MM-Hand: 3D-Aware Multi-Modal Guided Hand Generative Network for 3D Hand Pose Synthesis

Duc Hoang*
hoangd@tamu.edu
Texas A&M University

Zhenyu Wu*
wuzhenyu_sjtu@tamu.edu
Texas A&M University

Shih-Yao Lin
shihyaolin@tencent.com
Tencent America

Yusheng Xie†
yushx@amazon.com
Amazon Web Services

Liangjian Chen
liangjc2@ics.uci.edu
University of California, Irvine

Yen-Yu Lin
lin@cs.nctu.edu.tw
National Chiao Tung University

Zhangyang Wang
atlaswang@utexas.edu
University of Texas at Austin

Wei Fan
davidwfan@tencent.com
Tencent America

ABSTRACT

Estimating the 3D hand pose from a monocular RGB image is important but challenging. A solution is training on large-scale RGB hand images with accurate 3D hand keypoint annotations. However, it is too expensive in practice. Instead, we develop a learning-based approach to synthesize realistic, diverse, and 3D pose-preserving hand images under the guidance of 3D pose information. We propose a 3D-aware multi-modal guided hand generative network (**MM-Hand**), together with a novel geometry-based curriculum learning strategy. Our extensive experimental results demonstrate that the 3D-annotated images generated by *MM-Hand* qualitatively and quantitatively outperform existing options. Moreover, the augmented data can consistently improve the quantitative performance of the state-of-the-art 3D hand pose estimators on two benchmark datasets. The code will be available at *https://github.com/ScottHoang/mm-hand*.

KEYWORDS

3D Hand-Pose, Multi-Modal, Conditional Generative Adversarial Nets

ACM Reference Format:
Duc Hoang, Zhenyu Wu, Shih-Yao Lin, Yusheng Xie, Liangjian Chen, Yen-Yu Lin, Zhangyang Wang, and Wei Fan. 2020. MM-Hand: 3D-Aware Multi-Modal Guided Hand Generative Network for 3D Hand Pose Synthesis . In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413555

1 INTRODUCTION

3D hand pose estimation is an important and active research topic due to its versatile applications in sign language recognition [45], HCI (human-computer interaction, healthcare) [24–26], and entertainment [2, 17, 30]. Some HCI applications such as typing[1] highly rely on accurately estimated hand poses. Conventional HCI applications use depth sensors to capture hand information and infer hand poses [12, 22, 23, 41, 43]. In the past few years, there has been growing interest in HCI applications where pose estimation/tracking from single RGB images is carried out, usually to take advantage of the ubiquity of RGB cameras. For example, the exciting dancing application [5] and the popular personalized media creation app, Humen.ai[2], both require RGB-based body-hand pose estimation and tracking. However, without depth information, estimating the 3D hand pose from a monocular RGB image is an ill-posed problem. Recent studies [1, 3, 4, 10, 47, 54] resort to a large number of training images with the corresponding 3D hand pose annotations. Nevertheless, acquiring large-scale hand datasets with 3D annotations is labor-intensive and very expensive. Compared with 2D annotations, 3D annotations of real RGB hand images are much more difficult to be reliably labeled by humans.

A promising alternative emerges from synthesizing hand images. With 3D computer graphics software (*e.g.* Blender and Maya) and chroma key compositing, synthetic hand data can be generated with accurate 3D annotations [32, 54]. In this way, a large hand image dataset with various poses, skin textures, shapes background, lighting conditions, and object interactions, can be systematically synthesized. However, synthesizing hand images to train 3D hand pose estimators is limited in two aspects. Firstly, estimators trained on the synthetic data often fail to generalize due to the visual domain gap between non-photo-realistic synthetic data and real images. Such a domain gap has been widely documented and investigated in prior work [7, 8, 31, 37]. Secondly, building hand models with various textures or shapes requires laborious 3D geometric modeling and rendering, which are relatively less studied.

Recent work on pose guided person image generation [9, 27–29, 34, 35, 38, 40, 53] has made significant progress by human

*The first two authors Hoang and Wu contributed equally to this research.
†Work done prior to Amazon.

[1] https://uploadvr.com/frl-pinchtype-ar-vr-keyboard/
[2] https://www.humen.ai

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413555

body pose transfer, *e.g.* swapping the pose of a person image into a target pose, while maintaining other visual appearance details. Those generated realistic person images are blended into the training process to improve the person re-ID task. Inspired by those, our goal is to generate realistic and diverse hand images with accurate pose annotations. However, there are two challenges specific to the hand domain [20]: *occlusion* (i.e., various 3D hand movements will always make some finger parts invisible from 2D images) and *self-similarity* (e.g., the five fingers of the same hand share similar appearance and structure, making them indistinguishable). This paper makes the following three-fold contributions:

- Our proposed framework, *3D-Aware Multi-modal Guided Hand Generative Network* (**MM-Hand**), carries out the first efforts to generate hand images under the guidance of 3D poses, improving the realism, diversity and 3D pose preserving of the generated images simultaneously.
- *MM-Hand* is trained with a novel geometry-based curriculum learning strategy. Starting with easy pose-images pairs, we gradually increase the training task difficulty.
- Extensive experiments demonstrate that our generated hand images can consistently improve 3D hand pose estimation, across two strong pose estimators and two hand pose datasets.

2 RELATED WORK

2.1 Data Augmentation for Hand Pose Estimation

Generative adversarial networks (GANs) have demonstrated strong promise in synthesizing training data [6, 19, 51]. Shrivastava *et al.* [37] proposes SimGAN, which improves the realism of a simulator's rendered data by using unlabeled real data while preserving the annotation information from the simulator. The processed data by SimGAN are then leveraged to train a hand pose estimator. Mueller *et al.* [31] present GeoConGAN, whose generated image preserves hand pose structure by a geometric consistency loss. These approaches for data augmentation focus on image translation from the synthetic hands to real hands (based on an existing synthetic simulator). In contrast, we directly generate realistic hand images from 3D poses and synthetic depth maps, which is more challenging.

Zimmermann and Brox [56] introduce the first large-scale, multi-view hand image dataset, which includes both 3D hand pose and shape annotations. The annotation is achieved by an iterative, semi-automated "human-in-the-loop" approach, which includes hand fitting optimization to infer the 3D pose and shape for each sample.

2.2 Pose Guided Person Image Generation

Isola *et al.* [13] propose *Pix2Pix* to translate sketch to synthesize photos from label maps, reconstructing objects from edge maps and colorizing images, using paired data. Zhu *et al.* [52] present the *CycleGAN* to work under the unpaired settings by introducing a cycle consistency loss between source and target domains. Since then, (unpaired) image translation has been popular in various applications including image enhancement [14, 18], style transfer [49], interactive image editing [48], and domain adaptation [11, 33].

As the first work focusing on pose-guided human image generation, Ma *et al.* [28] proposes a two-stages network PG^2, where a person image under the target pose is first coarsely generated and then refined. Ma *et al.* [29] then improves the control over image generation by disentangling and separately encoding the three modes of variation (foreground, background, and pose information) into embedding features. Esser *et al.* [9] combine *Variational Auto-Encoder (VAE)* [16] and *U-Net* [36] to disentangle appearance and poses. Their work presents a *U-Net* that maps from shape to target image, conditioned on a *VAE* latent representation for preserving appearances.

