Using Deep Learning-based Hand Pose Estimation for handling occlusion

1st Roumaissa BEKIRI
Computer Science Department
LESIA Laboratory
Biskra University, Algeria
roumaissa.bekiri@univ-biskra.dz

2nd Mohamed Chaouki BABAHENINI
Computer Science Department
LESIA Laboratory
Biskra University, Algeria
mc.babahenini@univ-biskra.dz

***Abstract*—Hand pose estimation is a significant research topic for various computer vision applications. Nonetheless, reliable and robust pose estimation with existing methods remains challenging due to the complex anatomy of the hand and the varying shapes and sizes of hands. The traditional approach involved using depth sensors or multi-camera setups. However, with the advent of deep learning, there has been a shift towards using deep neural networks to learn, grasp, and manipulate objects accurately.**

In this paper, we propose an end-to-end framework called "ResUnet network" that can efficiently detect and estimate the position of a human hand from a monocular RGB image. The architecture includes three modules, feature extraction, 2D pose regression, and 3D hand estimation. The first module extracts the feature maps of the cropped hand to generate 2D heatmaps. The second module uses the previous outputs to regress the 2D pose coordinates employing Latent Heatmaps Representation (LHR). The last module concatenates the intermediate features with the upsampling block to process 3D regression and predict the 3D bones using a tree structure of the hand. Quantitative and qualitative results on three datasets GANerated, SynthHands, and Stereo Hand Pose Tracking Benchmark(STB), consistently demonstrate that our regression approach outperforms the current state-of-the-art hand pose estimation methods.

***Index Terms*—Hand pose estimation, occlusion, RGB image, deep learning, Human-Computer Interaction**

I. Introduction

Hand pose estimation has emerged as a critical focus within the field of computer vision, finding applications in various domains such as Virtual Reality (VR), Augmented Reality(AR), Mixed Reality(MR), and Human-Computer Interaction(HCI). These applications encompass areas such as sign language interpretation, activity detection, gesture recognition, and interactive gaming. The evolution of AR/VR/MR technology, including devices like VR headsets, Head-Mounted Displays(HMDs), wearable glasses, and technologies such as Microsoft Hololens, has empowered human hands to engage with virtual objects in real-time. Additionally, modern cameras are capable of capturing high-resolution depth images, due to commercial options such as Kinect or Intel Realsense, which are readily accessible and provide depth data.

In contrast, hand pose estimation adopts a regression approach to reconstruct the hand's skeletal structure in a 3D space. However, despite extensive research efforts (as evidenced by studies such as [1], [3], [12]), addressing these challenges remains difficult due to various inherent complexities. These challenges include issues such as occlusion, the similarity in appearance of fingers, rapid hand movements, and the high degree of freedom associated with hand articulation. Moreover, the presence of noise and variability stemming from diverse lighting conditions, camera angles, and hand shapes further accentuates the difficulty of these tasks.

The rapid progress of deep learning methodologies in the realm of computer vision has significantly mitigated many of the challenges previously mentioned in hand pose estimation. Researchers have dedicated substantial time and energy to tackling this pressing issue, resulting in a multitude of solutions.

In summary, our research makes the following significant contributions:

- We introduce an innovative deep-learning framework designed for the comprehensive estimation of both 2D and 3D hand poses from an individual's perspective, reducing the need for specialized equipment.
- We address the challenge of hand occlusion during interactions with objects by leveraging our "ResUnet network". This network is capable of reliably predicting hand poses by utilizing two key representations: the Latent Heatmap Representation (LHR) and the tree structure of the hand.
- To enhance the accuracy and quality of our training data, we apply data augmentation techniques based on this network. This augmentation improves the effectiveness of estimating hand interactions.
- Through extensive analysis, we demonstrate the efficiency and robustness of our approach. We validate our proposal by conducting tests on various synthetic benchmark datasets and comparing its performance with that of existing methods.

II. System Overview

A. Architecture

Our main objective is to propose a novel deep-learning architecture to estimate 2D and 3D poses from a single RGB

image destined to resolve the occlusion problem.
The 3D hand pose is represented by a sequence of 3D joint coordinates,$\Phi^{3D} = \{\phi\}_{k=1}^{K} \in T_{3D}$ where T_{3D} is our case 3D-dimensional hand joint space, with K= 21. The 2D hand pose estimation is depicted by a two-dimensional array joint coordinated,
where $\Phi^{2D} = \{\phi\}_{k=1}^{K} \in S_{2D}$ is the K-dimensional hand joint space with K= 21.

