TAGAN: Tonality-Alignment Generative Adversarial Networks for Realistic Hand Pose Synthesis

Liangjian Chen[1]
liangjc2@ics.uci.edu
Shih-Yao Lin[2]
shihyaolin@tencent.com
Yusheng Xie[2]
yushengxie@tencent.com
Hui Tang[2]
tanghui@tencent.com
Yufan Xue[2]
robinxue@tencent.com
Yen-Yu Lin[3]
yylin@citi.sinica.edu.tw
Xiaohui Xie[1]
xhx@ics.uci.edu
Wei Fan[2]
davidwfan@tencent.com

[1] University of California, Irvine
[2] Tencent Medical AI Lab
[3] Academia Sinica

Abstract

3D hand pose estimation from a single RGB image is important but challenging due to the lack of sufficiently large hand pose datasets with accurate 3D hand keypoint annotations for training. In this work, we present an effective method for generating realistic hand poses, and show that existing algorithms for hand pose estimation can be greatly improved by augmenting training data with the generated hand poses, which come naturally with ground-truth annotations. Specifically, we adopt an augmented reality simulator to synthesize hand poses with accurate 3D hand-keypoint annotations. These synthesized hand poses look unnatural and are not adequate for training. To produce more realistic hand poses, we propose to blend each synthetic hand pose with a real background and develop *tonality-alignment generative adversarial networks* (TAGAN), which align the tonality and color distributions between synthetic hand poses and real backgrounds, and can generate high-quality hand poses. TAGAN is evaluated on the *RHP*, *STB*, and *CMU-PS* hand pose datasets. With the aid of the synthesized poses, our method performs favorably against the state-of-the-arts in both 2D and 3D hand pose estimation.

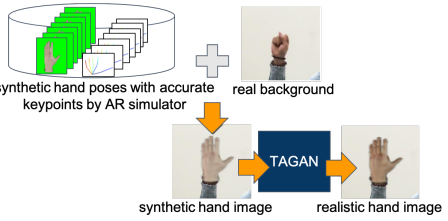


Figure 1: Overview of our method for realistic hand image synthesis. We blend a synthetic pose by an AR simulator with real background to yield a synthetic hand image, which is then fed to the proposed TAGAN to produce a more realistic hand image.

1 Introduction

Estimating hand poses from monocular RGB images has drawn increasing attention because it is essential to many applications such as virtual and augmented reality [17], and human computer interaction [25]. It has gained significant progress [6, 17] owing to the fast development of *deep neural networks* (DNN). These DNN-based methods learn hand representations and estimate poses jointly. Despite effectiveness, DNN-based methods highly rely on a vast amount of training data. However, it is expensive to collect all hand poses of interest with manual hand-keypoint annotations for training.

Synthesizing training data has been a feasible way to tackle the lack of training data. Recent studies, *e.g.* [17, 39], have adopted *augmented reality* (AR) simulators to generate large-scale training examples. In this way, plenty hand images with various poses, skin textures, and lighting conditions can be systematically synthesized. Moreover, accurate hand-keypoint annotations of these synthesized hand images are also available. Training with such synthetic images may not result in a much improved hand pose estimator because of the dissimilarity between the real and synthetic data. In this work, we suggest blending a synthetic hand pose (foreground) image with a real background image so that the blended images are realistic enough to serve as high-quality training data.

We are aware of the dissimilarity between synthetic hand pose images and real background images in styles and appearances. Thus, we present a GAN-based method, *tonality-alignment generative adversarial networks* (TAGAN), to eliminate the dissimilarity. TAGAN employs the image-to-image translation technique based on *conditional GAN* (CGAN) [12], where extra shape features serve as the input to GAN and constrain the object shape in the synthesized photo. In addition to the shape constraint, a tonality-alignment loss in TAGAN is designed to align the color distributions tonality of the input and generated images. It turns out that the hand pose images can be better blended into the background images, resulting in more realistic hand pose images. The hand pose estimator is then considerably improved by using the generated hand pose images as the augmented training data.

