Gesture-to-Gesture Translation in the Wild via Category-Independent Conditional Maps

Yahui Liu
yahui.liu@unitn.it
University of Trento, Trento, Italy

Marco De Nadai
denadai@fbk.eu
FBK, Trento, Italy

Gloria Zen
gloria.zen@unitn.it
University of Trento, Trento, Italy

Nicu Sebe
niculae.sebe@unitn.it
University of Trento, Trento, Italy

Bruno Lepri
lepri@fbk.eu
FBK, Trento, Italy

ABSTRACT

Recent works have shown Generative Adversarial Networks (GANs) to be particularly effective in image-to-image translations. However, in tasks such as body pose and hand gesture translation, existing methods usually require precise annotations, e.g. key-points or skeletons, which are time-consuming to draw. In this work, we propose a novel GAN architecture that decouples the required annotations into a category label - that specifies the gesture type - and a simple-to-draw category-independent conditional map - that expresses the location, rotation and size of the hand gesture. Our architecture synthesizes the target gesture while preserving the background context, thus effectively dealing with gesture translation *in the wild*. To this aim, we use an attention module and a rolling guidance approach, which loops the generated images back into the network and produces higher quality images compared to competing works. Thus, our GAN learns to generate new images from simple annotations without requiring key-points or skeleton labels. Results on two public datasets show that our method outperforms state of the art approaches both quantitatively and qualitatively. To the best of our knowledge, no work so far has addressed the gesture-to-gesture translation *in the wild* by requiring user-friendly annotations.

CCS CONCEPTS

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

KEYWORDS

GANs, image translation, hand gesture

ACM Reference Format:
Yahui Liu, Marco De Nadai, Gloria Zen, Nicu Sebe, and Bruno Lepri. 2019. Gesture-to-Gesture Translation in the Wild via Category-Independent Conditional Maps. In *Proceedings of the 27th ACM Int'l Conf. on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351020

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351020

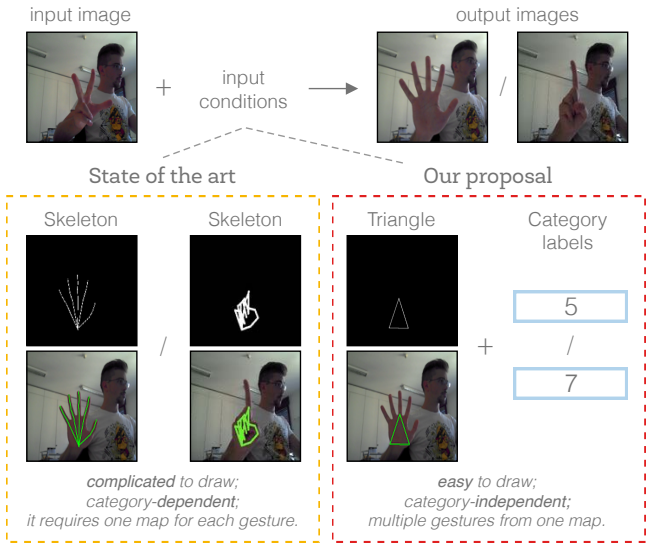

Figure 1: Our proposal decouples the *category label* that specifies the gesture type (e.g., gesture "5" or "7") from the *conditional map* (i.e. triangle) that controls the location, orientation and size of the target gesture. Existing works require a detailed conditional map (e.g. skeleton) that is gesture-dependent. In this example, we show that our method significantly lowers the drawing effort and expertise required by users. Our method can generate multiple output images with the same map for multiple gesture categories.

1 INTRODUCTION

Photo-editing software, fashion, and retail markets would enormously benefit from the possibility of modifying an image through a simple user input describing the changes to make (e.g. change the hand gesture of the person in the picture from "open hand" to "ok"). However, despite the significant advances of Generative Adversarial Networks (GANs) [1, 9, 10, 24], the generation of images *in the wild* without precise annotations (e.g. hand skeleton) is still an open problem. Previous literature on image-to-image translation has relied either on pixel-to-pixel mappings [13, 37, 42] or precise annotations to localize the instance to be manipulated, such as segmentation masks [21], key-points [30], skeletons [34] and facial landmarks [27]. However, obtaining such annotations is not trivial. On one hand, automatic methods for key-points extraction [5, 31] may fail or the reference gesture image/video [29] may be not available. On the other, drawing such annotations is

complicated, time-consuming, and their quality directly affects the performance of the network.

Moreover, existing methods often focus on foreground content, e.g., the target gesture or facial expression, generating blurred and imprecise backgrounds [18, 25, 30]. These methods are well suited in the cases where share fixed or similar spatial structures, such as facial expression datasets [7, 17]. Instead, in image-to-image translations *in the wild*, both the foreground and background between the source image and target image can vary a lot [34].

In this paper, we propose a novel method, named as $\triangle$-GAN, that requires a simple-to-draw annotation, such as a triangle, and focuses on the challenging task of hand gesture-to-gesture translation *in the wild*. In general, annotations such as key-points or skeletons are category-dependent since they provide four types of information at the same time, namely *category* (e.g., gesture "5"), *location*, *scale* and *orientation* of the hand gesture. Instead, our $\triangle$-GAN decouples the *category* from the *location-scale-orientation* information. Using a category-independent conditional map significantly lowers the annotation cost, allowing to generate multiple target images while requiring users to draw only a single map. In this work, we refer to "annotations" as the user effort to draw the desired target gesture also at deploy time, besides the effort needed for generating the training data for our model. The intuition of our approach is depicted in Figure 1. Furthermore, we propose a novel architecture that uses an attention module and a rolling guidance approach to perform gesture-to-gesture translation *in the wild*. Our research yields three main contributions:

- *Decoupled conditional map and category.* We designed a general architecture for gesture-to-gesture translations that separately encodes the category label (e.g., gesture "5") and the category-independent conditional map. This allows to perform several image translations with the same conditional map.
- *Rolling guidance and attention mask.* We propose a novel rolling guidance approach that allows generating higher quality output images by feeding the generated image back to the input as an additional condition to be considered. Also, $\triangle$-GAN learns unsupervisedly an attention mask that allows to preserve the details shared between the input and target image.
- *Simple-to-draw conditional maps.* We propose a triangle conditional map as the simplest and minimal necessary user provided condition for gesture-to-gesture translation. To the best of our knowledge, no work so far has addressed the gesture-to-gesture translation task *in the wild* by requiring user-friendly annotations. Furthermore, we assess the performance of our method with different shapes, such as boundary and skeleton maps. Finally, we enrich two public datasets with different conditional maps for each gesture image, specifically based on triangles and boundaries.

