# GestureGAN for Hand Gesture-to-Gesture Translation in the Wild

Hao Tang[1], Wei Wang[1,2], Dan Xu[1,3], Yan Yan[4], Nicu Sebe[1]
[1]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[2]Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[3]Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[4]Department of Computer Science, Texas State University, San Marcos, USA
{hao.tang, niculae.sebe}@unitn.it, wei.wang@epfl.ch, danxu@robots.ox.ac.uk, y_y34@txstate.edu

## ABSTRACT

Hand gesture-to-gesture translation in the wild is a challenging task since hand gestures can have arbitrary poses, sizes, locations and self-occlusions. Therefore, this task requires a high-level understanding of the mapping between the input source gesture and the output target gesture. To tackle this problem, we propose a novel hand Gesture Generative Adversarial Network (GestureGAN). GestureGAN consists of a single generator $G$ and a discriminator $D$, which takes as input a conditional hand image and a target hand skeleton image. GestureGAN utilizes the hand skeleton information explicitly, and learns the gesture-to-gesture mapping through two novel losses, the color loss and the cycle-consistency loss. The proposed color loss handles the issue of "channel pollution" while back-propagating the gradients. In addition, we present the Fréchet ResNet Distance (FRD) to evaluate the quality of generated images. Extensive experiments on two widely used benchmark datasets demonstrate that the proposed GestureGAN achieves state-of-the-art performance on the unconstrained hand gesture-to-gesture translation task. Meanwhile, the generated images are in high-quality and are photo-realistic, allowing them to be used as data augmentation to improve the performance of a hand gesture classifier. Our model and code are available at https://github.com/Ha0Tang/GestureGAN.

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Machine learning**;

## KEYWORDS

Generative Adversarial Networks; Image Translation; Hand Gesture

**ACM Reference Format:**
Hao Tang, Wei Wang, Dan Xu, Yan Yan, Nicu Sebe. 2018. GestureGAN for Hand Gesture-to-Gesture Translation in the Wild. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240704


*MM '18, October 22–26, 2018, Seoul, Republic of Korea*

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240704

## 1 INTRODUCTION

Hand gesture-to-gesture translation in the wild is a task that converts the hand gesture of a given image to a target gesture with a different pose, size and location while preserving the identity information. This task has many applications, such as human-computer interactions, entertainment, virtual reality and data augmentation. However, this task is difficult since it needs (i) handling complex backgrounds with different illumination conditions, objects and occlusions; (ii) a high-level semantic understanding of the mapping between the input and output gestures.

Recently, Generative Adversarial Networks (GANs) [7] have shown the potential to solve this challenging task. GAN is a generative model based on game theory, which has achieved impressive performance in many applications, such as high-quality image generation [12], video generation [56] and audio generation [30]. To generate specific kinds of images, videos and audios, Mirza et al. [28] proposed the Conditional GAN (CGAN), which comprises a vanilla GAN and other external information, such as class labels [3], text descriptions [37], images [10] and object keypoints [37].

In this paper, we focus on the image-to-image translation task using CGAN. Image-to-image translation tasks can be divided into two types: paired [10, 24, 41] and unpaired [1, 3, 58, 62]. However, existing image-to-image translation frameworks are inefficient in the multi-domain image-to-image translation task. For instance, given $m$ image domains, pix2pix [10] and BiCycleGAN [63] need to train $A_m^2 = m(m-1) = \Theta(m^2)$ models. CycleGAN [62], DiscoGAN [13] and DualGAN [58] need to train $C_m^2 = \frac{m(m-1)}{2} = \Theta(m^2)$ models, or $m(m-1)$ generator/discriminator pairs since one model has 2 different generator/discriminator pairs for these methods. ComboGAN [1] requires $m = \Theta(m)$ models. StarGAN [3] needs one model. However, for some specific image-to-image translation applications such as hand gesture-to-gesture translation, $m$ could be arbitrary large since gestures in the wild can have arbitrary poses, sizes, appearances, locations and self-occlusions.

To address these limitations, several works have been proposed to generate images based on object keypoints. For instance, Reed et al. [36] present an extension of Pixel Convolutional Neural Networks (PixelCNN) to generate images based on keypoints and text description. Siarohin et al. [41] introduce a deformable Generative Adversarial Network for pose-based human image generation. Ma et al. [25] propose a two-stage reconstruction pipeline that learns generates novel person images. However, the aforementioned methods always have the "channel pollution" problem that is frequently occurring in generative models such as PG$^2$ [24] leading to blurred generated images. To solve this issue, in this paper, we propose

a novel Generative Adversarial Network, *i.e.*, GestureGAN which treats each channel independently. It allows generating high-quality hand gesture images with arbitrary poses, sizes and locations in the wild, and thus reducing the dependence on environment and preprocessing operations. GestureGAN only consists of one generator and one discriminator, taking a conditional hand gesture image and a target hand skeleton image as inputs. In addition, to better learn the mapping between inputs and outputs, we propose two novel losses, *i.e.*, color loss and cycle-consistency loss. Note that the color loss can handle the problem of "channel pollution", making the generated images sharper and having higher quality. Furthermore, we propose the Fréchet ResNet Distance (FRD), which is a novel evaluation metric to evaluate the generated image of GANs. Extensive experiments on two public benchmark datasets demonstrate that GestureGAN can generate high-quality images with convincing details. Thus, these generated images can augment the training data and improve the performance of hand gesture classifiers.

Overall, the contributions of this paper are as follows:

- We propose a novel Generative Adversarial Network, *i.e.*, GestureGAN, which can generate target hand gesture with arbitrary poses, sizes and locations in the wild. In addition, we present a novel color loss to learn the hand gesture-to-gesture mappings, handling the problem of "channel pollution".
- We propose an efficient Fréchet ResNet Distance (FRD) metric to evaluate the similarity of the real and generated images, which is more consistent with human judgment. FRD measures the similarity between the real image and the generated image in a high-level semantic feature space.
- Qualitative and quantitative results demonstrate the superiority of the proposed GestureGAN over the state-of-the-art models on the unconstrained hand gesture-to-gesture translation task. In addition, the generated hand gesture images are of high-quality and are photo-realistic, thus allowing them to be used to boost the performance of hand gesture classifiers.

## 2 RELATED WORK

**Generative Adversarial Network (GAN)** [7] is an unsupervised learning method and has been proposed by Goodfellow et al. Recently, GAN has shown outstanding results in various applications, *e.g.*, image generation [2, 12], image editing [33, 40], video generation [26, 48], texture synthesis [17], music generation [57] and feature learning [54]. Recent approaches employ the idea of GAN for conditional image generation, such as image-to-image translation [10, 49], text-to-image translation [35, 59], image inpainting [6, 19], image blending [52], image super-resolution [16], as well as the applications of other domains like semantic segmentation [23], object detection [18, 50], human parsing [21], face aging [20] and 3D vision [31, 53]. The key point success of GANs in computer vision and graphics is the adversarial loss, which allows the model to generate images that are indistinguishable from real images, and this is exactly the goal that many computer vision and graphics tasks aim to optimize.