Figure 1 gives the overview of the proposed method. The main contribution of this work is three-fold: First, we propose to fuse synthetic hand poses and real background images so that the resulting synthesized hand images can be more realistic. Second, we present TAGAN which performs conditional adversarial learning and seamlessly blends synthetic hand poses into real backgrounds. Third, we demonstrate that existing pose estimators trained with the generated hand pose data gain significant improvements over the current state-of-the-arts on both 2D and 3D datasets.

2 Related Work

Data Augmentation via Simulator. Recent work, *e.g.* [14], for hand pose estimation has trained the models on synthetic training data. In [39], a synthetic hand pose dataset is generated by an open source simulator, and serves as augmented training data to improve pose estimator learning. However, the synthetic hand images produced by the AR simulator look artificial, leading to limited performance gains. To address this issue, recent work, *e.g.* [17, 23], leverages adversarial learning to enhance the quality of synthetic hand images.

Data Augmentation via Adversarial Learning. Generating realistic images by using *generative adversarial networks* (GAN) [8, 23] has been a research trend. Isola *et al.* propose *Pix2Pix Net* [13] to learn a mapping from a sketch to a realistic image, *e.g.* transferring a car sketch to a car image. Unlike GAN requiring paired training data, *CycleGAN* [38] employs cycle-consistent adversarial networks for translating images from a source domain to a target domain with unpaired examples. To increase the amount of training data, Shrivastava *et al.* present *SimGAN* [23], which employs simulated and unsupervised learning to improve the realism of the output of a simulator with unlabeled real data. However, the simulator's data include only objects, ignoring background scenes. Thus, the resulting synthetic images are filled objects, but the background information is usually crucial in practice. In this work, we explore techniques that directly regularize the foreground (hand) and the background (natural scenes where the hand appears [11, 36]).

Vision-based Hand Pose Estimation. Hand pose estimation has drawn increasing attention for decades [1, 2, 3, 5, 6, 7, 10, 16, 20, 26, 28, 29, 32, 33, 34, 37]. Research efforts can be categorized by their input data forms, which primarily include 2D RGB images and 3D RGBD images with depth information. Recent progress has tried to estimate the 3D hand pose from a monocular RGB image. For example, Oikonomidis *et al.* [18, 19] propose a hand tracking approach based on *particle swarm optimization*. Simon *et al.* [24] adopt multiview bootstrapping to calculate hand keypoints from RGB images. Zimmermann and Brox [39] propose a 3D pose regression net, enabling 3D hand pose estimation from an RGB image.

Domain Adaption via Adversarial Learning. Domain adaption has been introduced by Saenko *et al.* [22] for pairwise metric transforming and can be developed by the study of visual dataset bias [27]. For domain adaption, models are often designed to capture the invariant patterns between two different data distributions so that they can perform well cross domains. Hoffman *et al.* [9] adapt the CycleGAN loss with the task loss to further improve the results of image translation.

3 Methodology

This section describes our approach to hand pose image generation. We first explain how GAN and conditional GAN are applied to this problem, then depict synthetic hand image generation, and finally specify how our approach works to improve the synthetic images.

3.1 GAN and Conditional GAN

GAN learns a mapping from a random noise vector z to its generated image y, *i.e.*, $G : z \rightarrow y$, where G is the generator. The *conditional GAN* (CGAN) [15] is an extension of GAN. The

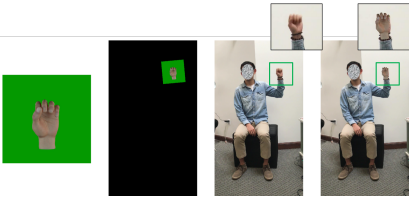


Figure 2: Hand image synthesis. From left to right: 1) a hand image $\mathbf{u}$ in our synthetic dataset, 2) its representation with an affine transformation $g(\mathbf{u})$, 3) target image (enclosed by the green box) $\mathbf{v}$ via hand detection, and 4) the synthesized hand image obtained by applying TAGAN to blend the synthetic hand pose into the real background image.

inputs to CGAN can be augmented with additional conditions so that CGAN can leverage the additional conditions to constrain the output image y. The conditions can be specified in the form of extra inputs to both the generator network and the discriminator network.