The proposed framework, the "ResUnet" network, combines ResNet-34 layers with Unet as a based backbone, which is particularly effective for tasks where input and output have similar sizes. Unet network comprises two main paths: The contracting path used a pre-trained ResNet-34, a 34-layer ResNet network to extract the main features from RGB cropped hand image $I \in \mathbb{R}^{128\times128\times3}$. The second path of Unet, called the Expansive path, includes four continuous multi-features combined with upsample Blocks named Unet-Block, as shown in Fig.1

The first Unet-Block uses the fusion of two features as input $\mathbb{F}_g = \{\mathbb{F}_4, \mathbb{F}_3\}$ and outputted the grouped features $\mathbb{F}_{out}$ and employ bilinear upsample to increase the quality of input images and acquire multi-scale features. After that, estimate 2D heatmaps before passing it to the upcoming Unet-Block. The remainder Unet-Blocks have a similar format with distinct input features. We concatenate for each block the upsampling layer with the respective feature vector to be fed later in the convolutional layers, which is denoted $\mathbb{F}_{skip}$.

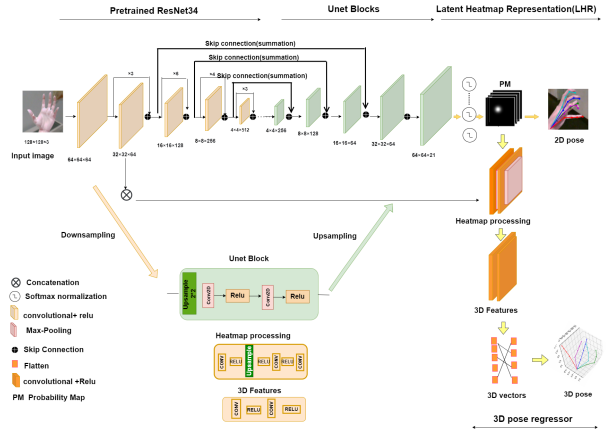


Fig. 1: The overall pipeline of our architecture for estimating 2D and 3D pose regression.

The output of the last block with a size of 64. Furthermore, we add upsample layer with bilinear mode followed by two convolution layers separate with RELU as an activation function to obtain nonlinear transformation and enhance our architecture ability to fit data. The output is a tensor of size (21, 64). We get the 2D feature heatmaps from the Unet-blocks, including pose data from intermediate outputs.

We employ Latent Heatmap Representation(LHR), which outperforms others and is deemed the more reliable method for predicting the 2D pose coordinates throughout the prediction of a 2D heatmap. This approach represents the hand pose as a 2D latent heatmap where each pixel corresponds to a specific joint location. The value at each pixel in the heatmap indicates the likelihood of the corresponding joint being present at that location. The input to the LHR network is a single image fitted into the network. Once the network processes the image, a 2D heatmap is generated using a UnetBlock with a skip connection. The outputted has saved the features of the learnable model as latent variables F_k^{2D} to approximate 2D latent heatmaps.
This function map converts the tensor values to a probability distribution, so the sum of the values in each channel is strictly added to one. This ensures that the probabilities assigned to each keypoint of the hand are normalized and can be interpreted as a 2D probabilities map. As mentioned in formula 2:

$$PM_i(p) = \frac{exp(\beta_i F_i^{2D(p)})}{\sum_{p'\in\Omega} exp(\beta_i F_i^{2D(p')})} \tag{1}$$

While p represents the probability map point,Ω is the sequence of all pixels on the 2D feature map F_i^{2D} of i^{th} joint, $\beta_i$is a factor that can be learned to control the probability map's spread.
The 2D joint coordinates of the k^{th} keypoint are then derived as the weighted mean for x and y coordinates, where weights are values from the normalized heatmaps and the generated x, y coordinates fall within the range[0, image width].

$$p_k = \sum_{p\in\Omega} PM(p).p \tag{2}$$

The second branch of our overall framework is to regress the 3D pose of a hand. We conduct such representation, a tree structure of a hand, as shown in Fig.4. This representation predicts the bones instead of joints because it is more accurate and stable to define.
For notation consistency, the bone is defined as $\beta_k = \{\beta_k | k = 1, ...k\}$

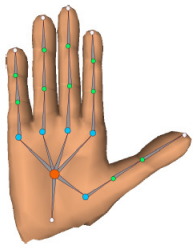


Fig. 2: The proposed hand bones representation.

