Fine-grained Image-to-Image Transformation towards Visual Recognition

Wei Xiong[1], Yutong He[1], Yixuan Zhang[1], Wenhan Luo[2], Lin Ma[2], and Jiebo Luo[1]

[1]Department of Computer Science, University of Rochester
[2]Tencent AI Lab

Abstract

Existing image-to-image transformation approaches primarily focus on synthesizing visually pleasing data. Generating images with correct identity labels is challenging yet much less explored. It is even more challenging to deal with image transformation tasks with large deformation in poses, viewpoints or scales while preserving the identity, such as face rotation and object viewpoint morphing. In this paper, we aim at transforming an image with a fine-grained category to synthesize new images that preserve the identity of the input image, which can thereby benefit the subsequent fine-grained image recognition and few-shot learning tasks. The generated images, transformed with large geometric deformation, do not necessarily need to be of high visual quality, but are required to maintain as much identity information as possible. To this end, we adopt a model based on generative adversarial networks to disentangle the identity related and unrelated factors of an image. In order to preserve the fine-grained contextual details of the input image during the deformable transformation, a constrained nonalignment connection method is proposed to construct learnable highways between intermediate convolution blocks in the generator. Moreover, an adaptive identity modulation mechanism is proposed to effectively transfer the identity information into the output image. Extensive experiments on the CompCars and Multi-PIE datasets demonstrate that our model preserves the identity of the generated images much better than the state-of-the-art image-to-image transformation models, and as a result significantly boosts the visual recognition performance in fine-grained few-shot learning.

1. Introduction

Image-to-image transformation is an important field of visual synthesis and has many successful applications [21, 46, 42, 17, 48]. One important application of image-to-image transformation is to synthesize new images that can benefit the visual recognition systems. For example, synthesized images can augment the original training data, and subsequently boost the performance of image classification tasks [1, 41]. Synthesized images that well preserve the categorical information of the input image have been successfully applied to boost face verification [45, 3], person re-identification [29] and so on.

In this paper, we focus on fine-grained image-to-image transformation for visual recognition, *i.e.*, transforming an image with a fine-grained category to synthesize new images that preserve the identity of the input image, so that the new samples can be used to boost the performance of recognition systems. We pay special attention to transformations with large geometric deformations in object pose, viewpoint, and scale, *e.g.*, face rotation [15], caricature generation [26] and object attribute editing [2, 23] without ruining the class/identity. For the ultimate goal of recognition, the generated images are not necessarily required to be of high visual quality, but should be correctly classified even under the scenarios of fine-grained generation. Achieving such a goal is difficult, since images from different fine-grained categories may exhibit only subtle differences. Transforming an image with geometric deformations can easily change the category of the image.

Conventional image-to-image transformation approaches primarily focus on synthesizing visually pleasing images. However, models that perform well in generating visually pleasing data do not necessarily generate identity-preserved data, thus may not benefit the subsequent recognition tasks. The problem is even more severe in fine-grained recognition because the differences between categories are inherently subtle. A possible reason is that existing generative models are not specifically designed for fine-grained image synthesis with identity preservation and visual recognition in mind.

Specifically, the performance of existing generators may be limited for the following reasons. 1) Typical generators for image-to-image transformation adopt an encoder-decoder architecture. The encoder maps the image to a con-

densed latent feature representation, which is then transformed to a new image by the decoder. During encoding, the latent feature fails to preserve the fine-grained contextual details of the input image, which contain rich identity information. An alternative way to preserve the contextual details is using skip-connections [34] to link feature blocks in the encoder and decoder. However, skip-connections can connect only pixels of the same spatial location in the feature blocks. It may fail on transformations with geometric deformations where there is no pixel-wise spatial correspondence between the input and output. 2) In a generator with a typical encoder-decoder architecture, the output image is decoded from the latent feature with long-range non-linear mappings. During decoding, the identity information contained in the latent feature can be weakened or even missing [23]. As a consequence, the identity of the output image is not well preserved.

To address deformable transformation problem while maintaining contextual details, we propose a constrained nonalignment connection method to build flexible highways from the encoder feature blocks to the decoder feature blocks. With learnable attention weights, each feature point in a decoder block can non-locally match and connect to the most relevant feature points within a neighborhood subregion of an encoder block. As such, rich contextual details from the encoder blocks can be transferred to the output image during the deformable transformation.

To address the second problem, we propose an adaptive identity modulation method which can effectively decode the latent feature and preserve identity information. Specifically, we embed the identity feature into each convolution block of the decoder with an adaptive conditional Batch Normalization. The identity information can then be incorporated into features at different spatial resolutions and can be transferred into the output image more effectively.

In order to generate images that better preserve the identity, we adopt a generative adversarial network (GAN) [10] based framework to disentangle the identity-related factors from the unrelated factors. We apply our proposed model to two large-scale fine-grained object datasets, *i.e.*, the CompCars car dataset [43] and the Multi-PIE face dataset [11]. Given an image with a fine-grained category, we alter the viewpoint of the image to generate new images, which are required to preserve the identity of the input image. These generated images can benefit the few-shot learning task when they are used for data augmentation.