CGAN-based methods can be applied to image-to-image translation. In the representative work *pix2pix net* [12], the condition is used to make the object shape in the output image similar to the additional input shape map x_s, which is pre-computed by applying the edge detector HED [31] to the image x. As the additional input, x_s is fed to both the generator and the discriminator. In this way, the condition x_s and the latent space representation z are transformed into a joint hidden representation. CGAN learns a mapping from x_s (inferred from x) and z to the generated output y, G: $\{x,z\} \rightarrow y$. The CGAN objective can be formulated as

$$\mathcal{L}_{CGAN}(G,D) = \mathbb{E}_{x_s,y}[\log D(x_s,y)] + \mathbb{E}_{x_s,z}[\log(1 - D(x_s,G(x_s,z)))], \tag{1}$$

where the discriminator D aims to distinguish the data generated by G from real data. Yet, the generator G generates the data to not only fool D but also fulfill the input condition. In Eq. (1), the generator G minimizes the differences between the real images and the generated images while G's adversary, D, tries to learn a discriminating function to maximize such differences. In addition, the shape of output y is constrained by x_s. Thus, G is optimized via

$$G^* = \arg\min_G \max_D \mathcal{L}_{CGAN}(G,D) + \lambda \cdot \mathcal{L}_S(G), \tag{2}$$

where $\mathcal{L}_s$ and λ are the shape loss and its weight, respectively. The shape loss can be calculated by using L_1 distance to reduce the unfavorable effect of blurring, and is defined by

$$\mathcal{L}_S(G) = \mathbb{E}_{x_s,y,z}[\|y - G(x_s,z)\|_1]. \tag{3}$$

Although the existing work, Pix2pix Net is able to contain the shape of the generated object by using the input shape map, it does not take color consistency between the object and background into account. Hence, its generated images might look unnatural.

3.2 Synthetic Hand Image Generation

Existing hand pose datasets are not large enough to stably learn a deep network for hand pose estimation. Moreover, manual hand-keypoint annotation is expensive and labor-intensive. Also, the annotated hand-keypoints are still error-prone and often not accurate enough. To address the quantitative and qualitative issues of hand pose training data, we firstly adopt an open-source AR simulator to produce large-scale *synthetic hand pose images* with *accurate*

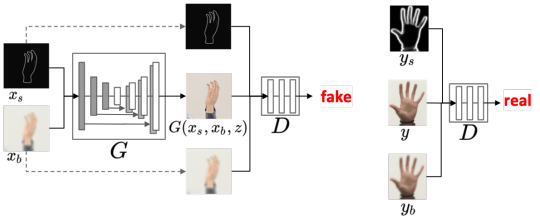


Figure 3: TAGAN derives a mapping from the shape map x_s and the color map x_b to the generated image $G(x_s, x_b, z)$. The generator G learns to produce realistic images to fool the discriminator D by blending a synthetic hand pose with a real background image, while the discriminator D aims to separate the fake (synthetic) images from the real images.

2*D/3D hand-keypoint labels*, which cover common and feasible hand poses by observing a group of subjects for a period of time. However, these synthetic hands look unnatural and cannot serve as training data.

We also collect a large-scale, daily-life, and unlabeled hand gesture videos performed by some subjects. Each image in these videos consists of *real hand(s) and background*. See the third image in Figure 2 as an example. We detect the hand region in the image using the pose estimation toolkit, OpenPose Library [23]. However, the estimated poses are not good enough to serve as the training data. Thus, we propose to *match* the hand pose images in the AR and real datasets. In this way, the accurate keypoints in the AR dataset and real backgrounds in the real dataset can complement each other, and the proposed TAGAN can produce more realistic hand poses with accurate keypoints.