2 RELATED WORK

Recently, there have been a lot of works on Generative Adversarial Networks (GANs) and, particularly, on conditional GANs for the task of image-to-image translation [6, 13, 23, 42, 43]. A significant line of these works have focused on translation tasks where the the input and target images are spatially aligned, as in the case of style transfer [13, 14, 42], emotional content transfer [40] or image inpaiting [38]. In general, these works aim to preserve the main image content while presenting them in various styles. Another line of works have tackled the more challenging task of image translation where the target object is spatially unaligned with respect to the original location, shape or size of the input original object. This is the case of image-to-image translation tasks like facial expression [27, 41], human pose [18, 30] or hand gesture translation [34]. To this aim, methods usually require geometry information as guidance to guarantee where and how to edit visual patterns corresponding to the image attributes, such as key-points [18, 19, 29], skeletons [34], object segmentation masks [21], facial landmarks [27], action units [23] or 3D models [41]. GANimation [23] learns an attention mask to preserve the background content of the source image for the spatially aligned translation of facial expressions. InstaGAN [21] performs multi-instance domain-to-domain image translation by requiring as input precise segmentation masks for the target objects. Pose Guided Person Generation Network (PG^2) [18] proposes a two stages generating framework to refine the output image given a reference image and a target pose. MonkeyNet [29] generates a conditioned video transferring body movements from a target driving video to a reference appearance image. The target key-points are obtained automatically using state-of-the-art detectors. Relatively few works have considered the challenging task of image translation *in the wild*, where both the foreground content and the background context undergoes significant variation [34]. In these cases, not only the networks learn to synthesize the target object or attribute, but they also have to correctly locate it in the image while the pixels depicting the content are preserved from the original image. GestureGAN [34] proposes a novel color loss to improve the output image quality and to produce sharper results. However, this approach does not aim to separately encode the foreground content and the background context.

Furthermore, the large majority of image-to-image translations works focus on one to one domain mapping. Recently, efficient solutions have been proposed to address the multi-domain translations [6, 41]. In particular, StarGAN [6] proposes the use of multiple mask vectors to generalize on different datasets and domain labels. 3DMM [41] decomposes an image into shape and texture spaces, and relies on an identity and target coefficients to generate images in multiple domains. They, however, focus on multi-domain face attributes where neither the background quality or the misalignment are considered.

Our work falls within the category of multi-domain and image-to-image translation *in the wild*. Still, rather than asking expensive annotation to the user, we propose a novel user-friendly annotation strategy. To the best of our knowledge, no works so far have investigated other approaches for gesture translation that also allow to reduce the annotation effort.

3 OUR APPROACH

The goal of this work is to perform gesture translation *in the wild*, conditioned on the user provided gesture attributes, such as the desired category, location, scale and orientation. Specifically, we seek to estimate the mapping $\mathcal{M}$: $(\mathbf{I}_X, \mathbf{S}_Y, C_Y) \rightarrow \mathbf{I}_Y$ that translates an input image $\mathbf{I}_X$ into an output image $\mathbf{I}_Y$, conditioned to a hand gesture category C_Y and a simple-to-draw conditional map $\mathbf{S}_Y$, which encodes the desired location, scale and orientation of the gesture.

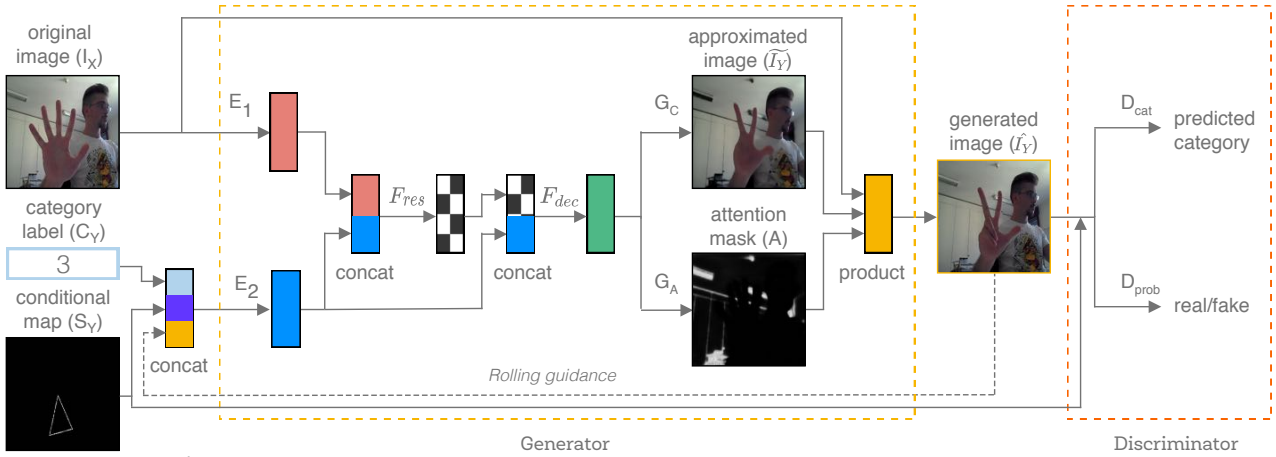

Figure 2: Overview of Δ-GAN, that allows to translate hand gestures *in the wild* by separately encoding the input image and the target gesture attributes including the category label (e.g., gesture 3) and the category-independent conditional map. We include an unsupervised attention mask to preserve the details shared between the input image and the target image. Specially, we feed the first reconstructed image back to the input conditions encoding module to improve quality of output image.

The generated image is discriminated through a Conditional Critic discriminator [23] that judges the photo-realism of the generated image and assess the category of the generated gesture (e.g. gesture "5"). Furthermore, in order to efficiently deal with the gesture translation *in the wild*, we propose a rolling guidance approach and an attention mask module. Figure 2 depicts the architecture of our approach. In this Section we further explain the details of the architecture and the loss functions used to train our framework.

