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 5
- [17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3
- [18] L. Kavan and J. Žára. Spherical blend skinning: A real-time deformation of articulated models. In *Symposium on Interactive 3D Graphics and Games*, 2005. 1, 2, 3
- [19] C. Keskin, F. Kırac, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012. 2
- [20] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon. Learning an efficient model of hand shape variation from depth images. In *CVPR*, 2015. 2, 3
- [21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [22] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *Symposium on Computer Animation*, 2017. 3
- [23] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *CVPR*, 2013. 5
- [24] P. Li, H. Ling, X. Li, and C. Liao. 3d hand pose estimation using randomized decision forest with segmentation index points. In *ICCV*, 2015. 2
- [25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG) (Proc. SIGGRAPH Asia)*, 2015. 3
- [26] A. Makris and A. Argyros. Model-based 3d hand tracking with on-line hand shape adaptation. 2015. 2
- [27] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Generated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 1, 2, 6
- [28] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*, 2017. 2, 5
- [29] J. Nocedal and S. J. Wright. *Nonlinear Equations*. Springer, 2006. 7
- [30] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, 2015. 2
- [31] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Markerless and efficient 26-dof hand pose recovery. In *ACCV*, 2010. 1, 2
- [32] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011. 2, 3, 7
- [33] P. Panteleris and A. Argyros. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. *Hands17 Workshop ICCV*, 2017. 1, 2
- [34] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, 2018. 2, 6, 7
- [35] T. Piumsomboon, A. Clark, M. Billingham, and A. Cockburn. User-defined gestures for augmented reality. In *IFIP Conference on Human-Computer Interaction*, 2013. 1
- [36] G. Poier, D. Schinagl, and H. Bischof. Learning pose specific representations by predicting different views. In *CVPR*, 2018. 1, 2
- [37] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *CVPR*, 2014. 2
- [38] D. A. R. McKee, D. McKee and E. Pailla. Nz sign language exercises. *Deaf Studies Department of Victoria University of Wellington*. 5
- [39] J. Romero, H. Kjellström, and D. Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010. 1, 2

- [40] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 2017. 1, 2, 3, 4, 6, 7
- [41] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3d hand pose reconstruction using specialized mappings. In *ICCV*, 2001. 1, 2
- [42] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 2004. 5
- [43] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *ACM CHI*, 2015. 2
- [44] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 1, 2, 4, 5, 7
- [45] A. Sinha, C. Choi, and K. Ramani. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *CVPR*, 2016. 2
- [46] J. Song, G. Sörös, F. Pece, S. R. Fanello, S. Izadi, C. Keskin, and O. Hilliges. In-air gestures around unmodified mobile devices. In *ACM Symposium on User Interface Software and Technology*, 2014. 1
- [47] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8
- [48] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015. 2
- [49] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016. 2, 5
- [50] S. Sridhar, H. Rhodin, H.-P. Seidel, A. Oulasvirta, and C. Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *3DV*, 2014. 1, 2, 3
- [51] B. Stenger, P. R. Mendonça, and R. Cipolla. Model-based 3d tracking of an articulated hand. In *CVPR*, 2001. 1, 2
- [52] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, 2015. 2
- [53] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *ICCV*, 2015. 2
- [54] J. Taylor, R. Stebbing, V. Ramakrishna, C. Keskin, J. Shotton, S. Izadi, A. Hertzmann, and A. Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *CVPR*, 2014. 3
- [55] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018. 3, 7
- [56] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 3, 7
- [57] A. Thayananthan, B. Stenger, P. H. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR*, 2003. 1, 2
- [58] A. Tkach, M. Pauly, and A. Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM TOG*, 2016. 3
- [59] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM ToG*, 2014. 2
- [60] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3
- [61] C. Wan, T. Probst, L. Van Gool, and A. Yao. Dense 3d regression for hand pose estimation. In *CVPR*, 2018. 1, 2
- [62] X. Wu, D. Finnegan, E. O'Neill, and Y.-L. Yang. Handmap: Robust hand pose estimation via intermediate dense guidance map supervision. In *ECCV*, 2018. 1, 2
- [63] Y. Wu, J. Y. Lin, and T. S. Huang. Capturing natural hand articulation. In *ICCV*, 2001. 1, 2
- [64] C. Xu and L. Cheng. Efficient hand pose estimation from a single depth image. In *ICCV*, 2013. 2
- [65] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 1, 2, 5, 6
- [66] Y. Zhou, J. Lu, K. Du, X. Lin, Y. Sun, and X. Ma. Hbe: Hand branch ensemble network for real-time 3d hand pose estimation. In *ECCV*, 2018. 1, 2
- [67] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2
- [68] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 1, 2, 5, 6, 7, 8