Given a hand image with the estimated pose $\mathbf{v}$ in the real dataset, our goal is to synthesize a hand image with its pose consistent with $\mathbf{v}$. To find the best match, we use $\mathbf{v}$ as a query to the AR dataset generated by the AR simulator, which covers millions of hand poses with accurate keypoint annotations. For each candidate hand pose $\mathbf{u}$ in the AR dataset, its similarity to the target pose $\mathbf{v}$ is defined as

$$K(\mathbf{u},\mathbf{v}) = \langle f \circ g(\mathbf{u}), f(\mathbf{v}) \rangle / (\|f \circ g(\mathbf{u})\| \|f(\mathbf{v})\|), \tag{4}$$

where function g is the affine transformation with which the transformed candidate pose $g(\mathbf{u})$ can best match $\mathbf{v}$, and function f is the feature representation of a pose. In this work, each hand pose is expressed as the concatenation vector of its 21 2D keypoints, *e.g.* $\mathbf{u} = [u_{x1}, u_{y1}, u_{x2}, u_{y2}, \ldots, u_{x21}, u_{y21}] \in \mathbb{R}^{42}$. The feature representation f of a hand pose is the ordered collection of pair-wise keypoint differences along the x and y axes, *i.e.,*

$$f(\mathbf{u}) = [\ldots, u_{xi} - u_{xj}, u_{yi} - u_{yj}, \ldots], \text{ for } 1 \le i < j \le 21. \tag{5}$$

The candidate pose $\mathbf{u}^* = \arg\max_{\mathbf{u}} K(\mathbf{u},\mathbf{v})$ is selected from the dataset yielded by the AR simulator. We superpose pose $\mathbf{u}^*$ over the scene covering $\mathbf{v}$, and apply the proposed TAGAN to better blend the selected pose into the background scene to produce a more realistic hand image. Figure 2 summarizes the process. The proposed TAGAN is elaborated below.

3.3 Tonality-Alignment GAN

Although data augmentation using AR simulators can relieve the lack of training data, the background of the synthesized images is artificial. The background tonality and color distributions between the synthetic and real hand poses are inconsistent. These issues make the

synthetic hand poses less qualified as training data. Inspired by the pix2pix net [13] that leverages the shape map to constrain the output image, we propose *tonality-alignment* GAN (TAGAN) to take the color distribution and shape features into account.

Given a superposed image x, we utilize its blurred counterpart x_b and shape map x_s as the color and shape reference, respectively. The blurred counterpart in our system is derived by applying an average filter to x, while the shape map x_s is obtained by using the HED detector. For the real image y, we adopt the same scheme to extract the shape map y_s and color maps y_b. In TAGAN, the shape map x_s and color map x_b are fed to both the generator and the discriminator as additional input layers such that the x_s, x_b and the output $G(x_s, x_b, z)$ are transformed into a joint hidden representation. Figure 3 illustrates the proposed TAGAN.

During training, the TAGAN learns a mapping from x_s, x_b and a random vector z to the generated output y, *i.e.,* G:$\{x_s, x_b, z\} \rightarrow y$. The objective of TAGAN is designed as

$$\mathcal{L}_{TAGAN}(G,D) = \mathbb{E}_{y_s,y_b,y}[\log D(y_s,y_b,y)] + \mathbb{E}_{x_s,x_b,z}[\log(1 - D(x_s,x_b,G(x_s,x_b,z)]. \quad (6)$$