3.1 Network Architecture

Generator. Our Generator G takes as input the conditional image $\mathbf{I}_X$, the category label C_Y and the conditional map $\mathbf{S}_Y$ and it outputs the translated image $\mathbf{I}_Y$. The original image and the target gesture conditions are encoded respectively through the encoder E_1 and E_2, and then concatenated and provided to F_{res}. Then, the output features of E_2 and F_{res} are concatenated and provided to F_{dec}. Similarly to [23], in the decoder we learn to predict in an unsupervised manner the approximated image $\widetilde{\mathbf{I}}_Y$ and an *Attention Mask* $\mathbf{A} \in [0, 1.0]^{H \times W}$. Since both the background and foreground between the input and target images can vary a lot, the learned attention mask $\mathbf{A}$ tends to preserve the shared part between $\mathbf{I}_X$ and $\mathbf{I}_Y$. Pixels in $\mathbf{A}$ with higher value indicate to preserve the corresponding pixels from $\mathbf{I}_X$, otherwise from $\widetilde{\mathbf{I}}_Y$. The final generated image is thus obtained as:

$$\widehat{\mathbf{I}}_Y = \mathbf{A} * \mathbf{I}_X + (1 - \mathbf{A}) * \widetilde{\mathbf{I}}_Y \tag{1}$$

Furthermore, the generated image $\mathbf{I}_Y$ is rolled back as an input condition and concatenated together with the category label C_Y and the conditional map $\mathbf{S}_Y$. More details are provided in Section 3.3.

Discriminator. Our Discriminator D takes as input the generated image $\widehat{\mathbf{I}}_Y$ and its conditional map $\mathbf{S}_Y$. The outputs of D consist of two parts: D_{cat} predicts the category label and D_{prob} classifies whether local image patches are real or fake. As in [23], D_{prob} is based on PatchGANs [13]. The conditional map $\mathbf{S}_Y$ is needed by D to verify that the generated gesture has also the right location, scale and orientation.

We refer to the Supplementary Material for additional details in the architecture.

3.2 Objective Formulation

The loss function of Δ-GAN is composed of four main components, namely *GAN Loss* that pushes the generated image distribution to the distribution of source images; *Reconstruction loss* that forces the generator to reconstruct the source and target image; *Category Label loss* that allows to properly classify the generated image into hand gesture classes; and *Total Variation loss* that indirectly learns the attention mask in an unsupervised fashion.

GAN Loss. To generate images with the same distribution of the source images, we adopt an adversarial loss:

$$\begin{aligned}\mathcal{L}_{GAN} =& \mathbb{E}_{\mathbf{I}_Y, \mathbf{S}_Y \sim \mathbb{P}}[\log D_{prob}(\mathbf{I}_Y, \mathbf{S}_Y)] + \\ & \mathbb{E}_{\mathbf{I}_X, \mathbf{S}_Y, C_Y \sim \mathbb{P}}[\log(1 - D_{prob}(G(\mathbf{I}_X, \mathbf{S}_Y, C_Y), \mathbf{S}_Y))]\end{aligned} \tag{2}$$

where $\mathbb{P}$ is the data distribution of the hand gesture images in the dataset, and where G generates an image $G(\mathbf{I}_X, \mathbf{S}_Y, C_Y)$ conditioned on both the input image $\mathbf{I}_X$, the conditional map $\mathbf{S}_Y$, and the target category C_Y, while D tries to distinguish between real and fake images. In this paper, we refer to the term $D_{prob}(\mathbf{I}, \mathbf{S})$ as a probability distribution over sources given by D.

Reconstruction Loss. The adversarial loss does not guarantee that the generated image is consistent with both the target conditional map $\mathbf{S}$ and category C. Thus, we first apply a *forward reconstruction loss* that ties together the target image $\mathbf{I}_Y$ with its target conditional map $\mathbf{S}_Y$ and category C_Y:

$$\mathcal{L}_{rec} = \|G(\mathbf{I}_X, \mathbf{S}_Y, C_Y) - \mathbf{I}_Y\|_1 \tag{3}$$

Then, instead of using perceptual features (e.g. extracted from VGG [32] networks) to force the model to reconstruct the source image, we propose a simplified *self-reconstruction (identity) loss*:

$$\mathcal{L}_{idt} = \|G(\mathbf{I}_X, \mathbf{S}_X, C_X) - \mathbf{I}_X\|_1 \tag{4}$$

where $\mathbf{S}_X$ is the conditional map of the source image and C_X the category label of the source image. Finally, we apply the *cycle*

consistency loss [15, 42] to reconstruct the original image from the generated one:

$$\mathcal{L}_{cyc} = \|G(G(\mathbf{I}_X, \mathbf{S}_Y, C_Y), \mathbf{S}_X, C_X) - \mathbf{I}_X\|_1 \quad (5)$$

Note that we apply the cycle reconstruction loss only in one direction to reduce computation, i.e., A-to-B-to-A, since a translation pair based on two images A and B may be sampled either as A-to-B or as B-to-A during the training.

Category Label loss. We enforce the generator to render realistic samples that have to be correctly classified to the hand gesture expressed by the input category label. Similarly to StarGAN [6], we split the *Category Label loss* in two terms: a gesture classification loss of the real image $\mathbf{I}_Y$ used to optimize D, and a gesture classification loss of the generated image $\hat{\mathbf{I}}_Y$, used to optimize G. Specifically:

$$\mathcal{L}_{cls} = \mathbb{E}_{\mathbf{I}_Y, C_Y}[-\log D_{cat}(C_Y|\mathbf{I}_Y, \mathbf{S}_Y)] \quad (6)$$

$$\mathcal{L}_{c\hat{l}s} = \mathbb{E}_{\hat{\mathbf{I}}_Y, C_Y}[-\log D_{cat}(C_Y|\hat{\mathbf{I}}_Y, \mathbf{S}_Y)] \quad (7)$$

where $D_{cat}(C_Y|\mathbf{I}_Y, \mathbf{S}_Y)$ and $D_{cat}(C_Y|\hat{\mathbf{I}}_Y, \mathbf{S}_Y)$ represent a probability distribution over the categories of hand gestures respectively in the real and generated images. In other words, these losses allow to generate images that can be correctly classified as the target hand gesture category.