**Image-to-Image Translation** frameworks use input-output data to learn a parametric mapping between inputs and outputs, *e.g.*, Isola et al. [10] build the pix2pix model, which uses a conditional GAN to learn a translation function from input to output image domains. Taigman et al. [46] propose the Domain Transfer Network (DTN) which learns a generative function between one domain and another domain. Zhu et al. [62] introduce the CycleGAN framework, which achieves unpaired image-to-image translation using the cycle-consistency loss. Moreover, Zhu et al. [63] present the BicycleGAN model based on CycleGAN [62] and pix2pix [10], which targets multi-modal image-to-image translation.

However, existing image-to-image translation models are inefficient and ineffective. For example, with $m$ image domains, CycleGAN [62], DiscoGAN [13], and DualGAN [58] need to train $2C_m^2 = m(m-1) = \Theta(m^2)$ generators and discriminators, while pix2pix [10] and BicycleGAN [63] have to train $A_m^2 = m(m-1) = \Theta(m^2)$ generator/discriminator pairs. Recently, Anoosheh et al. proposed ComboGAN [1], which only need to train $m$ generator/discriminator pairs for $m$ different image domains, having a complexity of $\Theta(m)$. Additionally, Choi et al. [3] propose StarGAN, in which a single generator and discriminator can perform unpaired image-to-image translations for multiple domains. Although the computational complexity of StarGAN is $\Theta(1)$, this model has only been validated on the face attributes modification task with clear background and face cropping. More importantly, for some specific image-to-image translation tasks such as hand gesture-to-gesture translation task, the image domains could be arbitrary large, *e.g.*, gesture in the wild can have arbitrary poses, sizes, appearances, locations and self-occlusions. The aforementioned approaches are not effective for solving these specific situations.

**Keypoint/Skeleton Guided Image-to-Image Translation.** To fix these limitations, several recent works have been proposed to generate person, bird or face images based on object keypoints [24, 37] or human skeleton [41, 56]. For instance, Di et al. [5] propose the Gender Preserving Generative Adversarial Network (GPGAN) to synthesize faces based on facial landmarks. Reed et al. [37] propose the Generative Adversarial What-Where Network (GAWWN), which generates birds conditioned on both text descriptions and object location. Ma et al. propose the Pose Guided Person Generation Network ($PG^2$) [24] and a two-stage reconstruction pipeline [25], which achieve person-to-person image translation using a conditional image and a target pose image (note that in these two models images are pre-cropped). Reed et al. [36] present an extension of Pixel Convolutional Neural Networks (PixelCNN) to generate images parts based on keypoints and text description. Sun et al. [45] propose a two-stage framework to perform head inpainting conditioned on the generated facial landmark in the first stage. Korshunova et al. [15] use facial keypoints to define the affine transformations of the alignment and realignment steps for face swap. Wei et al. [51] propose a Conditional MultiMode Network (CMM-Net) for landmark-guided smile generation. Qiao et al. [34] present the Geometry-Contrastive Generative Adversarial Network (GCGAN) to generate facial expressions conditioned on geometry information of facial landmarks. Song et al. [44] propose the Geometry-Guided Generative Adversarial Network (G2GAN) for facial expression synthesis guided by fiducial points. Yan et al. [56] propose a method to generate human motion sequence with simple background using CGAN and human skeleton information. Siarohin et al. [41] introduce PoseGAN for pose-based human image generation using human skeleton.

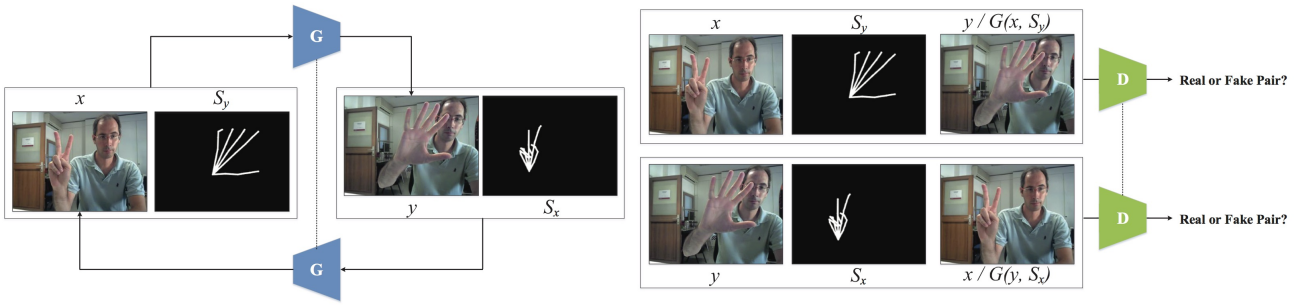


**Figure 1: Pipeline of the proposed GestureGAN model. GestureGAN consists of a single generator $G$ and a discriminator $D$, which takes as input a conditional hand image and a target hand skeleton image.**

The typical problem with the aforementioned generative models is that they suffer from "channels pollution" and thus they tend to generate blurry results with artifacts. To handle this problem, we propose GestureGAN, which allows generating high-quality hand gesture image with arbitrary poses, sizes and locations in the wild.

## 3 MODEL DESCRIPTION

### 3.1 GestureGAN Objective

The goal of vanilla GAN is to train a generator $G$ which learns the mapping between random noise $z$ and image $y$. The mapping function of GAN, $G(z) \mapsto y$ is learned via the following objective function,

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y \left[\log D(y)\right] + \mathbb{E}_z \left[\log(1-D(G(z)))\right]. \tag{1}$$

Conditional GANs learn the mapping $G(x, z) \mapsto y$, where $x$ is the input conditional image. Generator $G$ is trained to generate image $\widehat{y}$ that cannot be distinguished from "real" image $y$ by an adversarially trained discriminator $D$, while the discriminator $D$ is trained as well as possible to detect the "fake" images generated by the generator $G$. The objective function of the conditional GAN is defined as follows,

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y} \left[\log D(x, y)\right] + \mathbb{E}_{x,z} \left[\log(1 - D(x, G(x, z)))\right], \tag{2}$$

where generator $G$ tries to minimize this objective while the discriminator $D$ tries to maximize it. Thus, the solution is $G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D)$. In this paper, we try to learn two mappings through one generator. The framework of the proposed GestureGAN is shown in Figure 1.

**Adversarial Loss.** In order to learn the gesture-to-gesture mapping, we employ the hand skeleton information explicitly. We exploit OpenPose [42] to detect 21 hand keypoints denoted as $(p_i, q_i | i = 1, 2, \cdots, 21)$, where $p_i$ and $q_i$ represent pixel coordinates of keypoints. For each keypoint $(p_i, q_i)$, $c_i \in [0, 1]$ represents the confidence that the keypoint is correctly localized. Thus, the adversarial losses of the two mappings $G([x, K_y], z_1) \mapsto y$ and $G([y, K_x], z_2) \mapsto x$ are defined respectively, as follows:

$$\begin{aligned}\mathcal{L}_{K_y}(G, D, K_y) =& \mathbb{E}_{[x,K_y],y} \left[\log D([x, K_y], y)\right] + \\ & \mathbb{E}_{[x,K_y],z_1} \left[\log(1 - D([x, K_y], G([x, K_y], z_1)))\right],\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{K_x}(G, D, K_x) =& \mathbb{E}_{[y,K_x],x} \left[\log D([y, K_x], x)\right] + \\ & \mathbb{E}_{[y,K_x],z_2} \left[\log(1 - D([y, K_x], G([y, K_x], z_2)))\right],\end{aligned} \tag{4}$$

where $K_y$ and $K_x$ are the hand keypoints of image $y$ and $x$ respectively; $[\cdot, \cdot]$ represents the concatenate operation. $K_y$ and $K_x$ are defined by setting the pixels around the corresponding keypoint $(p_i, q_i)$ to 1 (white) with the radius of 4 and 0 (black) elsewhere. In other words, each keypoint is actually represented with pixels in a circle with a radius of 4. Therefore, the total adversarial loss based on hand keypoint can be defined as,

$$\mathcal{L}_K(G, D, K_x, K_y) = \mathcal{L}_{K_y}(G, D, K_y) + \mathcal{L}_{K_x}(G, D, K_x). \tag{5}$$