Our primary contributions are summarized as follows.

- We propose a constrained nonalignment connection method to preserve rich contextual details from the input image.
- We propose an adaptive identity modulation mechanism to better decode the identity feature to the output image, so that the identity is better preserved.
- Our model outperforms the state-of-the-art generative models in terms of preserving the identity and boosting the performance of fine-grained few-shot learning.

2. Related Work

Generative Image-to-Image Transformation. Existing works have adopted conditional GANs [31] for image-to-image transformation tasks, such as image inpainting [46, 42], super-resolution [25], and general-purpose image-to-image translation tasks [19, 49]. Many models mainly handle scenarios where the input image and output image have pixel-wise spatial correspondence, and tend to fail on geometric transformation tasks, which are specifically addressed by our work. Recent works have made attempts on geometric transformation tasks, including object rotation and deformation learning with spatial transformer networks [20] and deformable convolution [8], face viewpoint rotation [38, 16], person generation with different poses [29, 28] and vehicle generation with different viewpoints [50, 30].

However, existing works primarily aim at synthesizing data of high visual quality, and are not specifically designed to preserve the identity of the generated images especially under the scenarios of fine-grained image transformation, which is our primary goal. For example, StyleGAN [23] and PG-GAN [22] can generate high-quality faces, but the faces have no identity labels. There is a set of works that can synthesize fine-grained categorical images [2]. However, they are conditioned on category labels, which thereby cannot generalize to new categories.

Our work differs from the conventional image transformation works in the following aspects. 1) Our primary goal is to synthesize images with a correct identity, so that the generated images can benefit the subsequent fine-grained recognition tasks. Our model is specifically designed for preserving the fine-grained details that can benefit identity preservation. We emphasize that high visual quality is *not* necessarily required for identity preservation. 2) We address the task of image-to-image transformation with large geometric deformations. There is no pixel-wise correspondence between the input and the output images. 3) Our model can generalize to unseen categories. Therefore it can benefit the few-shot learning task by augmenting the data in new categories.

Non-Local Networks. Our proposed constrained nonalignment connection is related to non-local networks. The idea of non-local optimization has been proposed and used in many traditional vision tasks, such as filtering and denoising [4, 7]. Recently, such an idea has been extended to neural networks to compute the long-range dependencies within feature maps, such as non-local neural networks

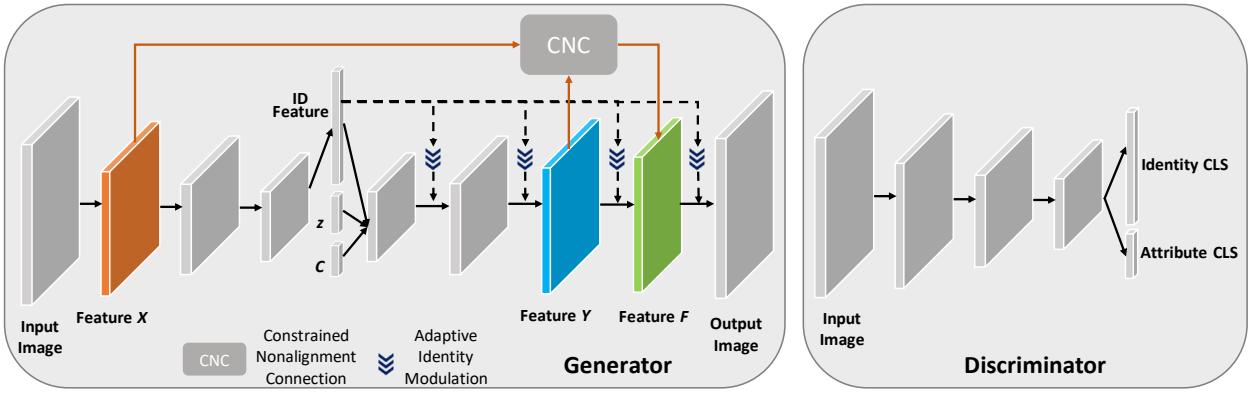

Figure 1: An overview of our model. In the generator, we use constrained nonalignment connections to preserve fine-grained contextual details from the input image, and use adaptive identity modulation to strengthen the identity information in the decoder. The discriminator predicts both the identity and attribute labels of a real or generated image (CLS: classifier).

[39, 27, 6] and self-attention GAN [47]. Our model differs from the existing non-local structure in two aspects. First, we compute non-local attention between different feature maps to construct information highways in a deep generator, while existing models typically calculate attentions within the same feature, *i.e.*, self-attention. Second, conventional non-local structures usually calculate the attention in the whole searching space, which may be difficult to optimize. On the contrary, our proposed constrained nonalignment connection reduces the non-local searching scope to capture the feature correspondences more effectively.

Network Modulation. Network modulation is a technique that modulates the behavior of network layers with a given conditioning feature [9], and has been proved effective in several tasks [44, 40, 33, 36, 5, 23]. It is typically realized by mapping the conditioning feature to the hidden variables of a layer, such as the re-scale factors of Batch Normalization [9] or Instance Normalization [23]. In our work, we propose a novel modulation method that can regularize the convolution layers by adaptively integrating the identity feature and the convolutional feature maps.