Total Variation loss. To prevent the final generated image having artifacts, we use a Total Variation Regularization, f_{tv}, as in GANimation [23]. However, differently from them, we calculate f_{tv} over the approximated image $\widetilde{\mathbf{I}}_Y$ instead of the attention mask $\mathbf{A}$, thus allowing to freely explore the shared pixels between the source and target images. The total variation loss is applied both to the *forward reconstruction* and *self-reconstruction* and is formulated as:

$$\mathcal{L}_{tv} = f_{tv}(G_C(\mathbf{I}_X, \mathbf{S}_X, C_X) + f_{tv}(G_C(\mathbf{I}_X, \mathbf{S}_Y, C_Y)) \quad (8)$$

The total variation regularization f_{tv} is defined as:

$$f_{tv}(\mathbf{I}) = \mathbb{E}_{\mathbf{I}}\left[\sum_{i,j}^{H-1,W-1} [(\mathbf{I}_{i+1,j} - \mathbf{I}_{i,j})^2 + (\mathbf{I}_{i,j+1} - \mathbf{I}_{i,j})^2]\right] \quad (9)$$

where $\mathbf{I}_{i,j}$ is the entry i, j of the image matrix $\mathbf{I}$.

Total loss. The final objective function used to optimize G and D is formulated as follows:

$$\mathcal{L}_D = \lambda_D \mathcal{L}_{GAN} + \lambda_{cls}\mathcal{L}_{cls} \quad (10)$$

$$\begin{aligned}\mathcal{L}_G = \lambda_G \mathcal{L}_{GAN} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{idt}\mathcal{L}_{idt} \\ + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{cls}\mathcal{L}_{c\hat{l}s} + \lambda_{tv}\mathcal{L}_{tv}\end{aligned} \quad (11)$$

where $\lambda_D, \lambda_G, \lambda_{rec}, \lambda_{idt}, \lambda_{cyc}, \lambda_{cls}$, and λ_{tv} are hyper-parameters that control the relative importance of each loss term.

3.3 Rolling Guidance

While the total variation loss $\mathcal{L}_{tv}$ also enforces the approximated images $\widetilde{\mathbf{I}}_Y$ to be smooth, the source and target images might contain edges and details that have to be preserved. Moreover, the E_1 and E_2 encoders mostly focus on the gesture, failing to learn important details of the context, which might result in blurred images. Inspired by previous works [18, 22, 39], we propose a Rolling Guidance approach to refine the generated image in a two-stage process. First, the network generates an initial version $\widehat{\mathbf{I}}_Y$ from input ($\mathbf{I}_X$, $\mathbf{S}_Y$, C_Y). Then, $\widehat{\mathbf{I}}_Y$ is fed back to E_2. Thus, the network generates a refined version of $\widehat{\mathbf{I}}_Y$ from input ($\mathbf{I}_X$, $\mathbf{S}_Y$, C_Y, $\widehat{\mathbf{I}}_Y$). Note that there exists some approaches, like PG^2 [18], feed the initial generated image back and concatenated to the source input to learn difference map to refine the results. However, images with a considerable variation in both the foreground and background between source and target images might result in an ill-posed problem. Gesture-to-gesture translation in the wild shows such an issue. For this reason, in $\triangle$-GAN, we feedback the generated image to E_2 to refine the generated image and to learn the condition related features of the target gesture at the same time. This results in better generalization and significantly improves the results as well.

4 EXPERIMENTS

We compare our model with the state of the art techniques on two hand gesture datasets. First, we evaluate our model quantitatively through various widely used metrics that compare the generated image with the ground truth. Then, we also evaluate our results qualitatively through a perceptual user study. We released the resulted dataset and annotations, source code and trained models are available at: https://github.com/yhlleo/TriangleGAN.

4.1 Datasets

The NTU Hand Gesture dataset [26] is a collection of 1,000 RGB-D images recorded with a Kinect sensor. It includes 10 gestures repeated 10 times by 10 subjects under a cluttered background. Image resolution is 640x480. The Creative Senz3D dataset [20] is a collection of 1,320 images, where 11 gestures are repeated 30 times by 4 subjects. Image resolution is 640x480.

Experimental setting. We consider two dataset setups in our experimental evaluation.

Normal uses the same setup as GestureGAN [34], in order to directly compare with the state of the art. GestureGAN authors used only a subset of the datasets (acquired by OpenPose [4]): 647 images out of 1,000 for NTU Hand Gesture and 494 out of 1,320 for Creative Senz3d. It shows that the state-of-the-art detector OpenPose [4] fails at detecting key-points for about 50% of the images. The resulting number of training and test data pairs obtained are respectively the following: 21,153 and 8,087 for NTU Hand Gesture; 30,704 and 11,234 for Creative Senz3D. These numbers are different from those reported in [34] since we here report only the unique pairs without considering flipping and A-to-B reverse ordering.

Challenging pushes the limits of our model by ensuring that all the translation pairs "A-to-B" to a specific image "B" are included either in the train or in the test, a condition not ensured by the *normal* setting and, thus, by state of the art. As a consequence, the model here generates multi-domain images without previous knowledge about it. Out of the possible translation pairs, we randomly select for the training and test data the following number of pairs: 22,050 and 13,500 for NTU Hand Gesture; 138,864 and 16,500 for the Creative Senz3D.

Conditional Maps. We consider three possible shapes of the conditional map to prove the effectiveness and generality of our method. Sample images are reported in Figure 3.

Triangle Map. In this type of annotation, the user has to provide an oriented triangle which outlines the size, base and orientation of the hand palm. This conditional map is easy to draw, as it is possible to provide a simple interface where users can draw a triangle simply specifying the three delimiting points of the triangle, plus its base. Moreover, the triangle conditional map is category independent, as it does not contain any information about the gesture. We annotated all images for both datasets with the corresponding triangle conditional maps.

Boundary Map. In the boundary map annotation, the user has to draw the contour of the desired target gesture. This type of annotation is weakly category dependent, since from the conditional map (see Figure 3 in the center) it may be possible to infer the target gesture category. This shape is considered in our experiments since it may be a valid alternative to the skeleton and triangle map, while still being a user-friendly shape to draw. We annotated all 1,320 images for the NTU Hand dataset with the corresponding boundary annotations.