In addition, to explore the influence of the confidence score $c_i$ to the generated image, we define a confidence keypoint image $\widehat{K}$ in which the pixels around the corresponding keypoint $(p_i, q_i)$ are set to $c_i$ in a radius of 4 pixels and 0 (black) elsewhere. Thus, Equation 5 can be expressed as $\mathcal{L}_{\widehat{K}}(G, D, \widehat{K_x}, \widehat{K_y}) = \mathcal{L}_{\widehat{K_y}}(G, D, \widehat{K_y}) + \mathcal{L}_{\widehat{K_x}}(G, D, \widehat{K_x})$.

Moreover, following OpenPose [42], we connect the 21 keypoints (hand joints) to obtain the hand skeleton, denoted as $S_x$ and $S_y$. The hand skeleton image visually contains richer hand structure information than the hand keypoint image. Next, the adversarial loss based on hand skeleton can be derived from Equation 5, *i.e.*, $\mathcal{L}_S(G, D, S_x, S_y) = \mathcal{L}_{S_y}(G, D, S_y) + \mathcal{L}_{S_x}(G, D, S_x)$. In hand skeleton image $S_x$ and $S_y$, the hand joints are connected by the lines with the width of 4 and with white color. Next, corresponding to the confidence keypoint image, we have also defined an adversarial loss using confidence hand skeleton as $\mathcal{L}_{\widehat{S}}(G, D, \widehat{S_x}, \widehat{S_y}) = \mathcal{L}_{\widehat{S_y}}(G, D, \widehat{S_y}) + \mathcal{L}_{\widehat{S_x}}(G, D, \widehat{S_x})$, where the line connections in $\widehat{S_x}$ and $\widehat{S_y}$ are filled with the confidence score of later point, *e.g.*, if the hand skeleton connects points 1 and 2, thus this line connection is filled with the confidence of point 2, *i.e.*, $c_2$ with the width of 4.

**Improved Pixel Loss.** Previous work indicated that mixing the adversarial loss with a traditional loss such as $L_1$ loss [10] or $L_2$ loss [32] between the generated image and the ground truth image improves the quality of generated images. The definition of $L_1$ and $L_2$ losses are:

$$\begin{aligned}\mathcal{L}_{L_{\{1,2\}}}(G, S_x, S_y) =& \mathbb{E}_{[x,S_y],y,z_1} \left[\|y - G([x, S_y], z_1)\|_{\{1,2\}}\right] + \\ & \mathbb{E}_{[y,S_x],x,z_2} \left[\|x - G([y, S_x], z_2)\|_{\{1,2\}}\right].\end{aligned} \tag{6}$$

However, we observe that the existing image-to-image translation models such as PG$^2$ [24] cannot retain the holistic color of the input images. An example is shown in Figure 2, where PG$^2$ is affected by the pollution issue and produces more unrealistic regions. Therefore, to remedy this limitation we introduce a novel channel-wise color loss. Traditional generative models convert a whole image to another, which leads to the "channel pollution"


**Figure 2: Illustration of the "channel pollution" issue on different methods.**

problem. However, the color loss treats $r$, $g$ and $b$ channels independently and generates only one channel each time, and then these three channels are combined to produce the final image. Intuitively, since the generation of a three-channel image space is much more complex than the generation of a single-channel image space, leading to higher possibility of artifacts, we independently generate each channel. The objective of $r$, $g$ and $b$ channel losses can be defined as follows,

$$\mathcal{L}_{Color^c_{\{1,2\}}}(G, S_x, S_y) = \mathbb{E}_{[x^c, S_y], y^c, z_1}\left[\|y^c - G([x^c, S_y], z_1)\|_{\{1,2\}}\right] + \mathbb{E}_{[y^c, S_x], x^c, z_2}\left[\|x^c - G([y^c, S_x], z_2)\|_{\{1,2\}}\right], \tag{7}$$

where $c \in \{r, g, b\}$, $x^r$, $x^g$ and $x^b$ denote the $r$, $g$ and $b$ channels of image $x$ respectively and similar to $y^r$, $y^g$ and $y^b$, $\|\cdot\|_1$ and $\|\cdot\|_2$ represent $L_1$ and $L_2$ distance losses. Thus, the color $L_1$ and $L_2$ losses can be expressed as,

$$\mathcal{L}_{Color_{\{1,2\}}}(G, S_x, S_y) = \mathcal{L}_{Color^r_{\{1,2\}}} + \mathcal{L}_{Color^g_{\{1,2\}}} + \mathcal{L}_{Color^b_{\{1,2\}}}. \tag{8}$$

When back-propagating the gradients of the $L_1$ loss, the partial derivatives are constants, *i.e.*, $\pm 1$. Therefore, the error from other channels will not influence the current one as the derivative is a constant. However, for the original $L_2$ loss, the derivative is not a fixed constant. Actually, the derivative for the variables in one channel is always influenced by the errors from other channels. We have listed the gradients of red channel of Equations 6 and 8. Let $\widehat{y}$ represent the generated target image $G([x, S_y], z_1)$, we have,

$$\begin{aligned}
&\frac{\partial}{\partial \widehat{y}_r^{i_o,j_o}} \mathcal{L}_{L_2}(G, S_x, S_y) \\
&= \frac{\partial}{\partial \widehat{y}_r^{i_o,j_o}} \sqrt{\sum_{i,j}(y_r^{i,j} - \widehat{y}_r^{i,j})^2 + \sum_{i,j}(y_g^{i,j} - \widehat{y}_g^{i,j})^2 + \sum_{i,j}(y_b^{i,j} - \widehat{y}_b^{i,j})^2} \\
&= \frac{y_r^{i_o,j_o} - \widehat{y}_r^{i_o,j_o}}{\sqrt{\sum_{i,j}(y_r^{i,j} - \widehat{y}_r^{i,j})^2 + \sum_{i,j}(y_g^{i,j} - \widehat{y}_g^{i,j})^2 + \sum_{i,j}(y_b^{i,j} - \widehat{y}_b^{i,j})^2}}.
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\frac{\partial}{\partial \widehat{y}_r^{i_o,j_o}} \mathcal{L}_{Color_2}(G, S_x, S_y) \\
&= \frac{\partial}{\partial \widehat{y}_r^{i_o,j_o}} \left( \sqrt{\sum_{i,j}(y_r^{i,j} - \widehat{y}_r^{i,j})^2} + \sqrt{\sum_{i,j}(y_g^{i,j} - \widehat{y}_g^{i,j})^2} + \sqrt{\sum_{i,j}(y_b^{i,j} - \widehat{y}_b^{i,j})^2} \right) \\
&= \frac{y_r^{i_o,j_o} - \widehat{y}_r^{i_o,j_o}}{\sqrt{\sum_{i,j}(y_r^{i,j} - \widehat{y}_r^{i,j})^2}}.
\end{aligned} \tag{10}$$