3. Our Approach

As shown in Fig. 1, our model is composed of a generator G and a discriminator D. The generator takes an image I, random noise z and a condition code C as inputs, and generates an image I_f. C is a vector encoding an attribute of image, such as viewpoint or pose. The discriminator takes an image as input, and outputs both the identity and attribute class probabilities. The identity of I_f is required to be the same as that of input image I, *i.e.*, identity preservation.

3.1. Generator

Our generator adopts an encoder-decoder architecture, *i.e.*, $G = \{Enc, Dec\}$. The encoder Enc maps the input image to an identity feature vector $f_{id} = Enc(I)$, which is then concatenated with noise z and target attribute code C to form the latent vector $f_l = cat[f_{id}, z, C]$. The latent vector is then decoded by the decoder Dec to the output image $I_f = Dec(f_l)$. To preserve the contextual details of the input image during deformable transformation, we propose a constrained nonalignment connection $CNC(X, Y)$ that can link the intermediate feature map X in the encoder and feature map Y in the decoder with non-local attention maps. To better preserve the identity, we propose an adaptive identity modulation method to effectively embed the identity feature f_{id} into the convolution blocks of the decoder.

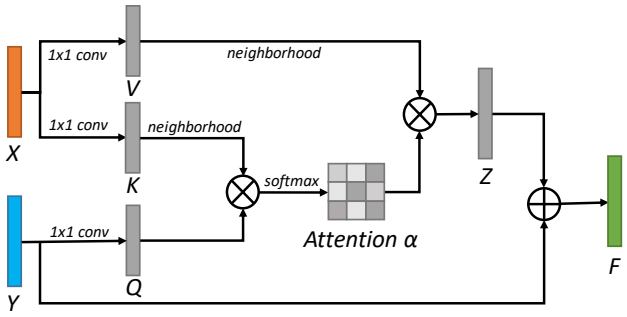

Figure 2: Structure of constrained nonalignment connection. $\otimes$ denotes matrix multiplication. $\oplus$ denotes concatenation.

3.1.1 Constrained Nonalignment Connection

Fig. 2 shows the structure of our constrained nonalignment connection. Consider an intermediate feature map $X \in \mathbb{R}^{C_X \times H_X \times W_X}$ in the encoder and an intermediate feature map $Y \in \mathbb{R}^{C_Y \times H_Y \times W_Y}$ in the decoder. (We ignore the batch size for simplicity.) Feature Y may lose fine-grained contextual details which are complementary

for identity preservation during layers of mapping in the generator [13]. To address this issue, we selectively link Y and X with a non-local attention map, so that the attended feature Z contains rich contextual details from X, while the generator still learns a correct geometric transformation.

Specifically, we first reshape the feature X to the shape $C_X \times N_X$, where $N_X = H_X \times W_X$. Similarly, we obtain the reshaped feature $Y \in \mathbb{R}^{C_Y \times N_Y}$. We then use several 1×1 convolutions to project X into key $K \in \mathbb{R}^{C_h \times N_X}$, value $V \in \mathbb{R}^{C_h \times N_X}$ and Y into query $Q \in \mathbb{R}^{C_h \times N_Y}$, so that they are in the same feature space.

Next, for each spatial location p in Q, we use the feature point Q_p to attend to the feature points in K and obtain a non-local attention map α_p. Conventional non-local networks typically calculate the attention map by matching Q_p with features of all the spatial locations in K, which is both time-consuming and difficult to optimize. Considering a point in the input image, in most situations, after the geometric transformation, the spatial location of that point is usually changed within a certain neighborhood region around the point. Inspired by this observation, we propose a *constrained non-local matching* between the query Q and the key K. As shown in Fig. 3, for each spatial location p in Q, we define a corresponding neighborhood region in K as $\mathcal{N}_p$, which is a square area with its center at location p. We define the radius of the neighborhood with a hyper-parameter r, then the spatial size of the neighborhood region is $(2r+1) \times (2r+1)$. For each location p, we extract all the features in neighborhood $\mathcal{N}_p$ from feature K, denoted as $K_{\mathcal{N}_p} \in \mathbb{R}^{C_h \times (2r+1)(2r+1)}$, then use Q_p to attend to $K_{\mathcal{N}_p}$ and calculate the constrained non-local attention as

$$\alpha_p = Q_p^T K_{\mathcal{N}_p}. \tag{1}$$

We normalize α_p using the softmax function so that the weights are summed to 1. Feature at location p of the attended feature Z is the weighted sum over all the feature points in neighborhood $\mathcal{N}_p$ of the value V, formulated as

$$Z_p = \sum_{i \in \mathcal{N}_p} \alpha_p^i V_{\mathcal{N}_p}^i. \tag{2}$$

We then concatenate the attended feature with the original feature Y, to obtain the final fused feature $F = [Y, Z]$.

3.1.2 Adaptive Identity Modulation

In the decoder, directly mapping the latent feature to an image with layers of convolution may not be optimal. During the long-range mapping, the identity information may be weakened [13, 23] or missing. To address this problem, we propose an adaptive identity modulation method to effectively transfer the identity information to the output image.