For each joint J^{th}, we refer to it as related bone as a direct vector pointing from a bone to its origin. As illustrated in the following equation:

$$B_i = J_i - J_{parent(i)} \tag{3}$$

Where the $J_{parent(i)}$ is a predefined function that returns the index of the parent joint of the current joint J_i. Our network backbone structure comprises two convolutional layers. Following each convolutional layer with a Rectified Linear Unit (ReLU) as the activation function that produces 3D feature maps. Every two convolutional layers are preceded by a max-pooling layer. Then, we concatenate the process 3D features with intermediate features. We flatten by employing two fully connected layers. The model outputs a tensor of size (20, 3) representing 3D coordinates of K=20 bones. When calculating the global coordinate system of a specific joint, the local coordinates of all bones along the path are summed together. During learning, bones are supervised. For that, we provide the following equation to produce the bones lossL_β:

$$L_\beta = \frac{1}{2}\sum_{j=1}^{J}\left\|\beta^{pred} - \beta^{gt}\right\|^2 \tag{4}$$

III. Implementation details

We conducted experiments for our method using a personal computer. To implement our deep learning approach, we employed PyTorch version 1.8 along with CUDA version 10.1 and cuDNN version 7.6.4. To accelerate the training process, we utilized the Nvidia GeForce GTX1070 graphics card with 64-bit support. The training process was carried out successfully on a system equipped with 16GB of memory and an AMD Ryzen 53600 6-Core processor.

A. Dataset Evaluation

We quantitatively assess our proposed framework using three distinct datasets:

Ganerated dataset: This dataset is among the most recent and noteworthy RGB-based datasets, specifically designed for approximating hand poses during interactions with obscured objects. It comprises approximately 330,000 images of synthetic hand poses, annotated in three dimensions with a model featuring 21 joints.

SynthHands: is an RGBD hand pose estimation dataset that includes 63.5K color and depth images with a resolution of 640x480. These images were captured from five egocentric viewpoints of male and female hands using an Intel RealSense camera. The dataset offers diverse variations, encompassing differences in skin color, shape, background clutter, wrist and arm rotation, and hand-object interactions involving seven distinct object shapes and 145 textures.

Stereo Hand Pose Tracking Benchmark (STB): STB is widely adopted for training and validating RGB-based 3D hand pose estimation methods. It contains 18,000 stereo and depth images, with 15,000 designated for training and 3,000 for testing. The stereo images were captured using a Point Grey Bumblebee2 stereo camera, while depth images were obtained using an Intel RealSense F200 depth camera, all with a resolution of 640x480.

B. Metric Evaluation

In order to assess the precision of our proposed method and benchmark it against state-of-the-art techniques, we employ the three most commonly used metrics in hand pose estimation:

EPE (End-Point-Error): EPE quantifies the average 3D Euclidean distance error between all joints calculated by our method and the ground truth. In the context of 3D hand estimation, these distances are expressed in millimetres (mm), while for 2D estimation, they are measured in pixels (px).

PCK (Percentage of Correct Keypoints): PCK is a commonly used error metric for 3D hand pose estimation, assessing the accuracy of localizing specific keypoints within a defined matching threshold.

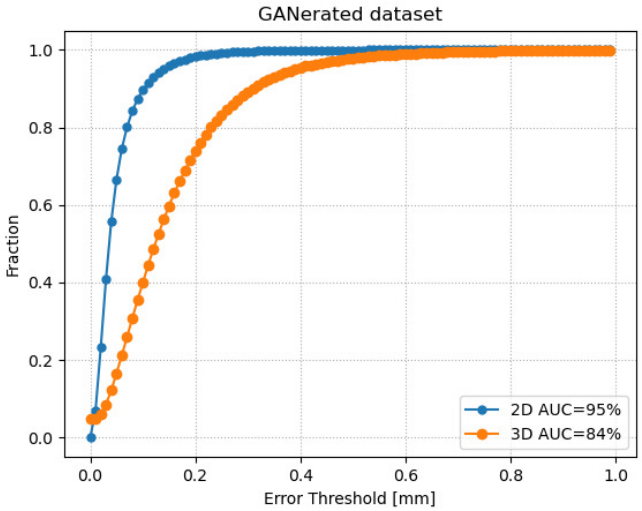
AUC (Area Under the Curve): AUC is regarded as the optimal criterion for evaluating a model's correctness, as it measures the proportion of true keypoints (PCK) across various error thresholds.