Therefore, we can calculate the gradient of original $L_2$ and color $L_2$ losses,

$$\begin{aligned}
\nabla \mathcal{L}_{L_2}(G, S_y) = &\frac{\partial}{\partial \widehat{y}_r^{i_o,j_o}} \mathcal{L}_{L_2}(G, S_y) + \\
&\frac{\partial}{\partial \widehat{y}_g^{i_o,j_o}} \mathcal{L}_{L_2}(G, S_y) + \frac{\partial}{\partial \widehat{y}_b^{i_o,j_o}} \mathcal{L}_{L_2}(G, S_y).
\end{aligned} \tag{11}$$

$$\begin{aligned}
\nabla \mathcal{L}_{Color_2}(G, S_y) = &\frac{\partial}{\partial \widehat{y}_r^{i_o,j_o}} \mathcal{L}_{Color_2}(G, S_y) + \\
&\frac{\partial}{\partial \widehat{y}_g^{i_o,j_o}} \mathcal{L}_{Color_2}(G, S_y) + \frac{\partial}{\partial \widehat{y}_b^{i_o,j_o}} \mathcal{L}_{Color_2}(G, S_y).
\end{aligned} \tag{12}$$

Clearly, in Equation 9 the red channel in original $L_2$ loss is polluted by green and blue channels. As a consequence, the error from other channels will also influence the red channel. On the contrary, if we compute the loss for each channel independently, we can avoid such influence as shown in Equation 10.

**Cycle-Consistency Loss.** It is worth noting that the CycleGAN [62] is different from pix2pix framework [10] as the training data in CycleGAN is unpaired. The CycleGAN introduces the cycle-consistency loss to enforce forward-backward consistency. The cycle-consistency loss can be regarded as "pseudo" pairs of training data even though we do not have the corresponding data in the target domain which corresponds to the input data from the source domain. However, in this paper we introduce the cycle-consistency loss for the paired image-to-image translation task. The cycle loss ensures the consistency between source images and the reconstructed image, and it can be expressed as,

$$\begin{aligned}
\mathcal{L}_{cyc}(G, S_x, S_y) = &\mathbb{E}_{x, y, S_x, S_y, z_1, z_2}\left[\left\|x - G(G([x, S_y], z_1), S_x, z_2)\right\|_1\right] + \\
&\mathbb{E}_{x, y, S_x, S_y, z_1, z_2}\left[\left\|y - G(G([y, S_x], z_2), S_y, z_1)\right\|_1\right].
\end{aligned} \tag{13}$$

Similar to StarGAN [3] we use the same generator $G$ two times, with the first time to convert an original image into the target one, then to recover the original image from the generated image.

**Identity Preserving Loss.** To preserve the person identity after image synthesis, we propose the identity preserving loss, which can be expressed as follows,

$$\begin{aligned}
\mathcal{L}_{identity}(G, S_x, S_y) = &\mathbb{E}_{x, S_y, z_1}\left[\left\|F(y) - F(G([x, S_y], z_1))\right\|_1\right] + \\
&\mathbb{E}_{y, S_x, z_2}\left[||F(x) - F(G([y, S_x], z_2))||_1\right],
\end{aligned} \tag{14}$$

where $F$ is a feature extractor. The feature extractor employs a VGG network [43] originally pretrained for face recognition. We minimize the difference between the feature maps which are generated from the real and the generated images via the pretrained CNN for identity preservation.

**Overall Loss.** The final objective of the proposed GestureGAN is,

$$\begin{aligned}
\mathcal{L} = &\mathcal{L}_S(G, D, S_x, S_y) + \lambda_1 \mathcal{L}_{Color_{\{1,2\}}}(G, S_x, S_y) + \\
&\lambda_2 \mathcal{L}_{cyc}(G, S_x, S_y) + \lambda_3 \mathcal{L}_{identity}(G, S_x, S_y),
\end{aligned} \tag{15}$$

where, $\lambda_1$ , $\lambda_2$ and $\lambda_3$ are three hyper-parameters controlling the relative importance of these four losses. In our experiments, we follow the same setup of pix2pix [10]. Instead of using the random noise vector $z$, we provide noise only in the form of dropout in generator $G$.

## 3.2 Network Architecture

**Generator.** We adopt the "U-shaped" network [10] as our generator. U-net has skip connections, which concatenate all channels at layer $l$ with those at layer $n-l$, where $n$ is the total number of layers.

**Discriminator.** We employ PatchGAN [10] as our discriminator architecture. The goal of PatchGAN is to classify each small patch in an image as real or fake. We run PatchGAN convolutationally across an image, then average all results to calculate the ultimate output of discriminator $D$.

### 3.3 Optimization

We follow the standard optimization method from [7] to optimize the proposed GestureGAN, *i.e.*, we alternate between one gradient descent step on discriminator $D$, and one step on generator $G$. In addition, as suggested in the original GAN paper [7], we train to maximize $\log D([x, S_y], \widehat{y})$ rather than to minimize $\log(1 - D([x, S_y], \widehat{y}))$. Moreover, in order to slow down the rate of $D$ relative to $G$ we divide the objective by 2 while optimizing $D$,

$$\begin{aligned}\mathcal{L}(D) =& \frac{1}{2}\left[\mathcal{L}_{bce}(D([x, S_y], y), 1) + \mathcal{L}_{bce}(D([x, S_y], G([x, S_y], z_1)), 0)\right] \\ &+ \frac{1}{2}\left[\mathcal{L}_{bce}(D([y, S_x], x), 1) + \mathcal{L}_{bce}(D([y, S_x], G([y, S_x], z_2)), 0)\right],\end{aligned} \tag{16}$$

where $\mathcal{L}_{bce}$ denotes the Binary Cross Entropy loss function. We also employ dual discriminators as in Xu et al. [55], Nguyen et al. [29] and CycleGAN [62], which have demonstrated that they improve the ability of discriminator to generate more photo-realistic images. Thus Equations 16 is modified as:

$$\begin{aligned}&\mathcal{L}(D_1, D_2) = \\ &\frac{1}{2}\left[\mathcal{L}_{bce}(D_1([x, S_y], y), 1) + \mathcal{L}_{bce}(D_1([x, S_y], G([x, S_y], z_1)), 0)\right] + \\ &\frac{1}{2}\left[\mathcal{L}_{bce}(D_2([y, S_x], x), 1) + \mathcal{L}_{bce}(D_2([y, S_x], G([y, S_x], z_2)), 0)\right].\end{aligned} \tag{17}$$

We employ the minibatch SGD algorithm and apply Adam optimizer [14] as solver. The momentum terms $\beta_1$ and $\beta_2$ of Adam are 0.5 and 0.999, respectively. The initial learning rate for Adam is 0.0002.