Specifically, we embed the identity feature into the convolution blocks, so that feature maps at each spatial resolution can perceive and utilize the identity knowledge. To this end, we adopt conditional batch normalization [9] to modulate the convolution layers [9, 32] with identity feature f_{id}.

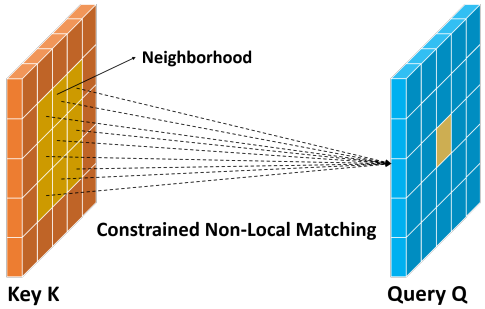

Figure 3: Illustration of constrained non-local matching between query Q and key K. Each feature point in Q can non-locally match over all the spatial locations inside a neighborhood region of K.

Given a mini-batch of features $\{B_{i,:,:,:}\}_{i=1}^N$ in a convolution layer, we first normalize each feature B_i with Batch Normalization (BN) [18]

$$\hat{B}_i = \frac{B_i - \mathbb{E}[B_{:,c,:,:}]}{\sqrt{var[B_{:,c,:,:}] + \epsilon}}, \tag{3}$$

where B_i is the feature map of the i-th sample in the batch, ϵ is a constant for numerical stability. In the vanilla BN, we re-scale the feature with two learnable parameters γ and β.

In order to better decode the identity feature, we adopt a conditional Batch Normalization (CBN) to learn the re-scale parameters γ and β on condition of the identity feature f_{id}. Then in each convolution block we have

$$\tilde{B}_i = \gamma(f_{id})\hat{B}_i + \beta(f_{id}), \tag{4}$$

where $\gamma(f_{id})$ and $\beta(f_{id})$ are functions of f_{id}.

In traditional CBN, the re-scale parameters γ and β usually depend only on the conditioning feature. However, we argue that different feature maps should perceive the conditioning feature in different ways. Features in different convolution layers exhibit different functionalities, and may pay different attention to the conditioning feature. In order to adaptively perceive and integrate the conditioning feature, we re-formulate γ and β to be conditioned on both the feature map to be modulated and the conditioning feature:

$$\tilde{B}_i = \gamma(f_{id}, B_i)\hat{B}_i + \beta(f_{id}, B_i), \tag{5}$$

where $\gamma(f_{id}, B_i)$ and $\beta(f_{id}, B_i)$ are functions of f_{id} and B_i.

Specifically, we first calculate the average feature B_f of B_i over spatial locations, *i.e.*, $B_f = \frac{1}{H \times W} \sum_{h,w} B_{i,:,h,w}$. Then we calculate an attention using B_f, formulated as $att_B = \tau(B_f)$, where τ can be realized with a MLP composed of several dense layers with the activation of the last

layer to be Sigmoid. We obtain the attended feature as:

$$f_{id}^{att} = f_{id} \odot att_B, \tag{6}$$

where $\odot$ denotes element-wise multiplication. As such, the identity feature is adaptively selected by the feature map B_i.

The attended identity feature f_{id}^{att} is then mapped to γ and β with two MLPs. By embedding the identity features into convolution layers on condition of the features to be modulated, the identity related information can be better integrated by the decoder.

3.2. Discriminator and Objective Functions

To encourage the model to generate identity-preserved images, our discriminator D adopts a similar architecture as ACGAN [31]. D is composed of several convolution blocks, followed by an identity classification layer D^i, and an attribute classification layer D^a.

We denote y_a^t as the target attribute label, which can be encoded into the one-hot code C. During training, the identity label y_i and the attribute label y_a of the input image I are provided to train the classifier in D, where $1 \leq y_i \leq N_i$ and $1 \leq y_a \leq N_a$. N_i and N_a are the number of identity and attribute categories in the training data, respectively.

Upon training the discriminator, we assign the ground-truth identity label of the fake image I_f as $N_i + 1$. In this way, the discriminator can not only classify the real image, but also distinguish the real image from the fake one. We use the following objective to optimize D:

$$\begin{aligned} \max_D J(G, D) = \mathbb{E}[logD_{y_i}^i(I) + \lambda \cdot logD_{y_a}^a(I)] \\ + \mathbb{E}[logD_{N_i+1}^i(G(I))], \end{aligned} \tag{7}$$

where J is the value function, D_k^i and D_k^a are the k-th element in D^i and D^a, respectively. λ denotes the weight of attribute classification loss.