The generator G in TAGAN is optimized via

$$G^* = \arg\min_G \max_D \mathcal{L}_{TAGAN}(G,D) + \mathcal{L}_{TA}(G,x_s,x_b), \quad (7)$$

where $\mathcal{L}_{TA}$ is the loss function for enforcing the shape similarity between x_s and y as well as the color consistency between x_b and y. The loss is defined by

$$\mathcal{L}_{TA}(G) = \mathbb{E}_{x_s,x_b,y,z}[\lambda_1 \cdot D_c(x_b,x_s,z,y) + \lambda_2 \cdot D_s(x_b,x_s,z,y)], \quad (8)$$

where $D_c(\cdot)$ and $D_s(\cdot)$ denote the color and shape distance functions, respectively. Constants λ_1 and λ_2 are the weights. The shape distance function D_s is expressed as

$$D_s(x_b,x_s,z,y) = \|y - G(x_s,x_b,z)\|_1. \quad (9)$$

In addition to the shape condition, we design a tonality-alignment loss to align the color distributions of the input and the generated images via defining D_c as

$$D_c(x_b,x_s,z,y) = -\sum_i h_g(i) \log\left(\frac{h_y(i)}{h_g(i)}\right), \quad (10)$$

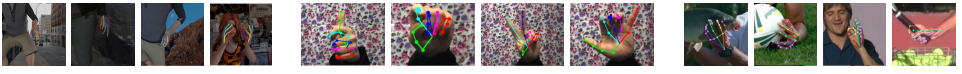
where h_y and h_g are the color histograms of y and $G(x_b,x_s,z)$, respectively. Thereby, $D_c(\cdot)$ in Eq. (10) is the Kullback-Leibler divergence between the two histograms. To train our TAGAN, we collect a real unlabeled hand dataset, called *RHTD*. The hand images in the RHTD are utilized to be real examples y. Some examples in RHTD are shown in Figure 4.

4 Experimental Setting

Hand Pose Estimators. Two existing hand pose estimators are used to assess the quality of the synthesized hand pose images in our experiments: *Hand3D* [39] and *convolutional pose machine* (CPM) [30]. The two estimators infer the 3D and 2D hand poses from a monocular RGB image, respectively. Both estimators are popular and often serve as the baselines for advanced estimators, such as [4, 17, 21, 24, 34]. Thus, if the synthetic hand pose images we generate can improve Hand3D and CPM, those images can also facilitate the follow-up research of Hand3D and CPM.

(a) (b) (c)

Figure 4: Four examples of the RHTD dataset upon which TAGAN is learned. Each example is composed of (a) an image x, (b) its color map x_b and (c) shape maps x_s.



(a) (b) (c)

Figure 5: Some examples in the three benchmark datasets. (a) The RHP dataset provides synthetic hand images with 3D hand keypoints. (b) The STB dataset contains real hand images with 3D keypoints. (c) The CMU-PS dataset offers real hand images with 2D keypoints.

Dataset for Training. To train the generators, including CycleGAN, pix2pix net, and the proposed TAGAN, for hand image synthesis, we collect 17,040 real unlabeled hand images captured from people performing various hand gestures. This dataset is called the *real hand training dataset* (RHTD), which contains hand images with various poses, perspective views, and lighting conditions. Some examples of RHTD are shown in Figure 4. To train the proposed TAGAN, images in RHTD come with the pre-computed edge [31] and color maps. In addition to RHTD, we adopt the AR simulator to generate 60,000 synthetic hand images with various poses, perspectives, and lighting conditions. Some synthetic hand image examples are displayed in the first column of Figure 6. By using our synthetic hand image generation process shown in Figure 2, the synthetic hand images are then present at the appropriate locations of the real images (background). The TAGAN is then applied to blend the synthetic hands with the real background images.

Datasets for Evaluation. To evaluate the quality and the efficacy of the synthesized data, we select three benchmark datasets for evaluation including the *Rendered Hand Pose* (RHP) [39], *Stereo Tracking Benchmark* (STB) [35], and *CMU Panoptic Studio* (CMU-PS) [24] datasets. Figure 5 displays some examples of the three datasets. 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 coordinates and 3D world coordinate positions. The RHP dataset is split into a validation set (R-val) and a training set (R-train). 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). The CMU-PS dataset provides 1,912 examples for training and 846 examples for testing. The 2D hand keypoints of these examples are available.