Skeleton Map. In the skeleton map, the user is required to draw either a complicated skeleton of the hand gesture or the exact position of the hand gesture key-points. However, when the target gesture image is available, it is sometimes possible to draw them automatically. As in [34], we obtain the skeleton conditional maps by connecting the key-points obtained through OpenPose, a state of the art hand key-points detector [4, 31]. In the case of *normal* experimental setting, the hand pose is detected for all the 647 and 494 images of the two datasets. Instead, in the case of *challenging* experimental setting, the key-points could not be obtained for over half of the image set. To this reason, the skeleton map is considered only in the *normal* experimental setting. This conditional map is hard to draw, and strongly dependent on the category of hand gesture.

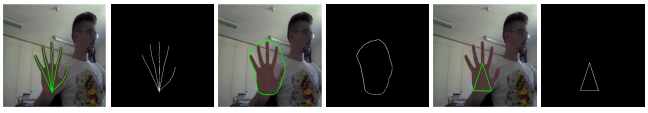

Figure 3: The three considered shapes of the conditional map, sorted by user drawing effort: from (left) the most difficult to (right) the easiest to draw.

4.2 Evaluation

Baseline models. As our baseline models we select the state of the art for hand gesture translation *in the wild* GestureGAN [34]. Also, we adopt StarGAN [6], GANimation [23], and PG^2 [18] as they showed impressive results on multi-domain image-to-image translation. Both StarGAN and GANimation learn to use attribute vectors to transfer facial images from one expression to another one. GestureGAN learns to transfer hand gestures via category-dependent skeleton maps.

Evaluation metrics. We quantitatively evaluate our method performance using two metrics that measure the quality of generated images, namely Peak Signal-to-Noise Ratio (PSNR) and Fréchet Inception Distance (FID) [12], and the F1-score, which measures whether the generated images depict a consistent category label.

Table 1: Quantitative comparison for the gesture-to-gesture translation task in *normal* experimental setting, using the same evaluation metrics as GestureGAN [34].

Model	NTU Hand Gesture [26]			Creative Senz3D [20]		
	MSE	PSNR	IS	MSE	PSNR	IS
PG^2 [18]	116.10	28.24	2.42	199.44	26.51	3.37
Yan *et al.* [36]	118.12	28.02	2.49	175.86	26.95	3.33
Ma *et al.* [19]	113.78	30.65	2.45	183.65	26.95	3.38
PoseGAN *et al.* [30]	113.65	29.55	2.40	176.35	27.30	3.21
GestureGAN [34]	105.73	32.61	**2.55**	169.92	27.97	**3.41**
Δ-GAN - no rolling	31.80	33.57	1.92	46.79	31.98	2.31
Δ-GAN	**15.76**	**36.51**	2.00	**21.73**	**35.39**	2.34

Moreover, to be comparable with GestureGAN [34], we employ the Mean Squared Error (MSE) between pixels and the Inception Score (IS) [10]. However, these two metrics were indicated as highly unstable [2, 3] and they are not strictly related to the quality of generated images. To this reason, we report their results on a separate table and do not further discuss them.

PSNR. It compares two images through their MSE and the maximal pixel intensity ($MAX_I = 255$). It is defined as: $PSNR = 20 \log_{10}(\frac{MAX_I}{\sqrt{MSE}})$.

FID. It is defined as the distance between two Gaussian with mean and covariance (μ_x, Σ_x) and (μ_y, Σ_y): $FID(x, y) = ||\mu_x - \mu_y||_2^2 + Tr(\Sigma_x + \Sigma_y - 2(\Sigma_x \Sigma_y)^{1/2})$, where the two Gaussians are defined in the feature space of the Inception model.

F1. The F1-score for binary classifiers is defined as $F_1 = (2pr)/(p + r)$ where p and r are the precision and recall. For multi-class classifiers it can be defined as the sum of the F1-scores of each class, weighted by the percentage of labels belonging to the class. The resulting measure ranges from 0 to 1, where 1 means that all the classes are correctly classified and 0 the opposite. Thus, we use the F1-score to evaluate the category consistence of our generated images. To compute the F1-score, we train a network to recognize hand gestures. Further details are provided in the Implementation Details.

Perceptual User study. We run a "fake" vs "real" perceptual user study following the same protocol as [34, 37]. Users are shown a pair of images, the original and the generated image, and they are asked to select the fake one. The images are displayed for only 1 second before the user can provide his or her choice. The image pairs are randomly shuffled in order to avoid introducing bias in the results. Overall, we collected perceptual annotations from 12 users and each user was asked to vote for 98 image comparisons. Specifically, 12 image translation pairs were selected for each dataset and experimental setting.

Implementation Details. Inspired by previous methods [6, 42], both E_1 and E_2 are composed of two convolutional layers with the stride size of two for downsampling, F_{res} refers to six residual blocks [11], and F_{dec} is composed of two transposed convolutional layers with the stride size of two for upsampling. We train our model using Adam [16] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and batch size 4. We use an n-dimensional one-hot vector to represent the category label ($n = 10$ for NTU dataset and $n = 11$ for Senz3D Dataset). For data augmentation we flip the images horizontally with a probability of 0.5 and we reverse the "A-to-B" direction with a probability of 0.5. The initial learning rate is set to 0.0002. We train for 20 epochs

Table 2: Quantitative results for the gesture-to-gesture translation task for the two experimental settings.

Model	Experimental setting	Easy to draw map	NTU Hand Gesture [26]			Creative Senz3D [20]		
			PSNR	FID	F1	PSNR	FID	F1
GANimation [23]	*Normal*	×	10.03	440.45	0.08	11.11	402.99	0.23
StarGAN [6]			17.79	98.84	0.09	11.44	137.88	0.07
PG^2 [30]			21.71	66.83	0.15	21.78	122.08	0.68
GestureGAN [34]			34.24	15.06	**0.93**	28.65	54.00	0.96
△-GAN - no rolling			34.07	20.29	0.90	31.98	58.28	0.88
△-GAN			**36.51**	**13.73**	**0.93**	**35.39**	**32.04**	**0.99**
PG^2 [18]	*Challenging*	✓	21.94	135.32	0.10	18.30	265.73	0.10
GestureGAN [34]			25.60	94.10	0.38	18.69	254.37	0.34
GestureGAN†			27.32	59.79	0.43	22.24	192.77	0.38
△-GAN - no rolling			27.14	61.14	0.43	22.77	134.54	0.33
△-GAN			**28.11**	**47.88**	**0.61**	**23.11**	**101.22**	**0.58**

and linearly decay the rate to zero over the last 10 epochs. To reduce model oscillation [8], we follow previous works [28, 42] and update the discriminators using a history of generated images rather than the ones produced by the latest generators. We use instance normalization [35] for the generator G. For all the experiments, the weight coefficients for the loss term in Eq. 10 and Eq. 11 are set to $\lambda_D = 1$, $\lambda_G = 2$, $\lambda_{cls} = 1$, $\lambda_{rec} = 100$, $\lambda_{idt} = 10$, $\lambda_{cyc} = 10$ and $\lambda_{tv} = 1e-5$. Baseline models are optimized using the same settings described in the respective articles. We used the source code released by the authors for all competing works, except for GestureGAN, which was implemented from scratch following the description of the original article [34]. △-GAN is implemented using the deep learning framework PyTorch.