Si *et al.* [38] adopts multistage adversarial losses separately for the foreground and background generation, which fully exploits the multi-modal characteristics of generative loss to generate more realistic looking images. Neverova *et al.* [34] proposes a combination of surface-based pose estimation and deep generative models to perform accurate pose transfer. Siarohin *et al.* [39] introduces deformable skip connections in the generator to deal with pixel-to-pixel misalignments caused by the pose differences, together with a nearest-neighbor loss. Pumarola *et al.* [35] proposes a pose conditioned bidirectional generator that maps back the initially rendered image to the original pose, hence being directly comparable to the input image without any training image. Li *et al.* [21] proposes to estimate dense and intrinsic 3D appearance flow to guide the transfer of pixels between poses better. They generate 3D flow by fitting a 3D model to the given pose pair and project them back to the 2D plane to compute the dense appearance flow for training. Song *et al.* [40] address unsupervised person image generation by decomposing it into semantic parsing transformation and appearance generation. Zhu *et al.* [53] propose a progressive pose attention transfer network composed of a cascaded *Pose-Attentional Transfer Blocks (PATBs)*. Liu *et al.* [27] tackle the human motion imitation, appearance transfer, and novel view synthesis within a unified framework. Unlike pose guided person images generation, pose guided hand generation can be much more subtle and difficult, due to the inherently strong self-similarity and the self-occlusion.

2.3 3D Hand Pose Estimation from a Single Image

Zimmermann and Brox [54] propose the first learning-based approach to estimate the 3D hand pose from a single RGB image. Their approach consists of three building blocks: *HangSegNet* for obtaining hand mask by segmentation, *PoseNet* for localizing a set of hand keypoints in score maps, and *PosePrior* network for estimating 3D structure conditioned on the score maps). Cai *et al.* [4] proposes a weakly-supervised method to generate depth maps from predicted 3D poses, which then serves as weak supervision for 3D pose regression. Chen *et al.* [6] presents a *Depth-image Guided GAN (DGGAN)* to generate realistic depth maps conditioned on the input RGB image and use the synthesized depth maps to regularize the 3D hand pose estimators. Some studies [1, 3, 10] tackle the challenging task of 3D hand shape *and* pose estimation from a single RGB image. Boukhayma *et al.* [3] combines a deep convolutional encoder with a generative hand model as the decoder. Ge *et al.* [10] proposes a graph CNN-based method to reconstruct 3D mesh of hand surface that contains rich information of both the 3D hand shape and pose. Baek *et al.* [1] adopts a compact parametric 3D hand model that represents deformable and articulated hand meshes. Yang *et al.* [47] proposes a disentangled VAE allowing for sampling and inference of variation factors, *e.g.* content, poses, and camera views.

3 OUR APPROACH

3.1 Problem Formulation

Goal. Given a target 3D hand pose $\boldsymbol{p}_t$, and a source image $\boldsymbol{I}_{p_s}$ under a source 3D pose $\boldsymbol{p}_s$, our goal is to generate a new image $\tilde{\boldsymbol{I}}_{p_t}$ following the appearance of $\boldsymbol{I}_{p_s}$, under the guidance of $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$. The generation can be formulated as:

$$< \boldsymbol{I}_{p_s}, \boldsymbol{p}_s, \boldsymbol{p}_t > \longrightarrow \tilde{\boldsymbol{I}}_{p_t} \tag{1}$$

Evaluation Protocols. The generated hand image $\tilde{\boldsymbol{I}}_{p_t}$ is expected to resemble the ground truth hand image $\boldsymbol{I}_{p_t}$ in both appearance realisticity and pose consistency. The appearance realisticity is evaluated by natural image quality metric (*e.g.* SSIM and IS). The pose consistency is approximated by pose joints alignment (*e.g.* PCKb). Considering both appearance realisticity and pose consistency, we further evaluate the quality of generated hand images on the visual task of 3D hand pose estimation. Details of the evaluation metrics are given in section 4.1.

Pose Representations. In this paper, we use the 21-joints hand model ($K = 21$). 3D poses are represented as $3 \times K$ matrices: $[J_1^{xyz}, \ldots, J_K^{xyz}] \in \mathcal{R}^{3\times K}(K = 21)$. 2D poses are represented as K probability heat maps: $\{\boldsymbol{H}_i\}_{i=1}^{K} \in \mathcal{R}^{K\times H\times W}(K = 21)$, where $\boldsymbol{H}_i$ is a Gaussian distribution centered at the location of the i_{th} joint.

3.2 Architecture Overview

As is shown in Figure 1, the proposed *3D-Aware Multi-modal Guided Hand Generative Network* (MM-Hand) is composed of 4 modules: *3D pose embedding*, *multi-modality encoding*, *progressive transfer* and *image modality decoding*.

3.2.1 3D Pose Embedding.

Contour Map Embedding. Given the camera's intrinsic matrix $\boldsymbol{K}$ and extrinsic matrix $[\boldsymbol{R} \mid -\boldsymbol{RC}]$, we obtain the matrix $\boldsymbol{P} = \boldsymbol{K}[\boldsymbol{R} \mid -\boldsymbol{RC}]$ which transforms homogeneous 3D world coordinates to homogeneous 2D image coordinates. Firstly, we represent the K joints with 2D coordinates with a sparse pose map, using erosion and dilation. Secondly, we connect the keypoints on finger with solid ellipsoid using different color. Lastly, A palm surrograte is formed by connecting a polygon from the basal of each finger and the wrist. The contours map of $\boldsymbol{c}_{p_s}$ and $\boldsymbol{c}_{p_t}$ are generated as the embeddings of the 3D pose $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$.

Depth Map Embedding. There are datasets that contain depth maps paired with annotated 3D hand poses, *e.g.* ICVL [42], NYU [43] and MSRA [41]. The annotated hand poses in ICVL, MSRA and NYU have 16 joints, 21 joints, and 14 joints. We choose the MSRA dataset annotated with 21 joints to be consistent with those benchmark datasets.

With the help of external datasets with paired depth maps and 3D hand poses, we can learn a mapping from 3D hand pose to depth map. In Figure 2, the depth map generator takes the raw 3D hand pose ($\boldsymbol{p}_s$ and $\boldsymbol{p}_t$) as input and outputs depth maps ($\boldsymbol{d}_{p_s}$ and $\boldsymbol{d}_{p_t}$).

The depth map generator is a Pix2Pix model regulated by a discriminator and a pair of 2D/3D key-points estimators. The key-points estimators are 6-stage HPM [44], and 1-stage HPM [6] , both are pre-trained on the MSRA's pose estimation task. Note that we first project the 3D hand pose onto a 2D image before feeding it to the generator.

3.2.2 Multi-Modality Encoding.

The modality encoder for the image modality, the contour map modality and the depth map modality are consistently adopting two convolution layers. Before encoding, we concatenate $\boldsymbol{c}_{p_s}$ with $\boldsymbol{c}_{p_t}$, and $\boldsymbol{d}_{p_s}$ with $\boldsymbol{d}_{p_t}$. Specifically,

$$\boldsymbol{c}_0 = f_c^e(\boldsymbol{c}_{p_s} \parallel \boldsymbol{c}_{p_t}), \boldsymbol{d}_0 = f_d^e(\boldsymbol{d}_{p_s} \parallel \boldsymbol{d}_{p_t}) \text{ and } \boldsymbol{I}_0 = f_I^e(\boldsymbol{I}_{p_s}), \tag{2}$$

where f_c^e, f_d^e and f_I^e are the modality encoders for the contour maps, the depth maps and the images respectively.