Evaluation Metrics. Following [4, 24, 39], we adopt two metrics for evaluating the estimated hand poses, including the average *End-Point-Error* (EPE) and the *Area Under the Curve* (AUC) on the *Percentage of Correct Keypoints* (PCK). We report the performance on both 2D and 3D hand pose estimation where the performance metrics are computed in pixels (px) and millimeters (mm), respectively. The performance of 3D hand joint prediction is measured by using the PCK curves averaged over all 21 keypoints. We use 2D PCK and 3D PCK to evaluate our approach on the RHP, STB, and CMU-PS datasets, respectively.

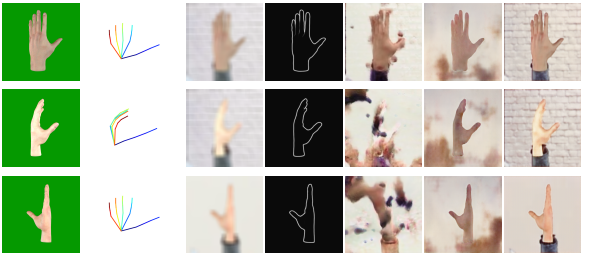


Figure 6: Comparison of the hand pose images synthesized by different methods. Images from left to right are 1) hand poses generated by the AR simulator, 2) their keypoint annotations, 3) color maps, 4) shape maps, the synthesized results by 5) CycleGAN [37], 6) Pix2pix Net [17], and 7) the proposed TAGAN, respectively.

Figure 7: Tonality inconsistency in the "AR Hand w/RBG" data.

5 Experimental Results

This section evaluates the proposed TAGAN. First, we visually inspect the hand images generated by TAGAN and compare them with those generated via CycleGAN [37] and Pix2pix Net [17]. Second, we perform an ablation study to independently verify the efficacy of additional hand pose data and the importance of using TAGAN as a background-aware generator. Third, we show that the hand pose estimators can be greatly improved by fine-tuning with the data generated by TAGAN.

5.1 Comparison of the Synthesized Hand Pose Images

Figure 6 compares the quality of the generated hand images by using CycleGAN, Pix2pix Net, and the proposed TAGAN, respectively. The CycleGAN learns one translation mapping from synthetic hand images to real images, and another mapping from real to synthetic images. The Pix2pix Net derives the translation from a shape map to a realistic image. TAGAN takes both foreground shapes and background tonality into account. In Figure 6, the first two columns show the synthetic hand images generated by the AR simulator and their hand-keypoint labels, respectively. The third and the forth columns show the pre-computed color maps and shape maps from those synthetic hand images. The last three columns display the results generated by CycleGAN, Pix2pix Net, and TAGAN, respectively.

As shown in Figure 6, the CycleGAN does not have the shape and color constraints, so it tends to generate unnatural hand images. Although the Pix2pix Net can successfully generate hand's shapes by using shape features, it does not consider the consistency of colors. The generated hand images still look unnatural. The proposed TAGAN leverages real background information and gains color and shape constrains. Hence, the generated results jointly maintain the color and shape features, making the synthesized images more realistic.

Table 1: 3D pose estimation results.

	AUC↑	EPE mean↓
RHP	0.42	35.6
RHP+AR Hand w/ CBG	0.56	26.4
RHP+AR Hand w/ RBG	0.57	26.4
RHP+AR Hand w/ TAGAN	**0.60**	**24.2**
STB	0.66	15.7
STB+AR Hand w/ CBG	0.67	8.2
STB+AR Hand w/ RBG	0.71	7.1
STB+AR Hand w/ TAGAN	**0.75**	**7.0**

Table 2: 2D pose estimation results.