To compute the F1-score, we train a network on hand gesture recognition using Inception v3 [33] network fine tuned on the NTU Hand Gesture and Creative Senz3D datasets. The network achieves F1-score 0.93 and 0.99 on Creative Senz3D and NTU Hand Gesture test sets, respectively. Additional details on the training can be found in the supplementary materials.

5 RESULTS

Quantitative Results. We begin by directly comparing △-GAN with the same experimental protocol and metrics used by our most similar competitor, GestureGAN [34]. Table 1 shows that △-GAN performs better than all competing works in terms of MSE and PSNR, especially when the rolling guidance is employed. In terms of IS GestureGAN performs better. However, the MSE and the IS are not directly related to the quality of the generated images. The MSE is indeed a metric of pixel difference, while the low reliability of the Inception Score is well known [2].

For this reason, we compare our method with competing works using the PSNR, F1-score and the FID score, both in *normal* and *challenging* experimental settings. These metrics compare the diversity, quality and hand-gesture consistency of the generated images. Table 2 shows that △-GAN outperforms all competing works, in both the experimental settings and for all the metrics. In the *normal* setting, compared to GestureGAN, we achieve higher PSNR (36.51 vs 34.24 and 35.39 vs 28.65), F1-score (0.99 vs 0.96) and lower FID (13.73 vs 15.06 and 32.04 vs 54.00) in both datasets. Other methods perform particularly poor in terms of F1-score. For example, GANimation and StarGAN perform near randomness in the F1-score (random baseline $\sim$ 0.09), while △-GAN achieves near-perfect performance in Creative Senz3D (0.99) and NTU Hand Gesture (0.93). These outcomes might be related to the fact that GANimation, StarGAN and PG^2 are not designed for image translation *in the wild*.

△-GAN significantly outperforms the competing methods in the *challenging* setting, where we enforce a stricter division of training and test gestures, and we use the 100% of the data, differently from GestureGAN's settings. In particular, the FID score reduces by 49% and 60% from GestureGAN in NTU Hand Gesture and Creative Senz3D, respectively. In terms of F1-score, the result improves by 61% and 71% in NTU Hand Gesture and Creative Senz3D, respectively. The PSNR improves by 10% and 24% in NTU Hand Gesture and Creative Senz3D, respectively. We note that the rolling guidance applied to GestureGAN (denoted in Table 2 as GestureGAN†) improves the original FID results by 36% and 24% in NTU Hand Gesture and Creative Senz3D, respectively.

Altogether, the quantitative results show that △-GAN outperforms the state of the art, both in the *normal* and *challenging* setting.

Qualitative Results. Figure 4 shows some randomly selected gesture-to-gesture translations in the *normal* experimental setting. Both GestureGAN [34] and △-GAN produce sharper output results while the output gestures from PG^2 [18] are very blurry. △-GAN also produces a better define sharper background than GestureGAN. StarGAN and GANimation, however, fail to produce gestures from the provided conditional map.

We further inspect the reason behind the poor results of PG^2 and GANimation. These methods focus on multi-domain translation and are specifically tailored to the task of facial expression translation and to the cases where the input and target objects are aligned. Figure 5 depicts two sample cases of the difference and attention masks generated by these methods. It can be seen that PG^2 fails at finding the difference mask, especially in the NTU Hand Gesture dataset (top row). Similarly, GANimation generates an empty attention mask, which might also be the cause of its poor results in both Figure 4 and Table 2. △-GAN, instead, learns to focus on the differences between the source and target images.

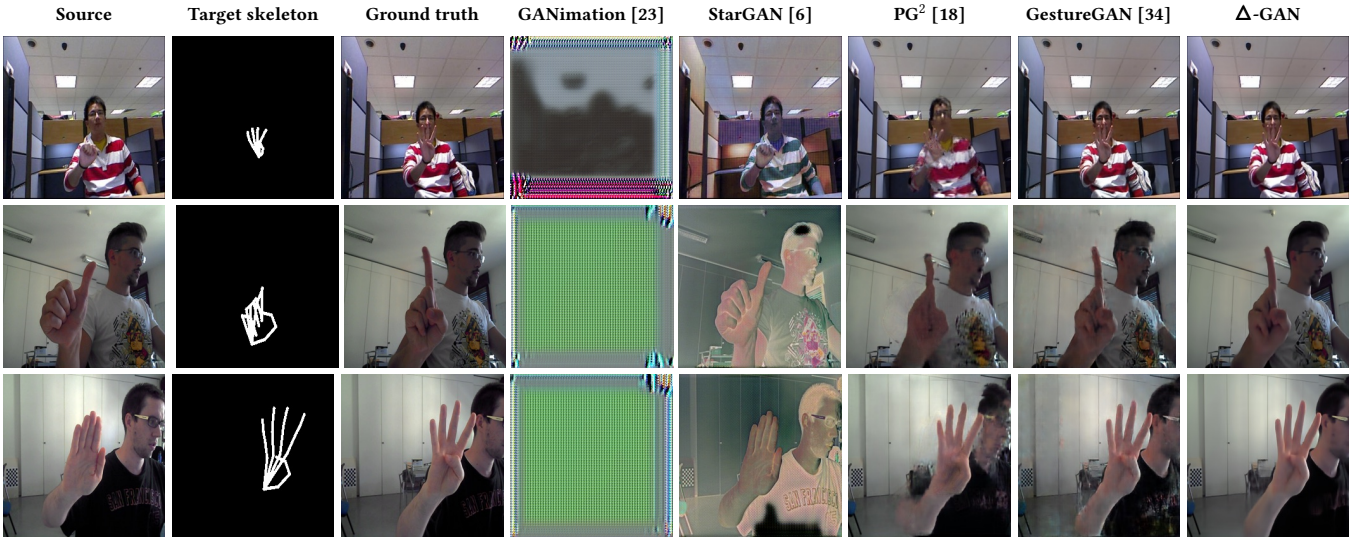

Figure 4: Qualitative comparison between △-GAN and competing works in the *normal* experimental setting. NTU Hand Gesture dataset (top two rows) and Creative Senz3D (bottom two rows).