When training the generator, we encourage the generated image to have the same identity label y_i as the input image as well as the target attribute label y_a^t by optimizing the following objective:

$$\max_G J(G, D) = \mathbb{E}[\lambda \cdot logD_{y_a^t}^a(G(I))] + \mathbb{E}[logD_{y_i}^i(G(I))], \tag{8}$$

4. Experiments

We evaluate our model on two challenging datasets, CompCars dataset [43] and Multi-PIE dataset [11]. CompCars dataset contains over 1,700 categories of car models and 100,000 images. Multi-PIE dataset contains face images of 337 identities. Both datasets are *quite large for fine-grained image generation and few shot learning*. We perform viewpoint morphing on both datasets. Given an image, a target viewpoint and random noise, our goal is to generate novel images belonging to the same identity/model category as the input image with the target viewpoint. We conduct two types of experiments. The first one is identity preservation. In this experiment, we derive a classifier on the real images, which are then used to classify the generated images. The second type is few-shot learning. In this experiment, we use the generated images to augment the training data, and test how the generative models can benefit the performance of the few-shot classifier.

4.1. Experiment Settings

Dataset. For Multi-PIE dataset, following the setting in [38], we use 337 subjects with neutral expression and 9 poses within ± 60 degree. The first 200 subjects form an auxiliary set, which is used for training the generative models. The rest 137 subjects form a standard set, which is used to conduct visual recognition experiments. We crop and align the faces and resize each image to 96×96.

The car images in the CompCars [43] dataset contain several viewpoints, including frontal view, frontal left side, rear view, rear side, side, and so on. Note that the same car model can have totally different colors. Since the rear views may contribute less to the identification of the car model, we remove all the images with rear views and keep only images with the following five viewpoints: frontal, frontal left, frontal right, left side, and right side. We also remove minor categories containing less than 10 samples. All the images are resized to 224×224. Similar to the setting in Multi-PIE, we split the filtered dataset into an auxiliary set which contains images of $1,181$ car models, and a standard set which contains images of another 296 car models. These two sets are disjoint in terms of the category label.

Existing Models to Compare. We compare our model with the state-of-the-art models DR-GAN [38], CR-GAN [37] and Two-step [12], which also aim at generating fine-grained objects given a target attribute as the condition. For fair comparison, we adjust the generator of each model to have a comparable amount of parameters. Note that there are other models for image-to-image transformation. However, many of them need pose masks or landmarks as guidance [26, 29], which differs from our setting. Therefore, it is not appropriate to compare them with our model. We also do not compare our model with StyleGAN [24], PG-GAN [22] or other similar models since they are unconditional models that cannot generate categorical images.

Evaluation Metric. Since our task is visual recognition oriented image transformation, we primarily evaluate the identity preservation performance of each model and report classification accuracy on identity preservation and few-shot learning experiments. We do not use FID [14] or Inception Score [35] to quantify the generated images, since they are mainly used to evaluate the visual quality of images.

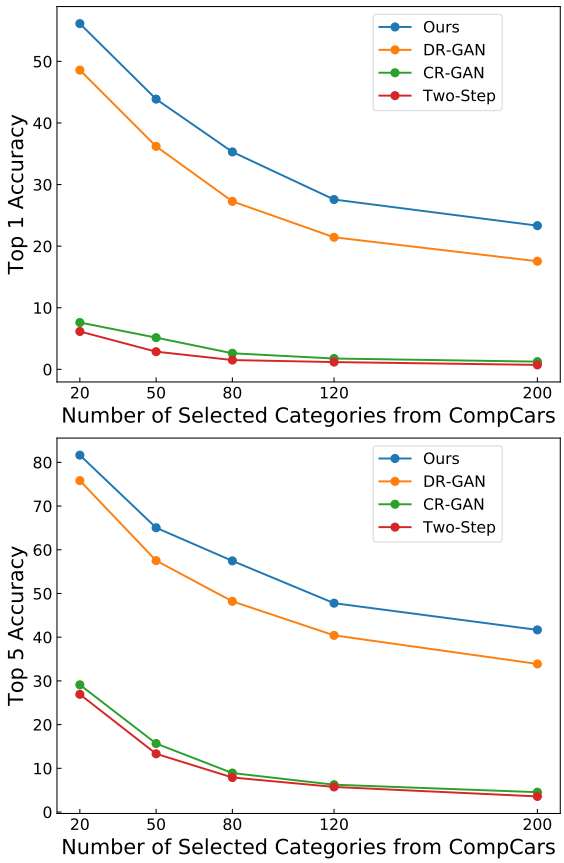

Figure 4: Classification accuracy on the generated images from CompCars dataset with 20, 50, 80, 120 and 200 categories.