	PCK@20 ↑	EPE mean ↓
RHP	88.45	10.76
RHP+AR Hand w/ RBG	90.00	9.80
RHP+AR Hand w/ TAGAN	**90.01**	**9.78**
STB	96.6	7.61
STB + AR Hand w/ RBG	96.7	7.59
STB + AR Hand w/ TAGAN	**97.0**	**7.55**

5.2 Ablation Study

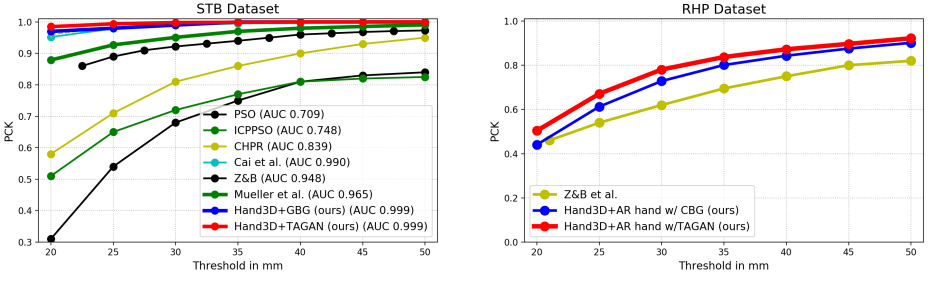
For analyzing the pose estimator learning by using our generated data, we conduct an ablation study on both 2D and 3D hand pose estimation. We adopt three comparative settings of training data on the STB datasets, including 1) STB data with hand images produced by the AR simulator with clear background "STB+AR hand w/ CBG", 2) STB data with AR hand images with real background "STB+AR hand w/ RBG", and 3) STB with hand images generated by TAGAN "STB+AR hand w/ TAGAN". Some examples of the "STB+AR hand w/ CBG" and "STB+AR hand w/ TAGAN" are shown in the first and the last columns of Figure 6, respectively. We train the hand pose estimators, Hand3D and CPM, on the training sets using the settings above, and test them on the validation sets, respectively. All the aforementioned experimental settings are also conducted on the RPH dataset.

Table 1 and Table 2 show the experimental results on 3D and 2D hand pose estimation, respectively. In EPE and PCK, we find that both 3D and 2D hand pose estimators can be greatly improved by using large-scale synthetic data. Besides, Table 1 and Table 2 show that the pose estimators trained with AR hands and real background images can be further improved. The reason is that the augmented AR hand images with real backgrounds are closer to the real images. Moreover, it can be observed that training hand pose estimators with data generated by TAGAN has higher PCK values and lower EPE errors than with the "AR Hand w/ RBG" data.

For 3D pose estimation task, the pose estimator trained with "AR Hand w/ CBG" is successfully improved in EPE mean 11.4 ($= 35.6 - 24.2$)mm and 8.7 ($= 15.7 - 7.0$)mm on the RHP and STB datasets, respectively, as shown in Table 1. For 2D pose estimation, the pose estimators are improved in EPE mean 0.98($= 10.76 - 9.78$) pixels and 0.06($= 7.61 - 7.55$) pixels on the RHP ans STB dataset, respectively, as shown in Table 2. The quality of the data generated by TAGAN is better than the synthetic AR hand superimposed with real backgrounds (AR hands w/ RBG). The reason is that the "AR Hand w/ RBG" data do not take the tonality consistency between synthetic hands and background into account. Some examples of tonality inconsistency are shown in Figure 7. To test the generation ability, we follow [24] where hand pose estimators are trained on the STB training set and evaluated on the CMU-PS validation set. We adopt CPM in this experiment. As shown in Table 3, training the pose estimator with augmented hand images (AR hand w/ RBG) can enhance the performance. Moreover, training it with the data generated by TAGAN can gain more significant improvement in both PCK accuracy and EPE error. The CPM is improved from 24.09 to 28.81 in PCK@20 accuracy and from 55.99 to 52.39 in EPE mean error.

Table 3: Results by training on STB and testing on CMU-PS.