IV. Experimental Results

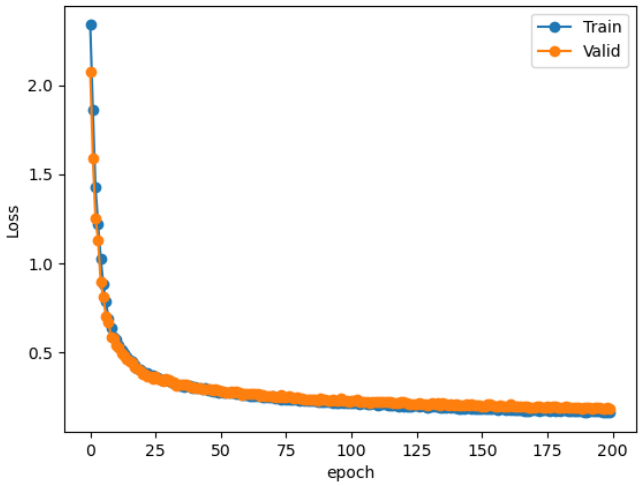
We systematically assess the performance of our method in both quantitative and qualitative terms, specifically focusing on its ability to learn 2D and 3D hand pose regression and address the challenge of occlusion. Additionally, we showcase the versatility of our approach by achieving accurate predictions for datasets involving single hands.

A. Quantitative Results

As illustrated in Figure 3(b), we employ the Mean Square Error (MSE) as a metric to evaluate the error during both training and validation phases on a GANerated dataset.



(a) 2D and 3D PCK metrics under the perspective threshold of our architecture.



(b) Mean Square Error(MSE) applied during training and validation test data.

Fig. 3: Quantitative Evaluation of our proposed approach on GANerated dataset.

SynthHANDS dataset

Comparing the 2D and 3D PCK curves as illustrated in Fig. 4. shows the superiority of our model prediction that aims to estimate 2D and 3D hand poses on the largest synthHands dataset. Additionally, because no previous work is employed in their comparison state on SynthHANDS, only the work of [7].

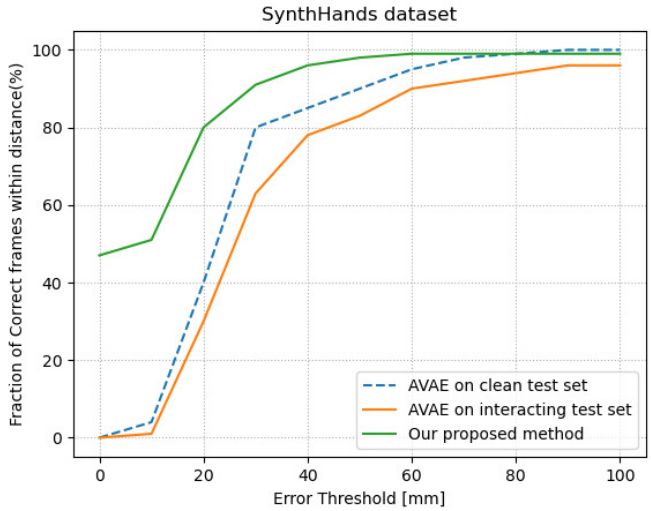


Fig. 4: 3D PCK on SynthHands datasets compared with Li et al. [7] and their differents training options and architectures.

Stereo Tracking Benchmark dataset

To confirm the enhanced performance of our approach, we conducted testing on a third benchmark dataset, Stereo DS, using the same metrics as in our previous experiments. On the other hand, we perform a comparison with the recent methods using the 3D PCK evaluation metrics as illustrated in Fig. 5 [4], [8]–[11], [13] that have gained a lot of attention due to the best results using STB dataset. We demonstrate that our AUC[20-50](mm)is about 0.999 is superior to all recent approaches.