## 4 EXPERIMENTS

### 4.1 Experimental Setup

**Datasets.** We evaluate the proposed GestureGAN on two public hand gesture datasets: NTU Hand Digit [38] and Creative Senz3D [27], which include different hand gestures. NTU Hand Digit dataset [38] contains 10 hand gestures (*e.g.*, decimal digits from 0 to 9) color images and depth maps collected with a Kinect sensor under cluttered background. The total images in this dataset are 10 gestures × 10 subjects × 10 times = 1000 images. All images are in 640×480 resolution. In our experiment, we only use the RGB images. We randomly select 84,636 pairs, each of which is comprised of two images of the same person but different gestures. 9,600 pairs are randomly selected for the testing subset and the rest of 75,036 pairs as the training set. Creative Senz3D dataset [27] includes static hand gestures performed by 4 people, each performing 11 different gestures repeated 30 times each in the front of a Creative Senz3D camera. The overall number of images of this dataset is 4 subjects × 11 gestures × 30 times = 1320. All images are in resolution 640×480. In our experiment, we only use the RGB images. We randomly select 12,800 pairs and 135,504 pairs as the testing and training set, each pair being composed of two images of the same person but different gestures.

**Implementation Details.** For both datasets, we do left-right flip for data augmentation and random crops are disabled in this experiment as was done in PG$^2$ [24]. For the embedding method, skeleton images are fed into an independent encoder similar to PG$^2$ [24], then we extract the fully connected layer feature vector to concatenate it with the image embedding at the bottleneck fully connected layer. For optimization, models are trained with a mini-batch size of 8 for 20 epochs on both datasets. Hyper-parameters are set empirically with $\lambda_1$=100, $\lambda_2$=10. $\lambda_3$ is 0.1 in the beginning and is gradually increased to 0.5. At inference time, we follow the same settings of PG$^2$ [24], Ma et al. [25] and PoseGAN [41] to randomly select the target keypoint or skeleton. GestureGAN is implemented using the public deep learning framework PyTorch. To speed up the training and testing processes, we use a Nvidia TITAN Xp GPU with 12G memory.

**Evaluation Metrics.** Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Inception Score (IS) [39], Fréchet Inception Distance (FID) [9] and the proposed Fréchet ResNet Distance (FRD) are employed to evaluate the quality of generated images. FRD approach provides an alternative method to quantify the quality of synthesis and is similar to the Fréchet Inception Distance (FID) proposed by [9]. FID is a measure of similarity between two datasets of images. The authors in [9] have shown that the FID is more robust to noise than IS [39] and correlates well with the human judgment of visual quality [9]. To calculate FID [9] between two image domains $x$ and $y$, they first embed both into a feature space $\mathcal{F}$ given by an Inception model. Then viewing the feature space as a continuous multivariate Gaussian as suggested in [9], the Fréchet distance between the two Gaussians to quantify the quality of the data and the definition of FID can be expressed as:

$$\text{FID}(x, y) = \|\mu_x - \mu_y\|_2^2 + \text{Tr}(\textstyle\sum_x + \sum_y - 2(\sum_x \sum_y)^{\frac{1}{2}}), \tag{18}$$

where $(\mu_x, \sum_x)$ and $(\mu_y, \sum_y)$ are the mean and the covariance of the data distribution and model distribution, respectfully.

Unlike FID, which regards the datasets $x$ and $y$ as a whole, the proposed FRD is inspired from feature matching methods [47, 60, 61], and separately calculates the Fréchet distance between generated images and real images from the semantical level. In this way, images from two domains do not affect each other when computing the Fréchet distance. Moreover, for FID the number of samples should be greater than the dimension of the coding layer, while the proposed FRD does not have this limitation. We denote $x_i$ and $y_i$ as images in the $x$ and $y$ domains, respectively. For calculating FRD, we first embed both images $x_i$ and $y_i$ into a feature space $\mathcal{F}$ with 4096×1 dimension given by a ResNet pretrained model [8]. We then calculate the Fréchet distance between two feature maps $f(x_i)$ and $f(y_i)$. The Fréchet distance $F(f(x_i), f(y_i))$ is defined as the infimum over all reparameterizations $\alpha$ and $\beta$ of $[0, 1]$ of the maximum over all $t \in [0, 1]$ of the distance in $\mathcal{F}$ between $f(x_i)(\alpha(t))$ and $f(y_i)(\beta(t))$, where $\alpha$ and $\beta$ are continuous, non-decreasing surjections of the range $[0, 1]$. The proposed FRD is a measure of similarity between the feature vector of the real image $f(y_i)$ and the feature vector of the generated image $f(x_i)$ by calculating the Fréchet distance between them. The Fréchet distance is defined as the minimum cord-length sufficient to join a point traveling forward along $f(y_i)$ and one traveling forward along $f(x_i)$, although the rate of travel for each point may not necessarily be uniform. Thus, the definition of FRD between two image domain $x$ and $y$ is,

$$\text{FRD}(x, y) = \frac{1}{N}\sum_{1}^{N} \inf_{\alpha, \beta} \max_{t \in [0,1]} \left\{d\big(f(x_i)(\alpha(t)), f(y_i)(\beta(t))\big)\right\}, \tag{19}$$

where $d$ is the distance function of $\mathcal{F}$, $N$ is the total number of images in $x$ and $y$ domains.