	PCK@20 ↑	EPE mean (mm) ↓
STB	24.09	55.99
STB+AR Hand w/ RBG	24.82	56.67
STB+AR Hand w/ TAGAN	**28.81**	**52.93**



(a) (b)

Figure 8: Comparison with the state-of-the-art approaches for 3D pose estimation on the (a) STB and (b) RHP datasets.

5.3 Comparison with State-of-the-Art Results

To compare with the existing 3D hand pose estimators, we select the methods, PSO, ICPPSO, and CHPR as the baselines, and choose the state-of-the art methods including Cai *et al.* in [4], Z&B in [39], and Mueller *et al.* in [17] for comparison on the STB dataset. We select Z&B [39] for comparison on the RHP dataset. We also provide the results by training our pose estimator with "AR hand w/ RBG" data for comparison. Figure 8 shows that training Hand3D with training data generated by TAGAN achieves the best performance. To explore the improvement by training a 2D/3D pose estimator with our generated data, we conduct two experiments on STB and RHP datasets. We adopt Hand3D and CPM as the 2D and 3D pose estimators. Table 1 and 2 show the results. We find that training either 2D or 3D pose estimators with the additional images generated by TAGAN reaches the best performance.

6 Conclusions and Future Work

This study presents a novel data augmentation approach for improving hand pose estimation task. To produce more realistic hand images for training pose estimators, we propose TAGAN, a conditional adversarial networks model, to blend the synthetic hand poses with real background images. Our generated results align the hand shape and color tonality distribution between synthetic hands and real background images. The experimental results show that the state-of-the-arts hand pose estimators can be greatly improved with the aid of the training data generated by our method.

Acknowledgement. This work was supported in part by Ministry of Science and Technology (MOST) under grants 107-2628-E-001-005-MY3 and 108-2634-F-007-009.

References

[1] Masoud Abdi, Ehsan Abbasnejad, Chee Peng Lim, and Saeid Nahavandi. 3d hand pose estimation using simulation and partial-supervision with a shared latent space. In *BMVC*, 2018.

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In *CVPR*, 2018.

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

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, volume 12, 2018.

[5] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018.

[6] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *CVPR*, 2018.

[7] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[9] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[10] Fuyang Huang, Ailing Zeng, Minhao Liu, Jing Qin, and Qiang Xu. Structure-aware 3d hourglass network for hand pose estimation from single depth image. In *BMVC*, 2018.

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), jul 2014.

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

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

[14] Aisha Urooj Khan and Ali Borji. Analysis of hand segmentation in the wild. In *CVPR*, 2018.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[16] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *CVPR*, 2018.

[17] 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 *CVPR*, 2018.

[18] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, volume 1, 2011.

[19] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, 2012.

[20] Rohit Pandey, Pavel Pidlypenskyi, Shuoran Yang, and Christine Kaeser-Chen. Efficient 6-dof tracking of handheld objects from an egocentric viewpoint. In *ECCV*, 2018.

[21] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, 2018.

[22] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

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

[24] Tomas Simon, Hanbyul Joo, Iain A Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[25] Srinath Sridhar, Anna Maria Feit, Christian Theobalt, and Antti Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *CHI*, 2015.

[26] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. *CVPR*, 2019.

[27] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *CVPR*, volume 1, page 7. Citeseer, 2011.

[28] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *CVPR*, 2018.

[29] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. In *Arxiv*, page 1905.07718, 2019.

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

[31] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *CVPR*, 2015.

[32] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. *CVPR*, 2019.

[33] Qi Ye and Tae-Kyun Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In *ECCV*, 2018.

[34] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *CVPR*, 2018.

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

[36] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. In *ICIP*, 2016.

[37] Yidan Zhou, Jian Lu, Kuo Du, Xiangbo Lin, Yi Sun, and Xiaohong Ma. Hbe: Hand branch ensemble network for real-time 3d hand pose estimation. In *ECCV*, 2018.

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

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