3.2.3 Progressive Transfer.

Our *progressive transfer* module inherits the ResNet Generator proposed in [13, 15]. It consists of multiple cascaded Multi-stream Attentional Blocks (MABs) (shown in Figure 1), variants of the ResNet blocks in [13, 15]. MAB is similar to the Pose Attention Transfer Block (PATB) proposed in [53]. Starting from the initial image modality $\boldsymbol{I}_0$, the contour map modality $\boldsymbol{c}_0$ and depth modality $\boldsymbol{d}_0$, *mm-hand* progressively updates these three modalities through a sequence of MABs. Then deconvolution is used to decode the output image modality $\boldsymbol{I}_N$ to generate $\tilde{\boldsymbol{I}}_{p_t}$. The final contour map modality $\boldsymbol{c}_N$ and depth map modality $\boldsymbol{d}_N$ are discarded after inference.

All MABs share the identical structure. One MAB's output becomes the input for the next MAB block. For example, the input of the n-th block consists of $\boldsymbol{I}_{n-1}$, $\boldsymbol{c}_{n-1}$ and $\boldsymbol{d}_{n-1}$. The specific modules within each MAB are described in the following paragraphs.

Attention Masks. Inspired by Zhu *et al.* [53], the attention masks $\boldsymbol{M}_n$, whose values are between 0 and 1, indicate the importance of every element in the image modality. $\boldsymbol{M}_n$ are computed from the contour modality $\boldsymbol{c}_{n-1}$ and the depth modality $\boldsymbol{d}_{n-1}$. The contour modality $\boldsymbol{c}_{n-1}$ incorporates both the source contour map $\boldsymbol{c}_{p_s}$ and the target contour map $\boldsymbol{c}_{p_t}$. Likewise, the depth modality $\boldsymbol{d}_{n-1}$ incorporates both the source depth map $\boldsymbol{d}_{p_s}$ and the target depth map $\boldsymbol{d}_{p_t}$. The attention mask is computed as element-wise product of $\sigma(f_c(\boldsymbol{c}_{n-1}))$ and $\sigma(f_d(\boldsymbol{d}_{n-1}))$, where σ is an element-wise sigmoid function and f_c and f_d are ResNet Blocks. Specifically, the attention masks $\boldsymbol{M}_n$ are obtained via:

$$\boldsymbol{M}_n = \sigma(f_c(\boldsymbol{c}_{n-1})) \odot \sigma(f_d(\boldsymbol{d}_{n-1})). \tag{3}$$

Image Modality Update. By multiplying the transformed image codes with the attention mask $\boldsymbol{M}_n$, image code $\boldsymbol{I}_n$ at certain locations are either preserved or suppressed. The output of the element-wise product is added by $\boldsymbol{I}_{n-1}$, via a residual connection. The residual connection helps preserve the original image modality. f_I is again a ResNet Block. The image modality $\boldsymbol{I}_n$ is updated by:

$$\boldsymbol{I}_n = \boldsymbol{M}_n \odot f_I(\boldsymbol{I}_{n-1}) + \boldsymbol{I}_{n-1}. \tag{4}$$

Discriminators. We adopt the same two discriminators in Zhu *et al.* [53]: appearance discriminator and pose discriminator. They are denoted as D_a and D_p, respectively. $D_a(\boldsymbol{I}_{p_s}, \tilde{\boldsymbol{I}}_{p_t})$ describes how well $\tilde{\boldsymbol{I}}_{p_t}$ resembles the source image $\boldsymbol{I}_{p_s}$ in appearance. $D_p(\boldsymbol{p}_t, \tilde{\boldsymbol{I}}_{p_t})$ shows how well $\tilde{\boldsymbol{I}}_{p_t}$ is aligned with the target pose $\boldsymbol{p}_t$.

3.2.4 Image Modality Decoding.

We take the output image modality $\boldsymbol{I}_N$ from the N-th MAB, and generated $\tilde{\boldsymbol{I}}_{p_t}$ from $\boldsymbol{I}_N$ via the image modality decoder.

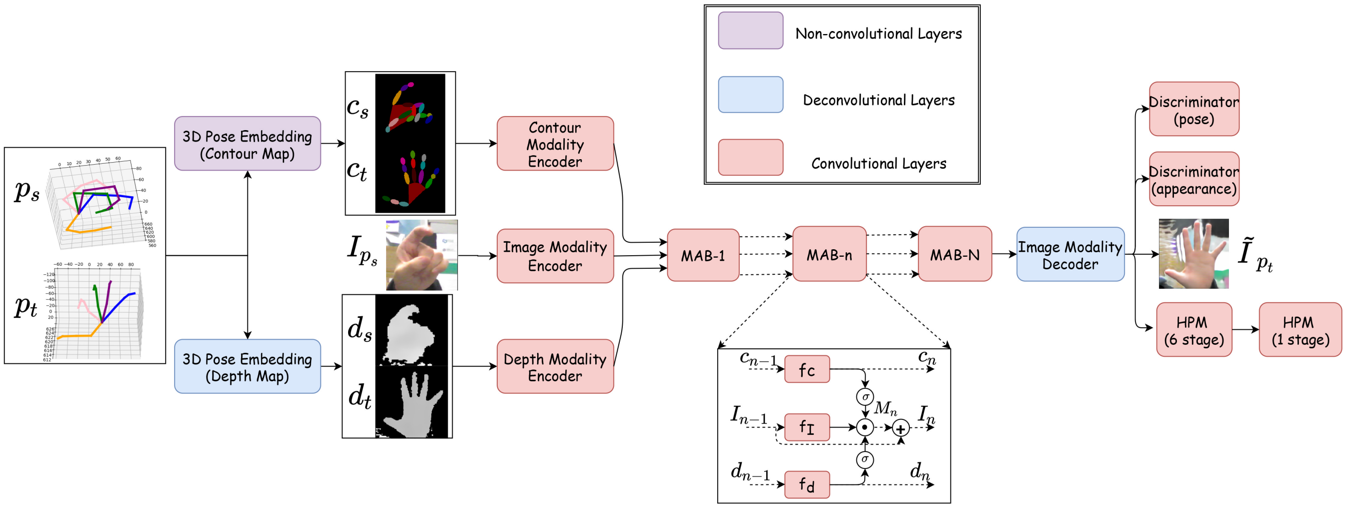

Figure 1: An overview of our proposed *mm-hand* Model. *mm-hand* mainly consists of four modules: *3D Pose Embedding*, *Multi-modality Encoding*, *Progressive Transfer* and *Image Modality Decoding*.