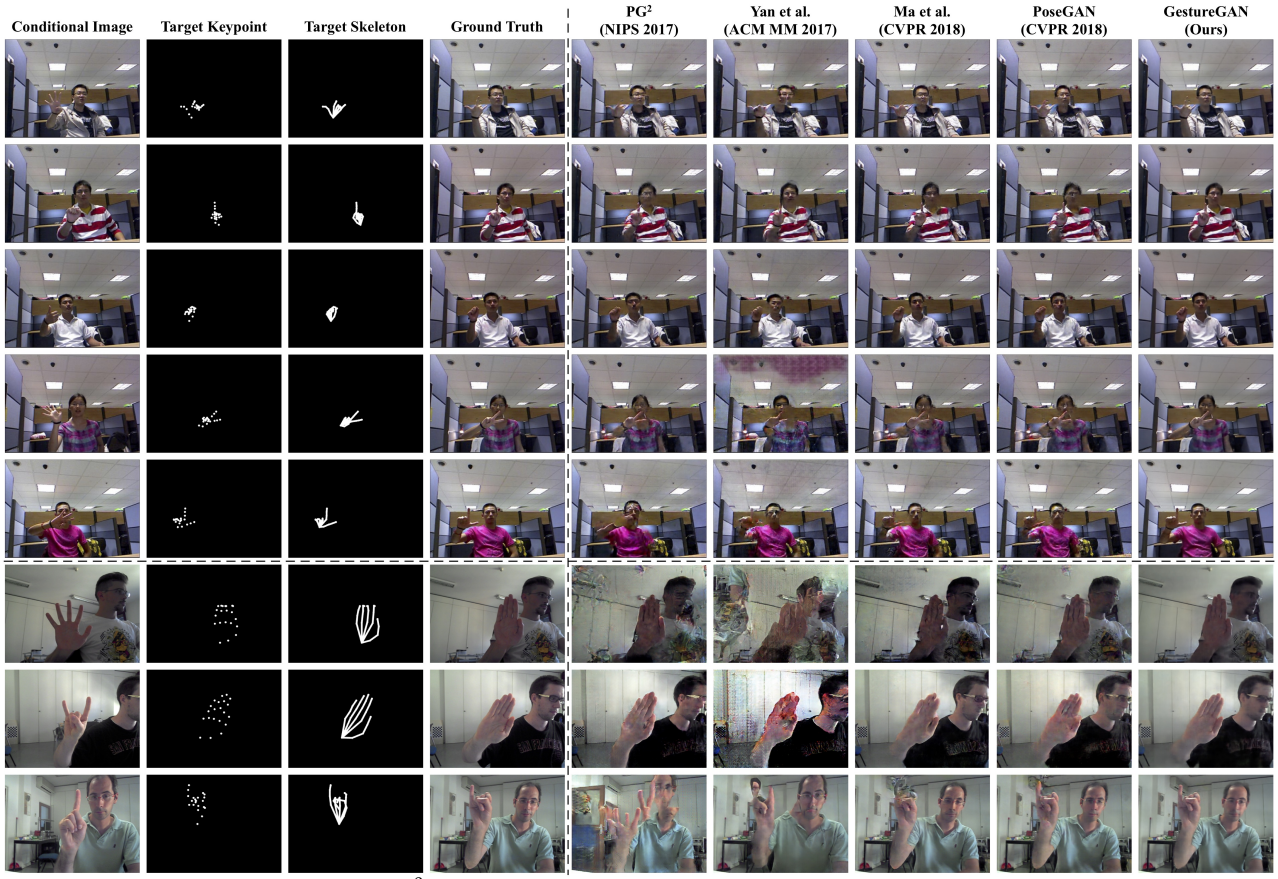


**Figure 3: Qualitative comparison with PG$^2$ [24], Ma et al. [25], Yan et al. [56] and PoseGAN [41] on the NTU Hand Digit (Top) and the Senz3D (Bottom) datasets. Zoom in for details.**

**Table 1: Quantitative results of different models on the NTU Hand Digit and Senz3D datasets. For PSNR and IS measures, higher is better. For MSE evaluation, lower is better.**

| Model | NTU Hand Digit [22] | | | Senz3D [27] | | |
|---|---|---|---|---|---|---|
| | MSE | PSNR | IS | MSE | PSNR | IS |
| PG$^2$ [24] (NIPS 2017) | 116.1049 | 28.2403 | 2.4152 | 199.4384 | 26.5138 | 3.3699 |
| Yan et al. [56] (ACM MM 2017) | 118.1239 | 28.0185 | 2.4919 | 175.8647 | 26.9545 | 3.3285 |
| Ma et al. [25] (CVPR 2018) | 113.7809 | 30.6487 | 2.4547 | 183.6457 | 26.9451 | 3.3874 |
| PoseGAN [41] (CVPR 2018) | 113.6487 | 29.5471 | 2.4017 | 176.3481 | 27.3014 | 3.2147 |
| GestureGAN (Ours) | **105.7286** | **32.6091** | **2.5532** | **169.9219** | **27.9749** | **3.4107** |

## 4.2 Qualitative & Quantitative Results

**Comparison against Baselines.** We compare the proposed GestureGAN with the most related four works, *i.e.*, PG$^2$ [24], Yan et al. [56], PoseGAN [41] and Ma et al. [25]. PG$^2$ [24] and Ma et al. [25] try to generate a person image with different poses based on a conditional person image and a target keypoint image. Yan et al. [56] and PoseGAN [41] explicitly employ human skeleton information to generate person images. Note that Yan et al. [56] adopt a CGAN to generate motion sequences based on appearance information and skeleton information by exploiting frame level smoothness. We re-implemented this model to generate a single frame for fair comparison. These four methods are paired image-to-image models and comparison results are shown in Figure 3 and Table 1. As we can see in Figure 3, GestureGAN produces sharper images with convincing details compared with other baselines. Moreover, it is obvious that our results in Table 1 are consistently much better than baseline methods on both datasets.

**Generated Results of Each Epoch.** Figure 5 (left) illustrates the convergence loss $\mathcal{L}$ of the proposed GestureGAN in Equation 15. Note that the proposed GestureGAN ensures a very fast yet stable convergence.

**Analysis of the Model Components.** In Figure 4 we conduct ablations studies of our model. We gradually remove components of the proposed GestureGAN, *i.e.*, Dual Discriminators (D), Identity Loss (I), Color Loss (P) and Cycle-consistency Loss (C). We find that removing the color loss and the cycle-consistency loss

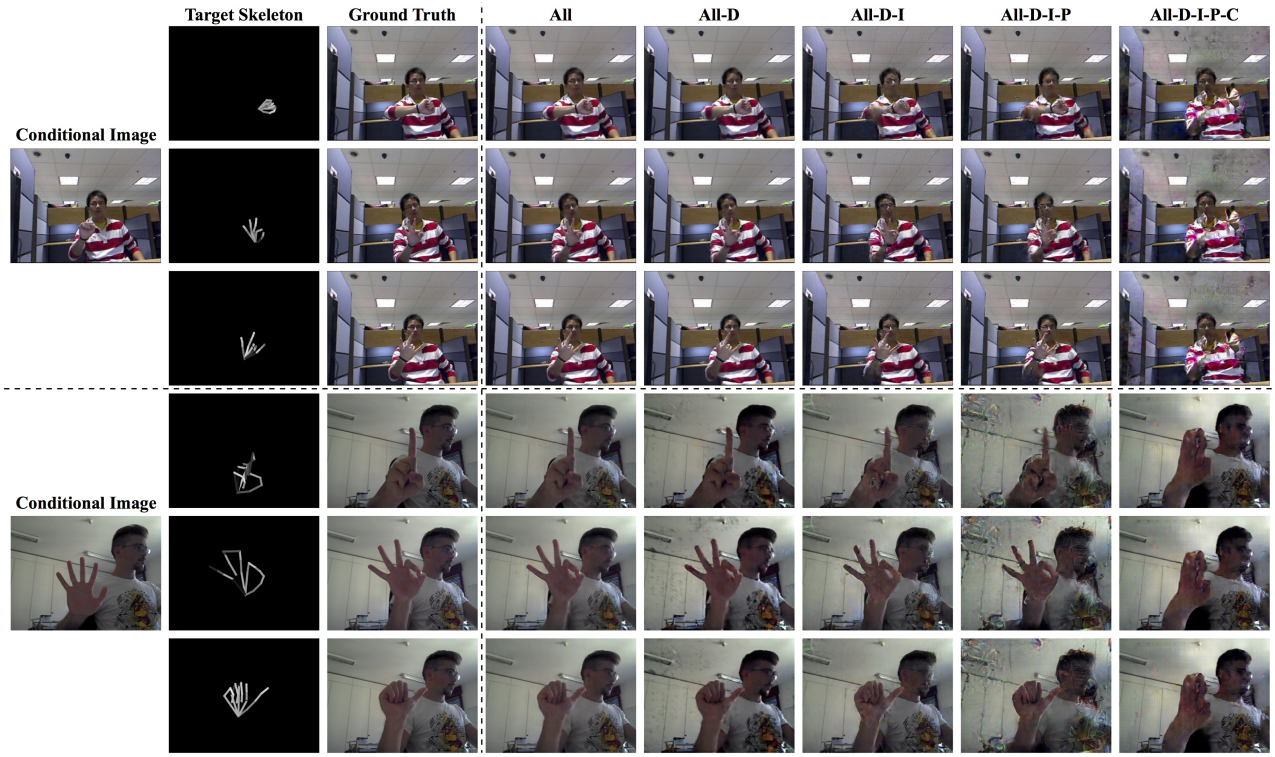


**Figure 4: Qualitative comparison using different components of GestureGAN on the NTU Hand Digit (Top) and the Senz3D (Bottom) datasets. All: full version of GestureGAN, D: Dual discriminators strategy, I: Identity preserving loss, P: Color loss, C: Cycle-consistency loss. "-" means removing. Zoom in for details.**