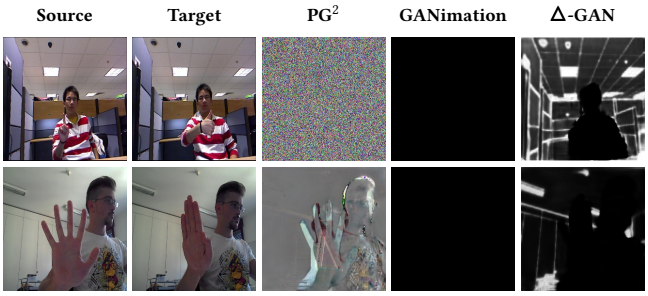

Figure 5: Masks computed by the various state of the art methods. Specifically, PG² computes a difference map that is noisy, GANimation fails at computing the attention mask, while Δ-GAN computes the attention of the pixels that stay constant from source to target images.

Figure 6 shows the results in the *challenging* setting. △-GAN generates sharper images with recognizable hand gestures, while other methods such as GestureGAN fails at it. This result is in line with the F1-scores reported in Table 2 (bottom), which range within 0.10 and 0.38 for the competing works, and within 0.58 and 0.61 in case of △-GAN. Both qualitative and quantitative results confirm that state-of-the-art methods are not adapted to perform gesture translation in the challenging settings, i.e. where a user-friendly category independent conditional map is provided, and where the network is asked to translate to an unseen gesture for a given user. We refer to the Supplementary Material for additional qualitative Figures.

Diversity Study. △-GAN decouples the desired hand gesture category, specified through a class number, from its location, scale and orientation, which are specified by a conditional map. This means that users can use the same conditional map with different hand gesture categories to generate multiple images. Figure 7 shows that △-GAN can generate three different distinguishable hand gestures by using the same *triangle* conditional map and different category numbers. Instead, when using alternative shapes that are not category-independent, such as boundary or skeleton, △-GAN fails to synthesize several hand gestures categories from the same conditional map. As aforementioned, *Boundary* maps are weakly category dependent, as their shape might suggest the type of gesture, while *Skeleton* and *Key-points* maps are category dependent.

Furthermore, we test the performance of △-GAN with the same source image and category, but with different conditional maps. For this test, we manually draw three triangle conditional maps with arbitrary size, location and orientation. Figure 8 shows that our method faithfully synthesizes the target gestures in all cases. Altogether, we show that users can generate hand gestures *in the wild* with much less effort than state-of-the-art models, that require complex annotations that are dependent on the specific hand gesture users want to generate.

Table 3: Perceptual user study. Percentage of times, on average, when the translated images are selected as "real" by users, in the "fake" vs. "real" comparison.

Model	NTU Hand Gesture [26]		Creative Senz3D [20]	
	Normal	*Challenging*	*Normal*	*Challenging*
GestureGAN	36.54%	3.21%	3.85%	0.64%
△-GAN	**44.32%**	**7.05%**	**16.03%**	**3.85%**

Perception User Study. Table 3 shows the outcome of the perceptual user study, where we report the percentage of times when the translated image wins against the original target image in the real vs fake comparison, i.e when the original image is selected as fake. It can be seen that △-GAN outperforms GestureGAN [34] in both experimental settings. This means that △-GAN generates higher quality images that can be mistaken with the real ones at higher rates than competitors.

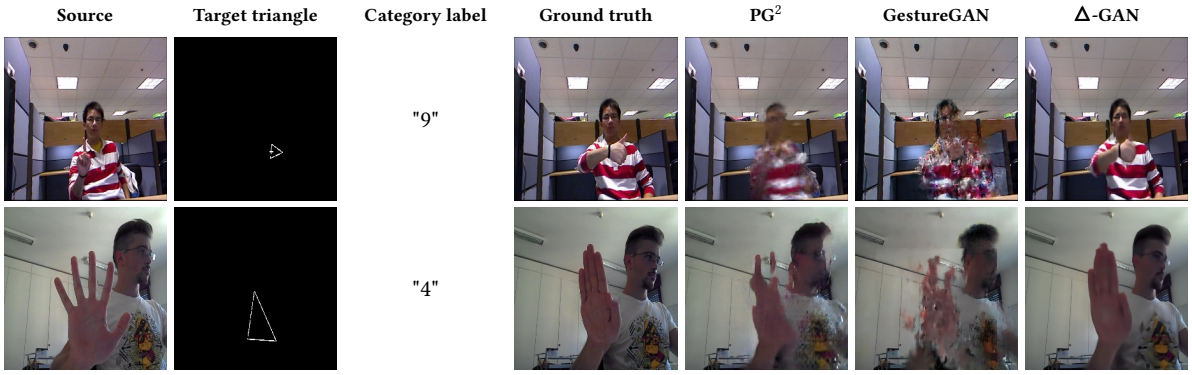

Figure 6: Qualitative comparison between △-GAN and competing works in the *challenging* experimental setting. NTU Hand Gesture dataset (top row) and Creative Senz3D (bottom row).

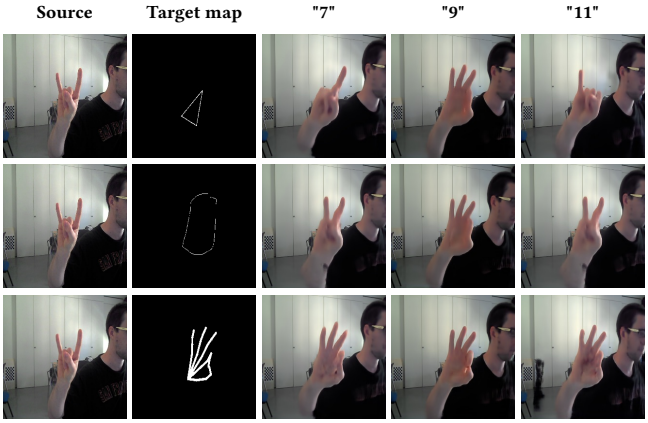

Figure 7: △-GAN decouples the conditional map from the category label of the hand gesture. The same conditional map can be used with different category labels to generate multiple images.