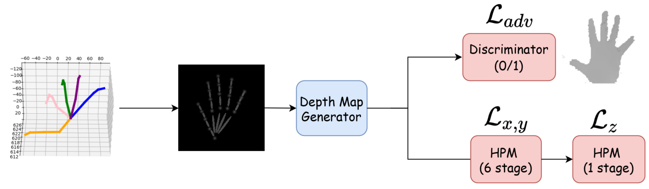

Figure 2: A detailed look into our the depth map generation model for 3D pose embedding

3.3 Training

3.3.1 Geometry-based Curriculum Training and Inference with Nearest Neighbor Match. Given two 3D hand poses $\boldsymbol{u}$ and $\boldsymbol{v}$, we define the pose distance between $\boldsymbol{u}$ and $\boldsymbol{v}$ as:

$$d(\boldsymbol{u},\boldsymbol{v}) = \frac{1}{\pi}cos^{-1}\left(\frac{< f(\boldsymbol{u}), f(\boldsymbol{v}) >}{\|f(\boldsymbol{u})\| \, \|f(\boldsymbol{v})\|}\right) \qquad (5)$$

where $f(\cdot)$ describes the "identity" of a hand pose. Each hand pose is expressed as the concatenation vector of its 21 3D keypoints, *i.e.* $\boldsymbol{u} = [u_x^1, u_y^1, u_z^1, \ldots, u_x^{21}, u_y^{21}, u_z^{21}] \in \mathbb{R}^{63}$. The "identity" $f(\boldsymbol{u})$ of $\boldsymbol{u}$ is defined as a vector of 3 components: $d_{tips}(\boldsymbol{u})$, $d_c(\boldsymbol{u})$ and $\sqrt{\mathcal{A}(\boldsymbol{u})}$.

$$f(\boldsymbol{u}) = [d_{tips}(\boldsymbol{u}), d_c(\boldsymbol{u}), \sqrt{\mathcal{A}(\boldsymbol{u})}] \qquad (6)$$

$d_{tips}(\boldsymbol{u})$ is a collection of pair-wise euclidean distance between the tip keypoint (thumb, index, mid, ring and pinky) and the palm. $d_c(\boldsymbol{u})$ is the distance between the centroid and the palm. $\sqrt{\mathcal{A}(u)}$ is the square root of the contour area of the convex hull formed by the 21 2D keypoints after projection.

Geometry-based Curriculum Training (GCT). Based on the observation that EPE is positively correlated with the pose distance (shown in Figure 3), we hypothesize that the level of difficulty to generate target hand image $\tilde{I}_{p_t}$ from source hand image I_{p_s} is positively correlated with the 3D pose distance between $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$. Hence in the training stage, we first randomly make pairs of hand images. Then we compute the 3D pose distance for each paired image. For each training epoch, *mm-hand* is fed by the data loader with hand pairs progressively from the easiest (smallest pose distance) pair to the hardest (largest pose distance) pair.

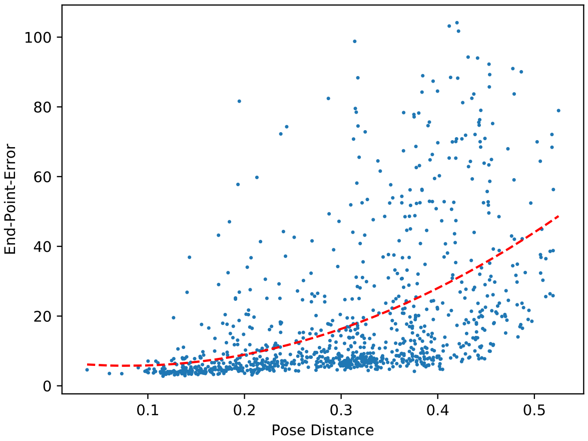

Figure 3: On STB, we first randomly select 1,000 hand images I_{p_s}s with their paired poses $\boldsymbol{p}_s$s from the testing set $\mathcal{X}_{te}$. Then we randomly select 1,000 hand images I_{p_t}s with their paired poses $\boldsymbol{p}_t$s from the training set $\mathcal{X}_{tr}$. We randomly make 1,000 pairs of $(I_{p_s}, \boldsymbol{p}_s)$ and $(I_{p_t}, \boldsymbol{p}_t)$. We generate 1000 $\tilde{I}_{p_t}$s from $< I_{p_s}, \boldsymbol{p}_s, \boldsymbol{p}_t >$. We empirically found the End-Point-Error (EPE) between the $\tilde{I}_{p_t}$ and the ground-truth I_{p_t} is positively correlated with the pose distance (Eq. (5)) between I_s and I_t.

Inference with Nearest Neighbor Match (INNM). In the inference stage, given a target 3D hand pose $\boldsymbol{p}_t$, we find the best matched source hand image I_{p_s} in the training hand images whose pose $\boldsymbol{p}_s$ is closest to $\boldsymbol{p}_t$ in pose distance.

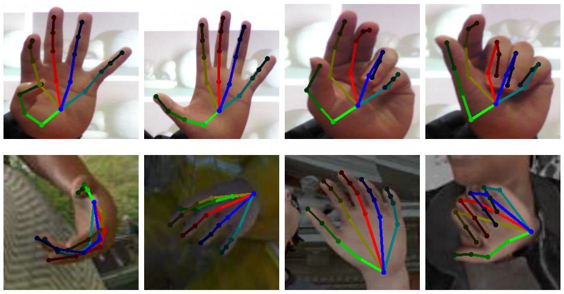

Figure 4: Some examples of the three benchmark hand datasets used for qualitative and quantitative evaluation. Top Row: The STB dataset [50] contains real hand images with 3D keypoints. Bottom Row: The RHP dataset [55] provides synthetic hand image with 3D hand keypoint annotations.

3.3.2 Loss Function.

The joint loss function is a nested sum of various types of loss functions. Specifically:

$$\mathcal{L}_{joint} = \alpha \mathcal{L}_{adv} + \mathcal{L}_{app} + \mathcal{L}_{pose} \tag{7}$$

$$\mathcal{L}_{app} = \tau_1 \mathcal{L}_1 + \tau_2 \mathcal{L}_p \tag{8}$$

$$\mathcal{L}_{pose} = \gamma_1 \mathcal{L}_{x,y} + \gamma_2 \mathcal{L}_z, \tag{9}$$

where $\mathcal{L}_{adv}$ denotes the adversarial loss, $\mathcal{L}_{app}$ measures the appearance difference of the generated hand image and the target hand image, and $\mathcal{L}_{pose}$ is a 3D hand pose estimation task loss. $\alpha, \tau_{1,2}, \gamma_{1,2}$ represent the corresponding weights and they are determined empirically. The adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_{adv} &= \mathbb{E}_{\boldsymbol{I}_{p_s}, \boldsymbol{I}_{p_t}, \boldsymbol{p}_t} \{\log[D_a(\boldsymbol{I}_{p_s}, \boldsymbol{I}_{p_t}) \cdot D_p(\boldsymbol{p}_t, \boldsymbol{I}_{p_t})]\}, \\ &+ \mathbb{E}_{\boldsymbol{I}_{p_s}, \tilde{\boldsymbol{I}}_{p_t}, \boldsymbol{p}_t} \{\log[(1 - D_a(\boldsymbol{I}_{p_s}, \tilde{\boldsymbol{I}}_{p_t})) \cdot (1 - D_p(\boldsymbol{p}_t, \tilde{\boldsymbol{I}}_{p_t}))]\}, \end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{I}}_{p_t} = G(\boldsymbol{I}_{p_s}, \boldsymbol{p}_s, \boldsymbol{p}_t)$. $G(\cdot)$ is the progressive transfer module. $\mathcal{L}_1$ denotes the pixel-wise ℓ_1 loss computed between the generated hand image and the target image, *i.e.* $\mathcal{L}_1 = \left\|\tilde{\boldsymbol{I}}_{p_t} - \boldsymbol{I}_{p_t}\right\|_1$. $\mathcal{L}_p$ is a perceptual loss [15] widely used in style transfer and super resolution defined as:

$$\mathcal{L}_p = \frac{1}{C_i H_i W_i} \left\|\phi_i(\tilde{\boldsymbol{I}}_{p_t}) - \phi_i(\boldsymbol{I}_{p_t})\right\|_2^2, \tag{11}$$

where ϕ_i is the i_{th} layer of a pretrained VGG-16 network. We empirically use the *conv3_3* layer. $\mathcal{L}_{x,y}$ denotes the 2D hand pose estimation loss:

$$\mathcal{L}_{x,y} = \frac{1}{6K} \sum_{s=1}^{6} \sum_{i=1}^{K} \left\|\boldsymbol{H}_i^s - \boldsymbol{H}_i^*\right\|_F^2, \tag{12}$$

where $\{\boldsymbol{H}_i^*\}_{i=1}^K (K = 21)$ is the ground truth 2D poses in heat maps and 6 is the number of stages in HPM. $\mathcal{L}_z$ denotes the depth estimation loss:

$$\mathcal{L}_z = \frac{1}{K} \sum_{i=1}^{K} \begin{cases} \frac{1}{2}(Z_i - Z_i^*)^2, |Z_i - Z_i^*| \leq 1 \\ |Z_i - Z_i^*| - 0.5, otherwise, \end{cases} \tag{13}$$

where $\{\boldsymbol{Z}_i^*\}_{i=1}^K (K = 21)$ is the ground truth relative depth.

4 EXPERIMENTS

4.1 Experimental Settings

Baselines Approaches. We select CycleGAN [52], Pix2pix [13], PG^2 [28], Pose-GAN [39], PATN [53] as the baseline methods to be compared with our proposed *mm-hand*.

The CycleGAN learns unpaired image-to-image translation, by enforcing cycle consistency to push the source domain of 2D pose maps to be consistent with the target domain of realistic hand images. The Pix2pix learns the translation from the 2D hand pose label maps to a real hand image. The PG^2 is a two-stage coarse-to-fine network that generates a person image under target pose from a source person image. The Pose-GAN introduces deformable skip connections in the generator. The PATN generates person images under the guidance of target pose progressively via attention mechanism.

3D Hand Pose Estimators. Unfortunately, several state-of-the-art estimators for further comparison [1, 3, 4, 10, 31, 46, 47, 54] has not released their training code. We adopt two hand pose estimators, Hand3D [54] and 3D-HPM (proposed), to assess the quality of our generated hand training images through the improvement of the model performance. We propose 3D-HPM[3] to evaluate 3D coordinates from a single monocular RGB image. Although not as good as the state-of-the-art estimators, Hand3D and 3D-HPM have achieved EPE of 9.81 and 15.71 on the STB dataset.

Datasets. We select two benchmark datasets for performance evaluation including the *Stereo Tracking Benchmark* (STB) [50], and the *Render Hand Pose* (RHP) [55]. The RHP dataset contains 41,258 training and 2,728 testing hand samples captured from 20 subjects performing 39 actions. Each sample consists of an RGB image, a depth map, and the segmentation masks for the background, person, and each finger. Each hand is annotated with its 21 keypoints in both 2D image coordinates and 3D world coordinate positions. The RHP dataset is split into a validation set (R-val) and a training set (R-train). This dataset is challenging due to the large variations in viewpoints and the low image resolution. The STB dataset provides 18,000 hand images. It is split into two subsets: the stereo subset (STB-BB) and the color-depth subset (STB-SK).

Evaluation Metrics. To evaluate the image realisticity and pose consistency, we propose PCKb and adopt SSIM and IS with their masked version for evaluation.

The PCKb is the percentage of correct keypoints within two-thirds of the average length of the bones in hand. Mask-SSIM and mask-IS are evaluated on the hand image with masked background.

We further adopt two metrics for evaluating the accuracy of estimated hand poses: the average End-Point-Error (EPE), and the Area Under the Curve (AUC) on the Percentage of Correct Keypoints (PCK). The performance metrics are computed in pixels (px) on the hand images and millimeters (mm) in 3D world coordinate, respectively. The performance of 3D hand joint prediction is measured by the PCK curves averaged over all 21 keypoints.

[3] 3D-HPM works by first passing the input through a series of convolution layers grouped in 6 stages. Next, it passes the encoded information through a series of Repeater modules, which in essence, is another series of convolution layers that are regularized by the ground truth 2D coordinates before being passed through a fully-connected layer to predict the z information.

Table 1: Quantitative comparison of the generated hand images using CycleGAN, Pix2pix, PG^2, Pose-GAN, PATN and our proposed MM-Hand on the two benchmark datasets STB and RHP. We adopt the metric of SSIM, IS, mask-SSIM, mask-IS and PCKb to quantitatively evaluate the quality of the generated hand images. IS and Mask-IS are not reliable due to lack of hand related classes in the ImageNet dataset, which the Inception v3 model is pretrained on. Similarly, SSIM's scores can be erratic due to background noise. ↑: higher is better.

	$\mathcal{X}_{STB}$					$\mathcal{X}_{RHP}$				
	SSIM ↑	IS ↑	mask-SSIM ↑	mask-IS ↑	PCKb ↑	SSIM ↑	IS ↑	mask-SSIM ↑	mask-IS ↑	PCKb ↑
CycleGAN	0.002	1.52	0.611	2.49	0.07	0.008	2.08	0.816	**2.98**	0.015
Pix2pix	0.027	2.24	0.625	**2.632**	0.527	0.010	2.67	0.816	2.85	0.119
PG^2	0.026	2.33	0.638	2.224	0.686	0.021	2.236	0.822	2.762	0.250
Pose-GAN	0.02	1.01	0.610	1.495	0.05	0.014	1.03	0.808	2.012	0.014
PATN	0.014	**2.371**	0.656	2.276	0.564	0.054	2.348	0.830	2.532	0.248
Ours (MM-Hand)	**0.115**	2.187	**0.677**	2.53	**0.688**	**0.078**	**2.376**	**0.844**	2.747	**0.619**

Table 2: The 3D hand pose estimation performance using $\mathcal{M}_{Hand3D}$ and $\mathcal{M}_{3D-HPM}$on $\mathcal{X}_{STB}$ and $\mathcal{X}_{RHP}$, augmented by images generated by different methods under different portion α of $\mathcal{X}$ as the reduced training set. *None* means no generative model is employed, thus not using data augmentation. We adopt the metric of mean EPE to evaluate the estimation performance. Lower EPE indicates better performance.