**Table 2: Ablation study: quantitative results with different components of GestureGAN on the NTU Hand Digit and Senz3D datasets. For PSNR and IS measures, higher is better. For MSE evaluation, lower is better. All: full version of GestureGAN, D: Dual discriminators strategy, I: Identity preserving loss, P: Color loss, C: Cycle-consistency loss. "-" means removing.**

| Component | NTU Hand Digit [22] | | | Senz3D [27] | | |
|---|---|---|---|---|---|---|
| | MSE | PSNR | IS | MSE | PSNR | IS |
| All | **105.7286** | **32.6091** | 2.5532 | **169.9219** | **27.9749** | 3.4107 |
| All - D (Dual Discriminators Strategy) | 118.7830 | 28.0189 | 2.5071 | 198.0646 | 26.7237 | 3.2740 |
| All - D - I (Identity Preserving Loss) | 198.7054 | 25.8474 | 2.5438 | 1319.3957 | 18.3892 | **4.0784** |
| All - D - I - P (Color Loss) | 406.1478 | 22.1564 | 2.5842 | 1745.3214 | 14.6598 | 3.4519 |
| All - D - I - P - C (Cycle-Consistency Loss) | 707.6053 | 20.2684 | **2.6114** | 2064.8428 | 15.5426 | 3.2064 |

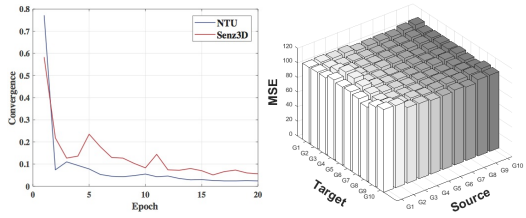


**Figure 5: Convergence loss $\mathcal{L}$ in Equation 15 (Left) and MSE of different gesture pairs on the NTU dataset (Right).**

substantially degrades results, meaning that the color loss and the cycle-consistency loss are critical to our results. In addition, the results without using the identity loss and the dual discriminators slightly degrade performance. We also provide quantitative results in Table 2, and we can see that the full version of GestureGAN produces more photo-realistic results that other variants on two measurements except IS. The reason could be that the datasets we used only include human images which do not fit into ImageNet classes [4]. Moreover, PG$^2$ [24] and other super-resolution works such as [11] also show the fact that sharper results have a lower quantitative value.

**User Study.** Similar to [24, 41, 62], we have also provided a user study. We follow the same settings as in [10] to conduct the Amazon Mechanical Turk (AMT) perceptual studies. The results of NTU Hand Digit [22] and Senz3D [27] datasets compared with the baseline models PG$^2$ [24], Ma et al [25], Yan et al. [56] and PoseGAN [41] are shown in Table 3. Note that the proposed GestureGAN consistently achieves the best performance compared with baselines.

**FID vs. FRD.** We also compare the performance between FID and the proposed FRD. The results shown in Table 4 and we can observe that FRD is more consistent with the human judgment in Table 3 than the FID metric. Moreover, we observe that the difference in

**Table 3: Comparison of AMT perceptual studies (%) on the NTU Hand Digit and Senz3D datasets.**

| Method | NTU Hand Digit [22] | Senz3D [27] |
|---|---|---|
| $PG^2$ [24] (NIPS 2017) | 3.5% | 2.8% |
| Yan et al. [56] (ACM MM 2017) | 2.6% | 2.3% |
| Ma et al. [25] (CVPR 2018) | 7.1% | 6.9% |
| PoseGAN [41] (CVPR 2018) | 9.3% | 8.6% |
| GestureGAN (Ours) | **26.1**% | **22.6**% |

**Table 4: Comparison of FID and the proposed FRD metrics on the NTU Hand Digit and Senz3D datasets. For both FID and FRD, lower is better.**

| Method | NTU Hand Digit [22] | | Senz3D [27] | |
|---|---|---|---|---|
| | FID | FRD | FID | FRD |
| $PG^2$ [24] (NIPS 2017) | 24.2093 | 2.6319 | 31.7333 | 3.0933 |
| Yan et al. [56] (ACM MM 2017) | 31.2841 | 2.7453 | 38.1758 | 3.1006 |
| Ma et al. [25] (CVPR 2018) | **6.7661** | 2.6184 | 26.2713 | 3.0846 |
| PoseGAN [41] (CVPR 2018) | 9.6725 | 2.5846 | 24.6712 | 3.0467 |
| GestureGAN (Ours) | 7.5860 | **2.5223** | **18.4595** | **2.9836** |

FRD between GestureGAN and the other methods is not as obvious as in the results from the user study in Table 3. The reason is that FRD calculates the Fréchet distance between the feature maps extracted from the real image and the generated image using CNNs which are trained with semantic labels. Thus, these feature maps are employed to reflect the semantic distance between the images. The semantic distance between the images is not very large considering they are all hands. On the contrary, the user study measures the generation quality from a perceptual level. The difference on the perceptual level is more obvious than on the semantic level, *i.e.*, the generated images with small artifacts show minor difference on the feature level, while being judged with a significant difference from the real images by humans.

**Data Augmentation.** The generated images are high-quality and are photo-realistic, and these images can be used to improve the performance of a hand gesture classifier. The intuition is that if the generated images are realistic, the classifiers trained on both the real images and the generated images will be able to boost the accuracy of the real images. In this situation, the generated images work as augmented data. We employ a pretrained ResNet-50 model [8] and feed the generated images to fine-tune it. For both datasets, we make a split of 70%/30% between training and testing sets. Specifically, the NTU Hand Digit dataset has 700 and 300 images for training and testing set. For the Senz3D dataset, the numbers of training and testing set are 924 and 396. The recognition results for the NTU Hand Digit and the Senz3D datasets are 15% and 34.34%, respectively. The term "real/real" in Table 5 represents the result without data augmentation. After adding the generated images by different methods to the training set, the performance improves significantly. Results compared with $PG^2$ [24], Yan et al. [56], Ma et al. [25] and PoseGAN [41] are shown in Table 5. Clearly, GestureGAN achieves the best result compared with baselines.

**Influence of Gesture Size and Distance.** We have also investigated the influence of the gesture size and the distance between source and target gestures. The training samples for the source and the target gesture from the same person are randomly paired and both gestures have different sizes and distances. Thus, the model is able to learn a robust translation *w.r.t* different hand size and distance. We show a qualitative example in Figure 6 in which the source images have different sizes and the target gestures have different locations. Note that GestureGAN can generate the target gesture from different hand sizes and distances with high quality.