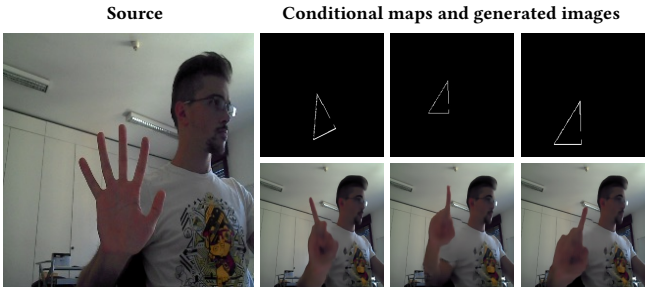

Figure 8: △-GAN generates images with the same conditional map that is rotated, shifted and resized.

Table 4: Performance degradation of Δ-GAN by removing the rolling guidance approach.

Conditional map	Category independent	Creative Senz3D [20] PSNR	FID	F1
Triangle	✓	1.53% ↓	24.77% ↑	75.76% ↓
Boundary	~	1.84% ↓	35.37% ↑	58.06% ↓
Skeleton	×	10.66% ↓	81.90% ↑	12.50% ↓

Ablation study. We further investigate the effect of the rolling guidance approach by removing this component from △-GAN. In the quantitative results, Table 2 it can be seen that the rolling guidance allows to significantly improves the quality of the generated images. In Table 4 we report the degradation of our method performance without rolling guidance, on Creative Senz3D dataset, for the three types of annotation maps. Specifically, we observe 24.77%, 35.37% and 81.90% worse (increased) FID score for the *Triangle*, *Boundary* and *Skeleton* maps, respectively. While the F1-score decreases by 75.76% and 58.06% on the *Triangle* and *Boundary* maps respectively, for *Skeleton* maps it decreases by 12.50%. In terms of PSNR the degradation is less significant for the *Triangle* and *Boundary*, but higher (10.66%) for *Skeleton* maps. Finally, we observed that a single rolling is sufficient to improve the generated results, while △-GAN does not benefit from additional rolling iterations.

6 CONCLUSION

We have presented a novel GAN architecture for gesture-to-gesture translation *in the wild*. Our model decouples the conditional input into a category label and an easy-to-draw conditional map. The proposed attention module and rolling guidance approach allow generating sharper images *in the wild*, faithfully generating the target gesture while preserving the background context. Experiments on two public datasets have shown that △-GAN outperforms state of the art both quantitatively and qualitatively. Moreover, it allows the use of simple-to-draw and category-independent conditional maps, such as triangles. This significantly reduces the annotation effort both at drawing time and allowing to use a single map for multiple gesture translations. The proposed framework is not limited to category labels but can also be used through embedding, learned from the data, that can express the gesture characteristics. In future work users could easily provide a reference image instead of a label, and translate the original image to the target gesture expressed by the reference image.

Our approach is especially important when the target image, and thus the target conditional map, is not available, which is the typical scenario in photo-editing software.

ACKNOWLEDGMENTS

We gratefully acknowledge NVIDIA Corporation for the donation of the Titan X GPUs and Fondazione Caritro for supporting the SMARTourism project.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*.

[2] Shane Barratt and Rishi Sharma. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973* (2018).

[3] Ali Borji. 2019. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding* 179 (2019), 41–65.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

[6] Yunjey Choi, Minje Choi, and Munyoung Kim. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. *IEEE International Conference on Computer Vision (CVPR)* (2018).

[7] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Ian Goodfellow. 2016. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160* (2016).

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*.

[15] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*. JMLR. org.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *IEEE international conference on computer vision (ICCV)*.

[18] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems (NIPS)*.

[19] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Alvise Memo and Pietro Zanuttigh. 2018. Head-mounted gesture controlled interface for human-computer interaction. *Multimedia Tools and Applications* 77, 1 (2018), 27–53.

[21] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2019. InstaGAN: Instance-aware Image-to-Image Translation. *International Conference on Learning Representations (ICLR)* (2019).

[22] Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. 2018. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3136–3145.

[23] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*.

[24] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[25] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *Advances in Neural Information Processing Systems (NIPS)*.

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

[27] Enrique Sanchez and Michel Valstar. 2018. Triple consistency loss for pairing distributions in GAN-based face synthesis. *arXiv preprint arXiv:1811.03492* (2018).

[28] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Aliaksandr Siarohin, Stéphane Lathuilière, Segey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating Arbitrary Objects via Deep Motion Transfer. *IEEE International Conference on Computer Vision (CVPR)* (2019).

[30] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[32] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Cision and Pattern Recognition (CVPR)*.

[34] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. 2018. GestureGAN for Hand Gesture-to-Gesture Translation in the Wild. In *ACM on Multimedia Conference*.

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

[36] Yichao Yan, Jingwei Xu, Bingbing Ni, Wendong Zhang, and Xiaokang Yang. 2017. Skeleton-aided articulated motion generation. In *ACM international conference on Multimedia*.

[37] Xuewen Yang, Dongliang Xie, and Xin Wang. 2018. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. *ACM on Multimedia Conference* (2018).

[38] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. 2018. Semantic Image Inpainting with Progressive Generative Networks. In *ACM Conference on Multimedia*.

[39] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. 2014. Rolling guidance filter. *European Conference on Computer Vision (ECCV)* (2014).

[40] Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer. 2018. EmotionGAN: Unsupervised Domain Adaptation for Learning Discrete Probability Distributions of Image Emotions. *ACM on Multimedia Conference* (2018).

[41] Sergey Tulyakov Zhenglin Geng, Chen Cao. 2019. 3D Guided Fine-Grained Face Manipulation. *IEEE International Conference on Computer Vision (CVPR)* (2019).

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*.

[43] 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 *Advances in Neural Information Processing Systems (NIPS)*.