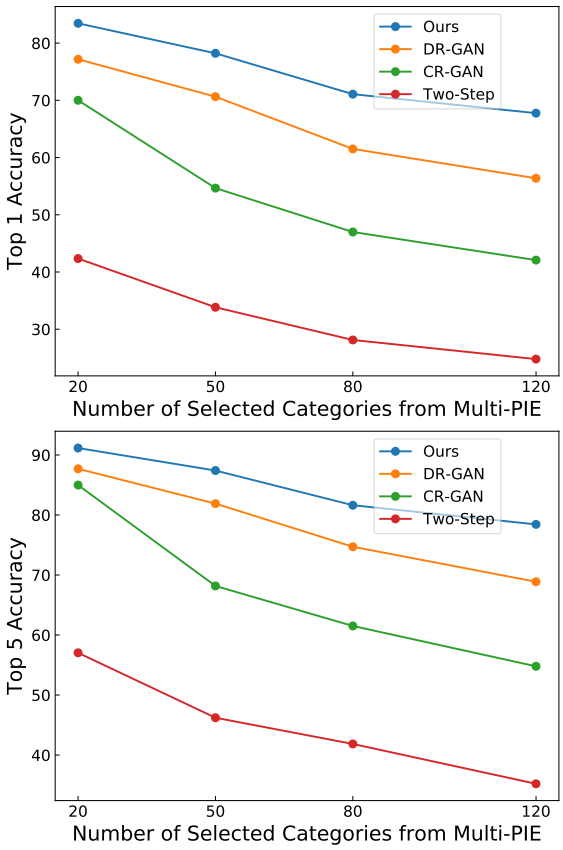

Figure 5: Classification accuracy on the generated images from Multi-PIE dataset with 20, 50, 80 and 120 categories.

Implementation Details. Our model is optimized with the Adam optimizer. The learning rate is 0.0002 and the batch size is 64. On CompCars dataset, the target viewpoint code C is a 5×1 one-hot vector. We empirically choose the radius of neighborhood $r = 7$ for feature maps with size 28×28 and $r = 14$ for feature maps with size 56×56. We set λ in Eq. (7) and (8) to be 5. On Multi-PIE dataset, the target viewpoint code is a 9×1 one-hot vector. We empirically choose the radius of neighborhood $r = 6$ for feature maps with size 24×24. The noise vector has a size of 128×1. We set λ in Eq. (7) and (8) to be 1.

4.2. Identity Preservation

In this section, we evaluate the identity preservation ability of each generative model on both CompCars and Multi-PIE datasets. On each dataset, we first train each model on the whole auxiliary set to learn the viewpoint transformation. We also train a Resnet18 [13] model on the auxiliary set, then use its features of the last pooling layer as the representation for identity classification experiments.

On CompCars dataset, we select N_c car models from all the 296 classes in the standard set, and choose all the images in the selected N_c classes to form the dataset on which the classification experiment will be conducted. We randomly split the selected dataset as train and test sets with the ratio $8:2$. Note that the train and test set contain images from all the N_c classes. We train a KNN classifier on the train set with the Resnet18 model as the feature extractor. Following that, for each image in the test set, we transform it with the generative model which outputs 5 images, one per specific target viewpoint. We then use the KNN classifier to classify all the generated images and report the top-1 and top-5 accuracies of each model. We choose the KNN classifier as it is parameter free, so that it can straightforwardly reveal the separability of the generated samples.

Fig. 4 shows the results of each model. We select N_c to be 20, 50, 80, 120 and 200. Our full model with both CNC and AIM significantly outperforms the existing models by a large margin (over **5%** accuracy gain under all settings), showing that our model can better preserve the identity of the generated images.

We conduct a similar identity preservation experiment on Multi-PIE dataset, except that we select N_c to be 20, 50, 80

and 120 from 137 classes in the standard set and generate 9 fake images (viewpoints ranging from -60 degree to 60 degree) from each input test image. Fig. 5 shows the classification results of each model on the generated face data. Our model again outperforms the existing models, further demonstrating the superiority of our model.

To make a more thorough analysis on the results, we investigate each model by showing their visual results straightforwardly, as shown in Fig. 6 on CompCars dataset and Fig. 7 on Multi-PIE dataset.

Observed from Fig. 6, DR-GAN, CR-GAN and our model can generate sharp images, while Two-Step can only generate blurry images. Although images generated by CR-GAN look realistic, the key regions that identify a car (such as bumper and lights) are quite different from the input image, showing that their identity is not well preserved. This observation is consistent with the classification performance in Fig. 4. The results further indicate that high-quality images do not necessarily stand for identity-preserved images. Our model can generate fine-grained details that are almost in accordance with the input image. Note that in some situations, our model fail to capture all the details of the input car. This is because we are dealing with fine-grained image transformation with large deformation, which is very challenging. Moreover, cars in our dataset contain many complex details, making the task more difficult to accomplish. Even though, images generated by our model still preserve many more details than all the existing methods, demonstrating the effectiveness of our model.

Fig. 7 shows an exemplar case from Multi-PIE dataset. We input the same image to the generative models, outputting images with 9 different viewpoints. DR-GAN, CR-GAN and Two-Step fail to preserve the identity very well, while our model can generate images whose identity is almost the same as the input image, with as many details preserved as possible, demonstrating the effectiveness of our model in identity preservation.

4.3. Few-shot Learning

In this section, we evaluate how well each generative model can boost the performance of the few-shot learning task when used as a data augmentation method. Experiments are conducted on CompCars dataset. Similar to the identity preservation experiment, we train the generative models on the whole auxiliary set.

We randomly select N_c car models from all the 296 model classes in the standard set. Then we select images of the N_c classes to form a selected dataset on which we will conduct the few-shot learning experiment. We randomly select s images from each car model (N_c car models in total) to form the few-shot train set, and use all the rest images as the test set. Under such a setting, the few-shot classification task can be named as "N_c way s shot" few-shot learning.