		$\mathcal{X}_{STB}$					$\mathcal{X}_{RHP}$				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
$\mathcal{M}_{Hand3D}$	None	60.28	48.84	30.23	18.74	9.81	90.12	60.27	36.58	25.29	20.52
	CycleGAN	80.27	82.57	75.39	72.56	9.81	90.29	78.29	60.29	40.29	20.52
	Pix2pix	72.57	71.27	54.13	50.25	9.81	82.39	76.48	62.16	80.29	20.52
	PG^2	74.27	68.22	58.23	52.28	9.81	85.29	72.83	64.49	40.25	20.52
	PATN	70.28	68.23	50.37	40.57	9.81	84.25	74.49	84.35	60.25	20.52
	Pose-GAN	72.58	69.27	52.85	39.56	9.81	94.59	84.38	67.83	45.59	20.52
	Ours (MM-Hand)	**52.39**	**32.37**	**27.49**	**16.48**	9.81	**80.29**	**54.38**	**28.49**	**24.38**	20.52
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
$\mathcal{M}_{3D-HPM}$	None	64.16	48.91	33.00	35.51	15.71	**52.21**	50.38	47.36	45.43	35.86
	CycleGAN	111.75	54.55	51.71	47.02	15.71	66.63	59.63	57.59	61.67	35.86
	Pix2pix	99.59	46.71	47.83	46.91	15.71	65.73	64.56	62.31	55.07	35.86
	PG^2	91.03	47.00	47.20	46.78	15.71	61.05	58.95	57.59	56.72	35.86
	PATN	99.66	46.95	47.84	40.18	15.71	56.11	50.26	50.64	51.92	35.86
	Pose-GAN	102.70	46.65	48.03	47.02	15.71	60.54	57.44	53.40	52.83	35.86
	Ours (MM-Hand)	**41.79**	**20.24**	**16.79**	**16.15**	15.71	52.47	**42.22**	**41.63**	**40.49**	35.86

4.2 Experimental Results

4.2.1 Qualitative and Quantitative Comparison of the Synthesized Hand Images. The source domain is 2D hand pose maps derived from 3D hand pose by projection, and the target domain is realistic hand images. The source domain is represented by 2D hand pose label maps. Some examples are shown in the first column in Figure 5. Figure 5 shows the visual quality comparison of the hand images generated by CycleGAN, Pix2pix, PG^2, Pose-GAN, PATN, and our proposed *mm-hand*, respectively. Our MM-Hand beats the baseline approaches on almost all the metrics on both STB and RHP.

4.2.2 Boosting the Performance of 3D Hand Pose Estimators when Training Data is Reduced. The hand image generation method can generate realistic hand images for improving the 3D hand pose estimator learning, especially when the original training data is insufficient. In the training set $\mathcal{X}_{tr}$ of the 3D hand pose dataset, we randomly select a portion α of $\mathcal{X}_{tr}$ as the reduced training set, denoted as $\mathcal{X}^R_\alpha$. The rest of the data is denoted as $\mathcal{X}^R_{1-\alpha}$ (*i.e.* , $\mathcal{X}^R_{1-\alpha} \cup \mathcal{X}^R_\alpha = \mathcal{X}_{tr}$).

Using the poses ($\boldsymbol{p}_t$s) in $\mathcal{X}^R_{1-\alpha}$ and the images with poses ($\boldsymbol{I}_{p_s}$s and $\boldsymbol{p}_s$s) in $\mathcal{X}^R_\alpha$, we build $\overline{\mathcal{X}}^R_{1-\alpha}$ by replacing each image $\boldsymbol{I}_{p_t}$ in $\mathcal{X}^R_{1-\alpha}$ with $\tilde{\boldsymbol{I}}_{p_t}$. Given $\boldsymbol{p}_t$, $\tilde{\boldsymbol{I}}_{p_t}$ is generated from the best matched source hand image $\boldsymbol{I}_{p_s}$ in $\mathcal{X}^R_\alpha$, whose pose $\boldsymbol{p}_s$ is closest to $\boldsymbol{p}_t$ in pose distance. Specifically,

$$\forall(\boldsymbol{I}_{p_t}, \boldsymbol{p}_t) \in \mathcal{X}^R_{1-\alpha}, \exists(\boldsymbol{I}_{p^*_s}, \boldsymbol{p}^*_s) \in \mathcal{X}^R_\alpha, \text{ s.t. } \boldsymbol{p}^*_s = \text{argmin}_{\boldsymbol{p}_s} d(\boldsymbol{p}_s, \boldsymbol{p}_t), \\ < \boldsymbol{I}_{p_s}, \boldsymbol{p}^*_s, \boldsymbol{p}_t > \longrightarrow \tilde{\boldsymbol{I}}_{p_t}. \tag{14}$$

Consequently, an augmented training set $\mathcal{X}^A_\alpha$ is formed from $\mathcal{X}^R_\alpha$ and $\overline{\mathcal{X}}^R_{1-\alpha}$ (*i.e.* , $\mathcal{X}^R_\alpha \cup \overline{\mathcal{X}}^R_{1-\alpha} = \mathcal{X}^A_\alpha$). We use $\mathcal{X}^A_\alpha$ instead of $\mathcal{X}_{tr}$ to train the two 3D hand pose estimators (Hand3D and 3D-HPM) respectively. All the numbers in Table 2 are obtained by evaluation on the testing set of $\mathcal{X}_{te}$.

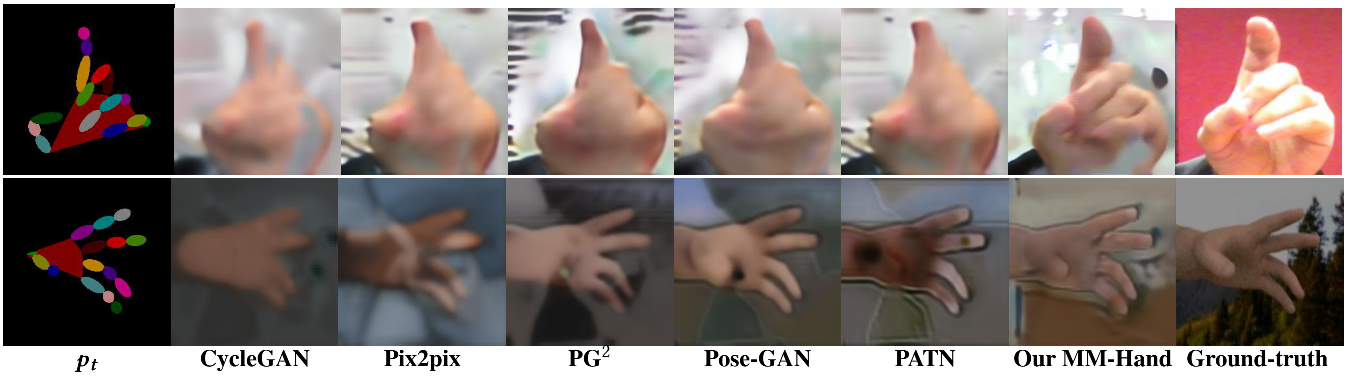

Figure 5: Qualitative comparison of the synthesized hand images using CycleGAN, Pix2pix, PG^2, Pose-GAN, PATN and our proposed MM-Hand on the two benchmark datasets STB and RHP. From top to bottom: The STB dataset; The RHP dataset.

Table 3: Ablation study of MM-Hand using 3D pose estimation on the RHD and STB datasets. AUC 20-50 is the area under the PCK curve between 20mm and 50 mm. ↑: higher is better.