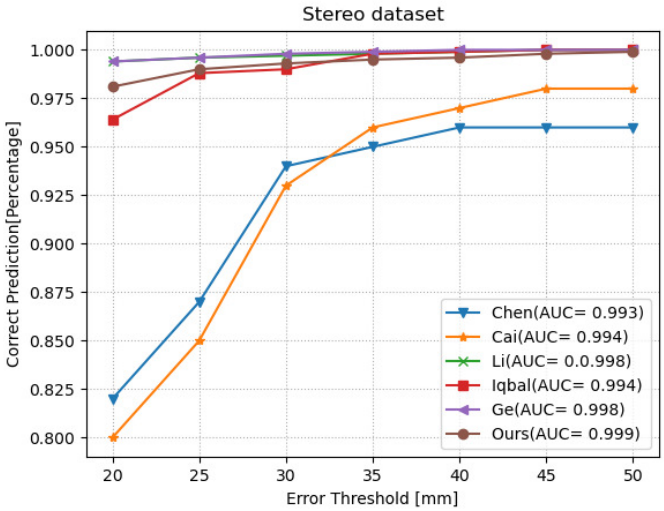


Fig. 5: Comparison with the SOTA methods [2], [3], [5]–[7] on the STB dataset using 3D PCK. The X-axis is the threshold values(i.e., maximum permitted distance between estimated and ground truth), and Y-axis is the 3D PCK over the perspective threshold. The ”AUC” shown in this curve is between 20 and 50[mm].

B. Qualitative Results

In addition to quantitative outputs, we also conduct a qualitative assessment on three readily available datasets: GANerated, SynthHands, and STB. This evaluation serves to showcase the effectiveness of our method in consistently predicting poses from diverse perspectives, encompassing 2D and 3D keypoints as well as 2D skeletal structures.

Furthermore, in order to underscore the practicality of our approach, we extend our analysis to include previously unseen images, thereby enhancing the reliability of our results in challenging scenarios characterized by significant occlusion, as depicted in Figures 6, 7, and 8.

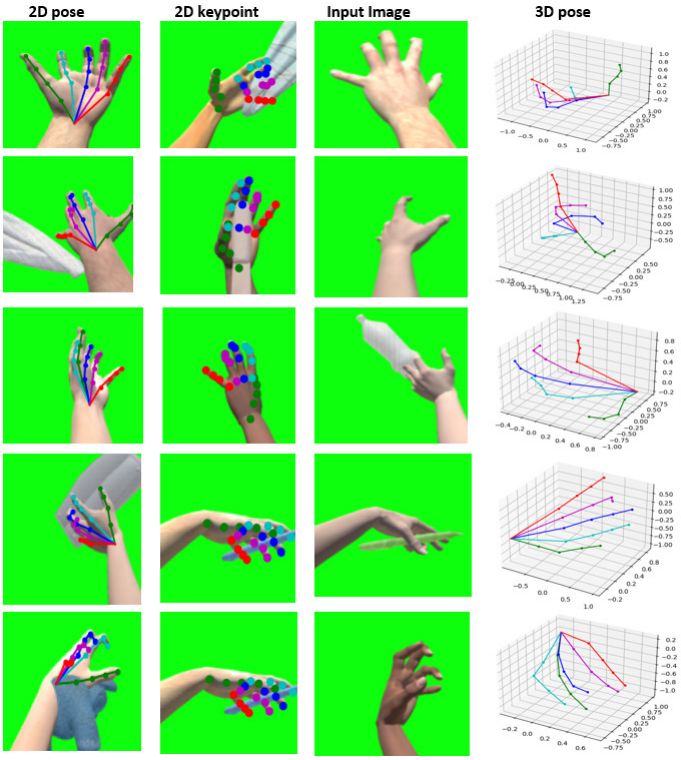


Fig. 6: Qualitative results on SynthHANDS dataset [10].