Figure 6: Exemplar images generated by different models on CompCars dataset. In each column, from the top to the bottom are: input image, and results of our model, DR-GAN [38], CR-GAN [37], Two-Step [12], respectively. Since all the models generate the correct viewpoint, we do not show the viewpoint here.

In this experiment, we adopt Resnet18 as the classifier * for few-shot learning. We first train the classifier only on the train set, which is then used to classify the images in the test set. Different from the setting in the identity preservation experiment, we classify the real images instead of the fake images. We then input the images in the train set to the generative model and generate 20 fake images per image in the few-shot train set and set their identity labels to be the same as the input image. To generate diverse images, we interpolate between different viewpoint codes and input the new code to the generator as the target viewpoint. The generated images are used to augment the train set.

We then retrain the Resnet18 on the augmented train set, and classify images in the original test set. Note that when training the Resnet18 classifier with the augmented data, we also input the real/fake label to the Resnet18, so that the model can balance the importance of generated data and real data. Specifically, when training the Resnet18 with a real image, we also input the label 1 (a 1-bit vector concatenated with the feature of global pooling layer in Resnet18) to the model. When training the Resnet18 with a fake image, we input label 0 to the model. During testing, since the test images are all real images, we input the label 1 along with the image to the classifier, to obtain the prediction.

We report the few-shot learning results boosted by different generative models under N_c classes, where $N_c = 20$ in our experiment. As shown in Table 1, without any augmented data, training on limited real samples leads to poor performance on the test data. Using the generated images by our model or DR-GAN to augment data can significantly

*the last layer of Resnet18 is modified to N_c nodes.

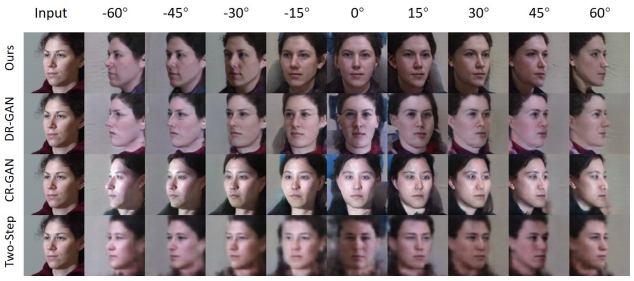

Figure 7: Exemplar images generated by different models on Multi-PIE dataset. From left to the right: input image, generated images with 9 different viewpoints. Please pay attention to the details such as face shape, hair and mouth.

Table 1: Classification accuracy of few-shot learning under different settings on CompCars dataset. "mw-ns" means m way n shot learning. "w/o" denotes "without", "$/w$" denotes "with".

Model	20w-5s %	20w-10s %
w/o augment	29.77	55.86
augment $/w$ Two-Step	32.04	52.53
augment $/w$ CR-GAN	27.61	39.67
augment $/w$ DR-GAN	47.85	60.01
augment $/w$ Ours	**52.44**	**66.93**

boost the performance of the classifier, indicating that it is an effective way to boost the few-shot learning by augmenting the data with generative models. Our model yields much better performance than DR-GAN. Interestingly, since the images generated by CR-GAN and Two-Step do not well preserve the identity, using them to augment data does not benefit few-shot classification. The results indicate that generators with better identity preservation ability can lead to more significant improvements on few-shot learning, while poor generator can even hurt the performance.

4.4. Ablation Study

We further analyze how each part of our model contributes to the overall performance. Specifically, we conduct the identity preservation experiment with the following versions of our model on CompCars dataset: 1) The vanilla model without constrained nonalignment connection (CNC) nor adaptive identity modulation (AIM). The vanilla model shares a similar architecture as DR-GAN. The generator has an encoder-decoder architecture (removing all the AIMs and CNCs), while the discriminator remains the same as our full model. 2) The vanilla model with deformable convolution [8] applied on the 28×28 feature block instead of the original convolution. 3) Model with unconstrained nonalignment connection, denoted as "Global-NC". Global-NC is a variant of CNC which modifies Eq. (1) to search over all the spatial locations in K, instead of merely searching a neighborhood region. 4) Model with only CNC. 5) Our full model with both CNC and AIM. The discriminator and the loss functions remain unchanged. We also study how the location of CNC influences the final performance. Therefore, we use CNC/Global-NC to connect convolution blocks with different spatial sizes. Specifically, as the structure of the encoder and the decoder in our model is symmetrical to each other, we choose to connect one block in the encoder with the corresponding symmetrical block in the decoder. We apply CNC and Global-NC on feature maps with a 28×28 or 56×56 spatial resolution.

Figure 8: Images generated by U-net (top) and our model (bottom). The first column shows the input image, and the rest columns are images generated with five different viewpoints as condition. Our model generates images with correct viewpoints while U-net fails to accomplish the task.

Results are shown in Table 2. Compared to the vanilla model, using deformable convolution benefits the performance. However, our model with CNC still outperforms deformable convolution. CNC significantly improves the performance of the model compared to Global-NC model and vanilla model by a large margin, demonstrating its effectiveness. Applying CNC to different feature blocks can influence the performance of the model. AIM also makes significant contributions in improving the identity preservation ability of the model.