	AUC 20-50 ↑
RHD Dataset	
a) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$	0.894
b) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $\boldsymbol{M}_n$	0.905
c) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + $\boldsymbol{M}_n$	0.909
d) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + $\boldsymbol{M}_n$ + GCT	0.915
e) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + $\boldsymbol{M}_n$ + GCT + INNM	0.929
STB Dataset	
a) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$	0.981
b) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $\boldsymbol{M}_n$	0.988
c) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + $\boldsymbol{M}_n$	0.994
d) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + $\boldsymbol{M}_n$ + GCT	0.997
e) base + $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + $\boldsymbol{M}_n$ + GCT + INNM	0.999

Inspired by Zhu *et al.* [53], this experiment is designed to assess the quality of our generated hand images, which would boost the performance of 3D hand pose estimation, as a data augmentation task.

Table 2 shows that augmenting the training data of STB and RHD using hand images generated by our *mm-hand* can achieve performance gain across different values of α. Moreover, the performance gain is more significant if the performance gap is relatively large.

We further compare our approach to PG^2, Pose-GAN, and PATN under the same experimental setting for 3D hand pose estimation. Table 2 shows that our approach achieves consistent improvements over all previous works for different αs, suggesting the proposed method can generate more realistic and pose-preserving hand images for the 3D hand pose estimation task.

4.2.3 Ablation Study of MM-Hand. To verify the effectiveness of the proposed 3D pose embeddings and the geometry-based curriculum learning, we incrementally evaluate them on the STB dataset. We choose PATN as the baseline method. Five variants are then constructed on top of the baseline:

a) ResNetGenerator (base) + contour map $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$;

b) ResNetGenerator (base) + contour map $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + attention mask $(\boldsymbol{M}_n)$;

c) ResNetGenerator (base) + contour map $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + depth map $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + attention mask $(\boldsymbol{M}_n)$;

d) ResNetGenerator (base) + contour map $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + depth map $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + attention mask $(\boldsymbol{M}_n)$ + geometry-based curriculum training (GCT);

e) ResNetGenerator (base) + contour map $(\boldsymbol{c}_{p_s}, \boldsymbol{c}_{p_t})$ + depth map $(\boldsymbol{d}_{p_s}, \boldsymbol{d}_{p_t})$ + attention mask $(\boldsymbol{M}_n)$ + geometry-based curriculum training (GCT) + inference with nearest neighbor match (INNM).

Table 3 presents the ablation study results, from which we can draw a conclusion that incrementally adding pose embedding (contour map + depth map), attention mask, geometry-based curriculum training, and inference nearest neighbor match consistently improve the 3D hand pose estimation performance.

4.2.4 Surpassing State-of-the-Art 3D Hand Pose Estimators under Standard Setting. We compare with the current state-of-the-art 3D hand pose estimators, including Z&B [54], Cai *et al.* [4], Mueller *et al.* [31], Yang *et al.* [47], Baek *et al.* [1], Ge *et al.* [10], Boukhayma *et al.* [3], Yang *et al.* [46]. We report our results by training the Hand3D model with additional hand images generated by our MM-Hand for augmentation. Unfortunately, all the state-of-the-art methods have not (or partially) released their code at the time of writing. Hence, we select Hand3D in our experiment, which provides both training and evaluation code. Figure 6 and Table 4 present that training Hand3D with additional hand images generated by MM-Hand achieves the best performance on STB and RHP in AUC of the PCK curve. We further show some examples of our generated hand images in Figure 7. These images are generated by MM-Hand trained on the STB dataset. The target poses $\boldsymbol{p_t}$s are generated using an augmented reality (AR) simulator, *i.e.* Blender[4].

5 CONCLUSION

Due to the inherent depth ambiguity, building a sizeable real-world hand dataset with accurate 3D annotations is one major challenge of 3D hand pose estimation. We propose a 3D-aware multi-model

[4] https://www.blender.org

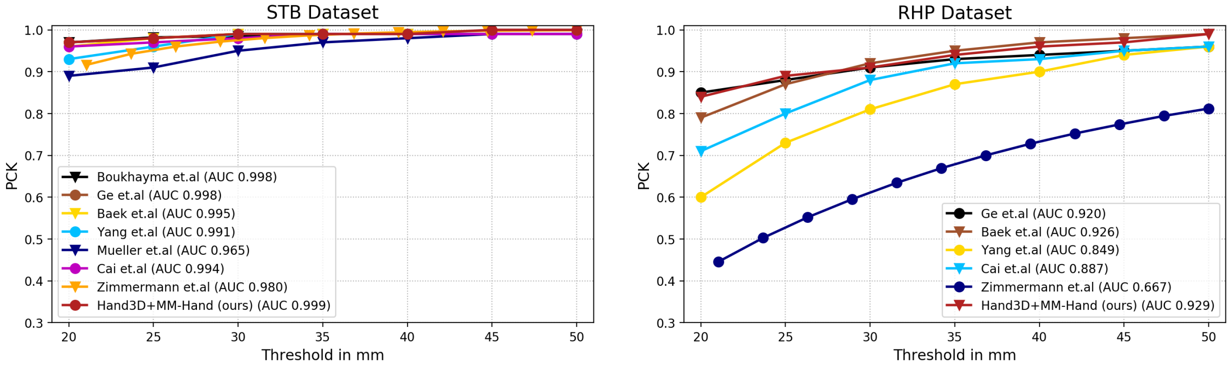

Figure 6: Our comparisons with state-of-the-art approaches on (a) the STB dataset in 3D hand pose estimation task; (b) the RHP dataset in 3D hand pose estimation task. We choose all the state-of-the-art approaches proposed after 2017.

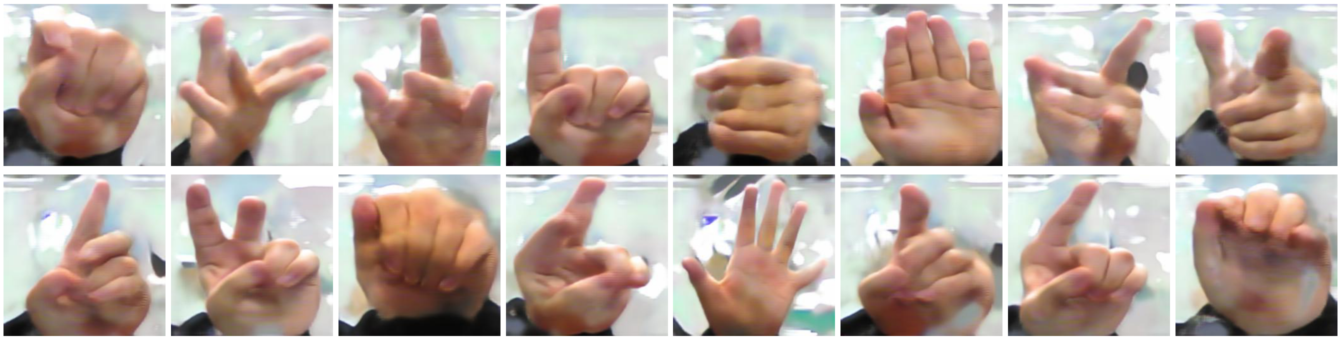

Figure 7: We randomly pick 16 hand images generated by MM-Hand trained on STB. Despite some missing texture details, the generated hand images look realistic and are consistent with the target poses p_ss.

Table 4: 3D pose estimation results on the RHD and STB datasets. AUC 20-50 is the area under the PCK curve between 20mm and 50 mm. ↑: higher is better.