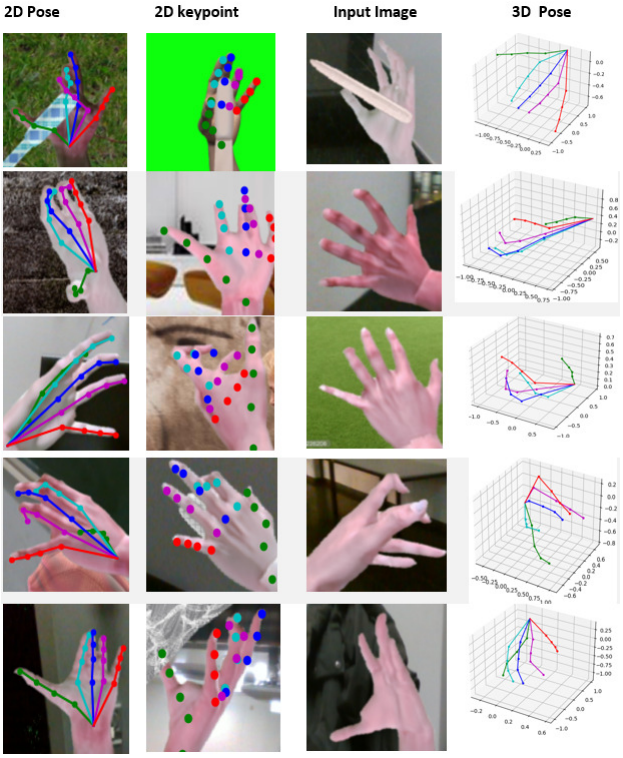


Fig. 7: Qualitative results on GANerated dataset [8].

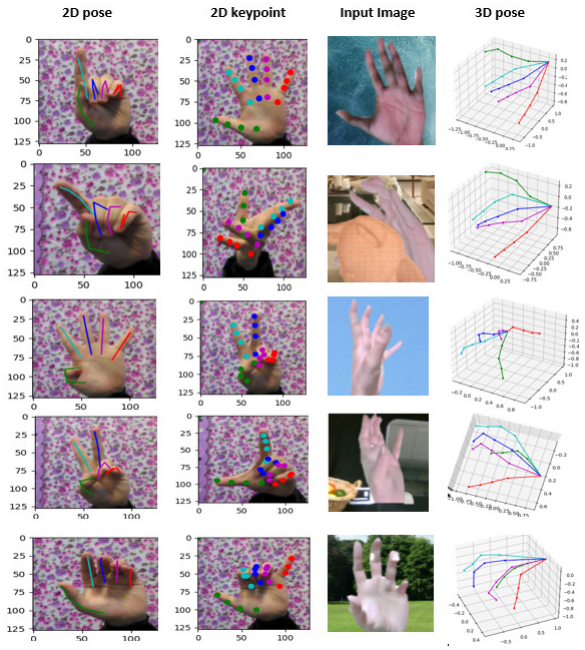


Fig. 8: Qualitative results on STB dataset [12].

V. Conclusion

In this study, we have tackled the issue of estimating the 2D and 3D hand pose from a single RGB image with solving the occlusion issue during hand interaction. Our contribution is based on developing a deep learning model ”ResUnet” network, combining the ResNet and Unet basic network. We

have introduced a 2D regression pose using a Latent Heatmap Representation(LHR) from RGB input for estimating 2D hand pose. We have applied a tree structure of the Hand to predict bones because it is more stable than joints and powerful. Quantitative and qualitative results show that our proposed framework significantly outperforms better estimation on different viewpoints pose and in difficult occlusion cases. Further, with this model, we can accurately forecast the joint angles. We compare the effectiveness of these components to other state-of-the-art methods and find that our approach is superior to previous research.

REFERENCES

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019.

[2] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018.

[3] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Hui Tang, Yufan Xue, Xiaohui Xie, Yen-Yu Lin, and Wei Fan. Generating realistic training images based on tonality-alignment generative adversarial networks for hand pose estimation. *arXiv preprint arXiv:1811.09916*, 2018.

[4] Endri Dibra, Silvan Melchior, Ali Balkis, Thomas Wolf, Cengiz Oztireli, and Markus Gross. Monocular rgb hand pose inference from unsupervised refinable nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1075–1085, 2018.

[5] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018.

[6] Moran Li, Jialong Wang, and Nong Sang. Latent distribution-based 3d hand pose estimation from monocular rgb images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4883–4894, 2021.

[7] Shile Li, Haojie Wang, and Dongheui Lee. Hand pose estimation for hand-object interaction cases using augmented autoencoder. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 993–999. IEEE, 2020.

[8] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[9] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[10] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1154–1163, 2017.

[11] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1106–1113, 2014.

[12] Yu Zhang, Chi Xu, and Li Cheng. Learning to search on manifolds for 3d pose estimation of articulated objects. *arXiv preprint arXiv:1612.00596*, 2016.

[13] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.