CNC versus Skip-Connection. We further analyze how constrained nonalignment connection is crucial to the success of fine-grained image transformation with large geometric deformation. On CompCars dataset, we compare our model with a counterpart, which uses a U-net as the generator with skip-connections to link the encoder and decoder. The other settings of the U-net model remain the same as our model. Fig. 8 shows the images generated by our model and the U-net model. Unsurprisingly, U-net model ignores the target viewpoint condition, and generates images that are almost the same as the input image. It only accomplishes the task of reconstruction without changing the views. In contrast, our model can generate images with correct viewpoints, demonstrating the superiority of the proposed constrained nonalignment connection over skip-connection.

Table 2: Identity preservation experiment results with different versions of our model on CompCars dataset. Experiments are done with 20, 50, and 80 categories from the standard set. We report both top-1 and top-5 accuracies.

Model	20c-top1 %	20c-top5 %	50c-top1 %	50c-top5 %	80c-top1 %	80c-top5 %
vanilla	48.59	75.82	36.20	57.52	27.27	48.20
vanilla + Deformable Conv	49.75	76.08	37.26	58.53	28.82	48.81
vanilla + Global-NC(56)	50.37	76.25	37.45	58.21	29.23	49.39
vanilla + CNC(56)	52.45	78.31	39.42	60.52	31.38	52.88
vanilla + Global-NC(28)	53.12	77.08	38.30	59.12	30.40	52.13
vanilla + CNC(28)	55.05	80.16	42.24	63.49	34.68	56.09
vanilla + CNC(28) + AIM	**56.13**	**81.65**	**43.87**	**65.04**	**35.30**	**57.46**

5. Conclusion

We study fine-grained image-to-image transformation with the goal of generating identity-preserved images that can boost the performance of visual recognition and few-shot learning. In particular, we adopt a GAN-based model that learns to encode an image to an output image with different viewpoints as conditions. To better maintain the fine-grained details and preserve the identity, we propose constrained nonalignment connection and adaptive identity modulation which are demonstrated effective in our extensive experiments on the large-scale fine-grained CompCars and Multi-PIE datasets. Our model outperforms the state-of-the-art image transformation methods in identity preservation and data augmentation for few-shot learning tasks.

References

[1] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 1

[2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 1, 2

[3] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *CVPR*, pages 6713–6722, 2018. 1

[4] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 2

[5] T. Chen, M. Lucic, N. Houlsby, and S. Gelly. On self modulation for generative adversarial networks. In *ICLR*, 2019. 3

[6] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng. Aˆ 2-nets: Double attention networks. In *Advances in Neural Information Processing Systems*, pages 352–361, 2018. 2

[7] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[8] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2, 8

[9] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In *NeurIPS*, pages 6594–6604, 2017. 3, 4

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2, 5

[12] N. Hadad, L. Wolf, and M. Shahar. A two-step disentanglement method. In *CVPR*, pages 772–780, 2018. 5, 7

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 6

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[15] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun. Pose-guided photorealistic face rotation. In *CVPR*, pages 8398–8406, 2018. 1

[16] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, pages 2439–2448, 2017. 2

[17] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 1

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2

[20] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. 2

[21] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 1

[22] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 5

[23] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 1, 2, 3, 4

[24] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

[25] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 2

[26] W. Li, W. Xiong, H. Liao, J. Huo, Y. Gao, and J. Luo. Carigan: Caricature generation through weakly paired adversarial learning. *arXiv preprint arXiv:1811.00445*, 2018. 1, 5

[27] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018. 2

[28] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NeurIPS*, pages 406–416, 2017. 2

[29] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *CVPR*, pages 99–108, 2018. 1, 2, 5

[30] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019. 2

[31] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651. JMLR. org, 2017. 2, 5

[32] E. Perez, H. De Vries, F. Strub, V. Dumoulin, and A. Courville. Learning visual reasoning without strong priors. *arXiv preprint arXiv:1707.03017*, 2017. 4

[33] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3

[34] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 6

[36] F. Strub, M. Seurin, E. Perez, H. De Vries, J. Mary, P. Preux, and A. CourvilleOlivier Pietquin. Visual reasoning with multi-hop feature modulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018. 3

[37] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. Cr-gan: learning complete representations for multi-view generation. *arXiv preprint arXiv:1806.11191*, 2018. 5, 7

[38] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1415–1424, 2017. 2, 5, 7

[39] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[40] X. Wang, K. Yu, C. Dong, and C. Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 3

[41] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan. Low-shot learning from imaginary data. In *CVPR*, pages 7278–7286, 2018. 1

[42] W. Xiong, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo. Foreground-aware image inpainting. In *CVPR*, 2019. 1, 2

[43] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015. 2, 5

[44] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018. 3

[45] X. Yin and X. Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975, 2017. 1

[46] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018. 1, 2

[47] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 2

[48] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 1

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2

[50] J.-Y. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. Tenenbaum, and B. Freeman. Visual object networks: image generation with disentangled 3d representations. In *NeurIPS*, pages 118–129, 2018. 2