	AUC 20-50 ↑
RHD Dataset	
Z&B [55]	0.667
Yang *et al.* [47]	0.849
Cai *et al.* [4]	0.887
Ge *et al.* [10]	0.92
Baek *et al.* [1]	0.926
Hand3D + MM-Hand (Ours)	**0.929**
STB Dataset	
Z&B [55]	0.980
Yang *et al.* [47]	0.991
Cai *et al.* [4]	0.994
Baek *et al.* [1]	0.995
Ge *et al.* [10]	0.998
Hand3D + MM-Hand (Ours)	**0.999**

guided hand generative network (*MM-Hand*), and a novel geometry-based training and inference strategy to generate hand images under the guidance of 3D hand poses. With the help of an external dataset with paired depth maps and 3D hand poses, we train the depth map generator to synthesize depth maps based on any given 3D poses. Our proposed *mm-hand* can generate realistic, diverse, and pose preserving hand images based on any given 3D poses and synthetic depth maps. Qualitative results show that the hand images generated by *mm-hand* resemble the ground truth hand images in both appearance and pose. We further evaluate the quality of our generated hand images on two benchmark datasets. Our experimental results show that training 3D hand pose estimator with our augmented data can outperform the state-of-art methods on both STB and RHP datasets.

ACKNOWLEDGMENTS

The work was mainly done during Z. Wu and D. Hoang's internships at Tencent America. The research of Y.-Y. Lin was supported by Ministry of Science and Technology under grants 107-2628-E-009-007-MY3, 109-2634-F-007-013, and 109-2221-E-009-113-MY3. The research of Z. Wang was partially supported by US Army Research Office Young Investigator Award W911NF2010240.

REFERENCES

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. 2019. Pushing the Envelope for RGB-based Dense 3D Hand Pose Estimation via Neural Rendering. In *CVPR*.

[2] Manas Kamal Bhuyan, Mithun Kumar Kar, and Debanga Raj Neog. 2011. Hand pose identification from monocular image for sign language recognition. In *ICSIPA*.

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 2019. 3d hand shape and pose from images in the wild. In *CVPR*.

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. 2018. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*.

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. 2019. Everybody Dance Now. In *ICCV*.

[6] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Wei Fan, and Xiaohui Xie. 2020. DGGAN:Depth-image Guided Generative Adversarial Networks for Disentangling RGB and Depth Images for 3D hand Pose Estimation. In *WACV*.

[7] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Hui Tang, Yufan Xue, Yen-Yu Lin, Xiaohui Xie, and Wei Fan. 2019. TAGAN: Tonality-Alignment Generative Adversarial Networks for Realistic Hand Pose Synthesis. In *BMVC*.

[8] Wuyang Chen, Zhiding Yu, Zhangyang Wang, and Anima Anandkumar. 2020. Automated Synthetic-to-Real Generalization. *arXiv preprint arXiv:2007.06965* (2020).

[9] Patrick Esser, Ekaterina Sutter, and Björn Ommer. 2018. A variational u-net for conditional appearance and shape generation. In *CVPR*.

[10] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 2019. 3D Hand Shape and Pose Estimation from a Single RGB Image. In *CVPR*.

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *ICML*.

[12] Yi-ping Hung and Shih-Yao Lin. 2016. Re-anchorable virtual panel in three-dimensional space. US Patent 9,529,446.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

[14] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2019. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972* (2019).

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv* (2013).

[17] Philip Krejov. 2016. *Real time hand pose estimation for human computer interaction*. Ph.D. Dissertation. University of Surrey (United Kingdom).

[18] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*.

[19] Chongxuan Li, Kun Xu, Jiashuo Liu, Jun Zhu, and Bo Zhang. 2019. Triple generative adversarial networks. *arXiv preprint arXiv:1912.09784* (2019).

[20] Rui Li, Zhenyu Liu, and Jianrong Tan. 2019. A survey on 3D hand pose estimation: Cameras, methods, and datasets. *Pattern Recognition* (2019).

[21] Yining Li, Chen Huang, and Chen Change Loy. 2019. Dense intrinsic appearance flow for human pose transfer. In *CVPR*.

[22] Shih-Yao Lin, Yun-Chien Lai, Li-Wei Chan, and Yi-Ping Hung. 2010. Real-time 3D model-based gesture tracking for multimedia control. In *ICPR*.

[23] Shih-Yao Lin, Yen-Yu Lin, Chu-Song Chen, and Yi-Ping Hung. 2017. Learning and inferring human actions with temporal pyramid features based on conditional random fields. In *ICASSP*.

[24] Shih-Yao Lin, Chuen-Kai Shie, Shen-Chi Chen, and Yi-Ping Hung. 2012. Action recognition for human-marionette interaction. In *MM*.

[25] Shih-Yao Lin, Chuen-Kai Shie, Shen-Chi Chen, and Yi-Ping Hung. 2013. AirTouch panel: a re-anchorable virtual touch panel. In *MM*.

[26] Wanhong Lin, Lear Du, Carisa Harris-Adamson, Alan Barr, and David Rempel. 2017. Design of hand gestures for manipulating objects in virtual reality. In *ICHCI*.

[27] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. *arXiv* (2019).

[28] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *NeurIPS*.

[29] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled person image generation. In *CVPR*.

[30] Asanterabi Kighoma Malima, Erol Özgür, and Müjdat Çetin. 2006. A fast algorithm for vision-based hand gesture recognition for robot control. In *ICPCA*.

[31] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*.

[32] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2017. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*.

[33] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. 2018. Image to image translation for domain adaptation. In *CVPR*.

[34] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. 2018. Dense pose transfer. In *ECCV*.

[35] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *CVPR*.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

[37] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *CVPR*.

[38] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Multistage adversarial losses for pose-based human image synthesis. In *CVPR*.

[39] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *CVPR*.

[40] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. 2019. Unsupervised Person Image Generation with Semantic Parsing Transformation. In *CVPR*.

[41] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. 2015. Cascaded hand pose regression. In *CVPR*.

[42] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. 2014. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR*.

[43] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-time continuous pose recovery of human hands using convolutional networks. *TOG* 33, 5 (2014), 169.

[44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.

[45] Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks* 125 (2020), 41–55.

[46] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. 2019. Aligning Latent Spaces for 3D Hand Pose Estimation. In *ICCV*.

[47] Linlin Yang and Angela Yao. 2019. Disentangling Latent Hands for Image Synthesis and Pose Estimation. In *CVPR*.

[48] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. 2020. Deep Plastic Surgery: Robust and Controllable Image Editing with Human-Drawn Sketches. *arXiv preprint arXiv:2001.02890* (2020).

[49] Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. 2019. Controllable artistic text style transfer via shape-matching gan. In *CVPR*.

[50] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 2016. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214* (2016).

[51] Xiaofeng Zhang, Zhangyang Wang, Dong Liu, and Qing Ling. 2019. Dada: Deep adversarial data augmentation for extremely low data regime classification. In *ICASSP*.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

[53] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. 2019. Progressive Pose Attention Transfer for Person Image Generation. In *CVPR*.

[54] Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3d hand pose from single rgb images. In *ICCV*.

[55] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *CVPR*.

[56] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. 2019. FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images. *arXiv* (2019).