**Table 5: Comparison of hand gesture recognition accuracy (%) on the NTU Hand Digit and Senz3D datasets.**

| Method | NTU Hand Digit [22] | Senz3D [27] |
|---|---|---|
| real/real | 15.000% | 34.343% |
| $PG^2$ [24] (NIPS 2017) | 93.667% | 98.737% |
| Yan et al. [56] (ACM MM 2017) | 95.333% | 99.495% |
| Ma et al. [25] (CVPR 2018) | 95.864% | 99.054% |
| PoseGAN [41] (CVPR 2018) | 96.128% | 99.549% |
| GestureGAN (Ours) | **96.667%** | **99.747%** |

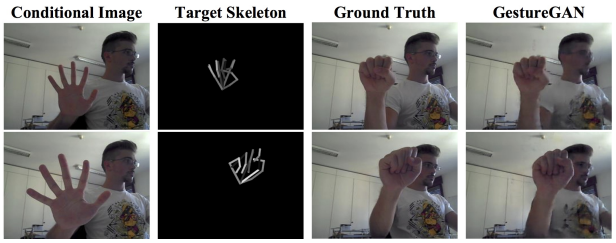


**Figure 6: Two samples with different hand sizes and distances.**

**Influence of Gesture Pairs.** To evaluate the influence of gesture pairs, we searched all the translations between every possible category combinations including the translation within each category. In Figure 5 (right), we show the MSE for the translation from the source to the target gesture types on NTU dataset. Note that the MSE for the generation of different gesture pairs has a small variance, showing that the influence of different gesture pairs is very low. This proves that our model is stable.

## 5 CONCLUSIONS

In this paper, we focus on a challenging task of hand gesture-to-gesture translation in the wild. To this end, we propose a novel Generative Adversarial Network (GAN), *i.e.*, GestureGAN, which can generate hand gestures with different poses, sizes and locations in the wild. We also propose two novel losses to learn the mapping from the source gesture to the target gesture, *i.e.*, the color loss and the cycle-consistency loss. It is worth noting that the proposed color loss handles the "channel pollution" problem while back-propagating the gradients, which frequently occurs in the existing generative models. In addition, we present the Fréchet ResNet Distance (FRD) metric to evaluate the quality of generated images. Experimental results show that GestureGAN achieves state-of-the-art performance. Lastly, the generated images of GestureGAN are of high-quality and are photo-realistic, and they can thus be used to improve the performance of hand gesture classifiers. Future work will focus on designing a GAN model which can handle the situation where the background is total different between source and target gestures.

## ACKNOWLEDGEMENTS

We want to thank the Nvidia Corporation for the donation of the TITAN Xp GPUs used in this work.

# REFERENCES


[1] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. 2018. ComboGAN: Unrestrained Scalability for Image Domain Translation. In *CVPR Workshops*.

[2] David Berthelot, Tom Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[5] Xing Di, Vishwanath A Sindagi, and Vishal M Patel. 2018. GP-GAN: gender preserving GAN for synthesizing faces from landmarks. In *ICPR*.

[6] Brian Dolhansky and Cristian Canton Ferrer. 2018. Eye In-Painting with Exemplar Generative Adversarial Networks. In *CVPR*.

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

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NIPS*.

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

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

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.

[13] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.

[14] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[15] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast face-swap using convolutional neural networks. In *ICCV*.

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

[17] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*.

[18] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual generative adversarial networks for small object detection. In *CVPR*.

[19] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative Face Completion. In *CVPR*.

[20] Si Liu, Yao Sun, Defa Zhu, Renda Bao, Wei Wang, Xiangbo Shu, and Shuicheng Yan. 2017. Face aging with contextual generative adversarial nets. In *ACM MM*.

[21] Si Liu, Yao Sun, Defa Zhu, Guanghui Ren, Yu Chen, Jiashi Feng, and Jizhong Han. 2018. Cross-domain human parsing via adversarial feature and label adaptation. In *AAAI*.

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*.

[23] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. 2016. Semantic segmentation using adversarial networks. In *NIPS Workshops*.

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose Guided Person Image Generation. In *NIPS*.

[25] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled Person Image Generation. In *CVPR*.

[26] Michael Mathieu, Camille Couprie, and Yann LeCun. 2016. Deep multi-scale video prediction beyond mean square error. *ICLR* (2016).

[27] Alvise Memo and Pietro Zanuttigh. 2016. Head-mounted gesture controlled interface for human-computer interaction. *Springer Multimedia Tools and Applications* (2016), 1–27.

[28] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[29] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. 2017. Dual discriminator generative adversarial nets. In *NIPS*.

[30] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *SSW*.

[31] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. 2017. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*.

[32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *CVPR*.

[33] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. 2016. Invertible Conditional GANs for image editing. In *NIPS Workshops*.

[34] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. 2018. Geometry-Contrastive Generative Adversarial Network for Facial Expression Synthesis. *arXiv preprint arXiv:1802.01822* (2018).

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *ICML*.

[36] Scott Reed, Aäron van den Oord, Nal Kalchbrenner, Victor Bapst, Matt Botvinick, and Nando de Freitas. 2016. Generating interpretable images with controllable structure. *Technical Report* (2016).

[37] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *NIPS*.

[38] Zhou Ren, Junsong Yuan, Jingjing Meng, and Zhengyou Zhang. 2013. Robust part-based hand gesture recognition using kinect sensor. *IEEE TMM* 15, 5 (2013), 1110–1120.

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *NIPS*.

[40] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017. Neural Face Editing with Intrinsic Image Disentangling. In *CVPR*.

[41] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. 2018. Deformable GANs for Pose-based Human Image Generation. In *CVPR*.

[42] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[43] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[44] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. 2017. Geometry Guided Adversarial Facial Expression Synthesis. *arXiv preprint arXiv:1712.03474* (2017).

[45] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Natural and Effective Obfuscation by Head Inpainting. In *CVPR*.

[46] Yaniv Taigman, Adam Polyak, and Lior Wolf. 2017. Unsupervised cross-domain image generation. In *ICLR*.

[47] Hao Tang and Hong Liu. 2016. A Novel Feature Matching Strategy for Large Scale Image Retrieval.. In *IJCAI*.

[48] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *CVPR*.

[49] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*.

[50] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*.

[51] Wang Wei, Alameda-Pineda Xavier, Xu Dan, Ricci Elisa, Fua Pascal, and Sebe Nicu. 2018. Every Smile is Unique: Landmark-Guided Diverse Smile Generation. In *CVPR*.

[52] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2017. Gp-gan: Towards realistic high-resolution image blending. *arXiv preprint arXiv:1703.07195* (2017).

[53] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*.

[54] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable Invariance through Adversarial Feature Learning. In *NIPS*.

[55] Runze Xu, Zhiming Zhou, Weinan Zhang, and Yong Yu. 2017. Face Transfer with Generative Adversarial Network. *arXiv preprint arXiv:1710.06090* (2017).

[56] Yichao Yan, Jingwei Xu, Bingbing Ni, Wendong Zhang, and Xiaokang Yang. 2017. Skeleton-aided Articulated Motion Generation. In *ACM MM*.

[57] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation. In *ISMIR*.

[58] Zili Yi, Hao Zhang, Ping Tan Gong, et al. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *ICCV*.

[59] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In *ICCV*.

[60] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. 2014. Packing and padding: Coupled multi-index for accurate image retrieval. In *CVPR*.

[61] Liang Zheng, Shengjin Wang, Wengang Zhou, and Qi Tian. 2014. Bayes merging of multiple vocabularies for scalable image retrieval. In *CVPR*.

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

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *NIPS*.