



TileGAN: category-oriented attention-based high-quality tiled clothes generation from dressed person

Wei Zeng¹ · Mingbo Zhao^{1,2} · Yuan Gao² · Zhao Zhang³

Received: 14 January 2020 / Accepted: 6 April 2020 / Published online: 8 May 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

During the past decades, applying deep learning technologies on fashion industry are increasingly the mainstream. Due to the different gesture, illumination or self-occasion, it is hard to directly utilize the clothes images in real-world applications. In this paper, to handle this problem, we present a novel multi-stage, category-supervised attention-based conditional generative adversarial network by generating clear and detailed tiled clothing images from certain model images. This newly proposed method consists of two stages: in the first stage, we generate the coarse image which contains general appearance information (such as color and shape) and category of the garment, where a spatial transformation module is utilized to handle the shape changes during image synthesis and an additional classifier is employed to guide coarse image generated in a category-supervised manner; in the second stage, we propose a dual path attention-based model to generate the fine-tuned image, which combines the appearance information of the coarse result with the high-frequency information of the model image. In detail, we introduce the channel attention mechanism to assign weights to the information of different channels instead of connecting directly. Then, a self-attention module is employed to model long-range correlation making the generated image close to the target. In addition to the framework, we also create a person-to-clothing data set containing 10 categories of clothing, which includes more than 34 thousand pairs of images with category attribute. Extensive simulations are conducted, and experimental result on the data set demonstrates the feasibility and superiority of the proposed networks.

Keywords Generative adversarial network (GAN) · Image-to-image translation · Attention · Clothes generation

1 Introduction

People are increasingly pursuing high-quality life with the improvement of happiness, and the fashion industry has gradually entered everyone's sight. Online shopping has

become an indispensable ingredient of people's lives because of its convenience, which allows buyers to purchase what they want without going out. However, there are still some unsettled problems in purchasing clothing online. For example, when someone noticed favorite costumes in magazines or videos, they felt like buying it, but it is not so easy to get access to it. To handle this problem, most e-commerce platforms like Taobao now support image retrieval, which makes it possible to find the target garment by feeding a photo into the search engine. If the raw images are used directly for retrieval, the retrieval performance is difficult to meet people's requirements, as in some cases, the photographs may have poor shooting angles or distorted deformations so that degrade the retrieval performance. In addition, another disadvantage of purchasing clothes online is that trying on the desired garment is an impossible task. People don't know whether they are satisfied with the dressing. Both issues will have a negative impact on the consumer shopping experience.

✉ Mingbo Zhao
mzhao4@dhu.edu.cn

Wei Zeng
2181339@mail.dhu.edu.cn

Yuan Gao
ethan.y.gao@my.cityu.edu.hk

Zhao Zhang
cszzhang@gmail.com

- ¹ Donghua University, Shanghai, People's Republic of China
- ² City University of Hong Kong, Kowloon, Hong Kong, SAR
- ³ Hefei University of Technology, Anhui, People's Republic of China

In order to handle the above problem, people need to generate a clear and tiled clothing picture of the clothes in the picture by means of image-to-image translation [13, 15, 22, 26, 39, 41, 52]. This is because for most garments, dozens of characteristic information of the garment is on the front, and the back or side view contain less information. As a result, the tiled image has most of the information that is valuable for retrieval [23–25, 27, 48] or other learning tasks [46]. In addition, a tiled clothes can also be further extended to virtual try-on [10, 53] by replacing the garment on the buyer with the generated tiled clothing picture, so that the buyer can have an general impression with the appearance of wearing this dress, which will greatly improve consumer shopping experience.

Details of the generated images are of critical importance for the generative model, and the synthetic result of generative adversarial nets [8] demonstrates its superiority in producing images that are consistent with human perception. Previous GAN-based methods like [15] have been employed in scene translations such as day to night, sketches to photos, etc. These tasks all have one thing in common: the input and output are structurally similar. However, when the input and output differ greatly in structure, the network cannot achieve satisfying results. The tough problem we have to deal with is to translate the picture of the model wearing the costume into a flat garment that was taken off the model. This inevitably encounters structural deformation problems. What's more, the input picture of a model may contain multiple kinds of clothing (tops and pants, jackets, T-shirts, etc.), the network does not know which type is our concern, which will lead to the generated clothing in a wrong category that is unexpected to us. In addition, when these images with multi-category apparel are fed as input to refine details, plenty of superfluous information (undesired portion such as unexpected clothes) will undoubtedly affect the generation performance.

In this paper, motivated by the recently developed image-to-image translation technique, we propose a two-stage image generation method for high-quality tiled clothes generation from dressed person. The first stage generates an image picture containing the appearance and category, and is dedicated to solving the shape change and controllability of the translation result in the image translation process. The second stage is to refine the coarse picture produced in the first stage, committed to solving the intractable issue that exists in the majority of rest methods: lack of details. The spatial transformer module [16] is introduced to deal with the negative impact of convolutional neural networks on structural changes that contributes to a poor performance. We introduce this module into the first phase, without adding too much computational overhead, and can implicitly transform the feature map in

an end-to-end manner. We constrain the generation process toward our conditional category by adding a category condition to the generator and employing an extra classifier [31] to calculate a category loss. In the second stage, the adversarial learning method is adopted to refine the rough picture obtained in the first stage. Both the generated rough image and the model image are regarded as input of network of the second stage. Therefore, we propose a novel bi-path attention-based generator which incorporates both the channel module and the self-attention module. When the model pictures are fed into the network, quantities of redundant information (limbs and head) will have a devastating effect on the network performance if all the information directly shuttled from the encoder to the subsequent decoder like [15]. We introduce attention modules [12, 50], which allows the network to focus on more valuable information and transmit the information in a more efficient manner.

In summary, our contributions are mainly as follows:

1. We propose a novel two-stage class-supervised attention-based image generation method.
2. In order to capture the structural and category information, a spatial transformation module is introduced in the first stage to mitigate the negative impact of structural deformation during image transformation. In addition, steering the network to synthesize clothing according to the conditional category, and overcoming the ambiguity of the generated result.
3. In order to refine details and texture, a dual path generator combined with channel attention module and self-attention module is proposed. In addition, the channel attention module is introduced instead of skip connection and a self-attention module is employed in the second stage to learn global dependency.
4. A supervised image-to-image translation data set of more than 34 thousand image pairs is established, each picture pair containing corresponding category.

In the following few sections, we introduce our paper in terms of related work, our method, experimental result analysis, and conclusion.

2 Related work

2.1 Generative model

Before the emergence of GANs, the image generation model is mainly based on the auto-encoder (AE), which consists of an encoder and a decoder. The variational auto-encoder (VAE) [21] solves the problem that AE can only be reconstructed and cannot generate new images, while the conditional variational auto-encoder (CVAE) [37] is

further extended from VAE to overcome the randomness of the generated result. However, all auto-encoders produce very blurry results and lack of detail. Ian Goodfellow et al. proposed GAN [8] which consists of two parts: a generator and a discriminator, and adopts adversarial learning to synthesize images that are indistinguishable from real images. GAN is promising for whose generator produces images by fitting data distribution. [3, 45] make efforts to boost the quality of generated pictures. Other efforts made by [1, 2, 9, 19, 29, 32, 35] concentrate on making training stable and converge early. There are also abundant derivatives of GAN which has been proved feasible in computer vision, such as text-to-image transformation [33, 49, 53] and image-to-image translation [13, 15, 22, 26, 39, 41, 52], high-quality image generation [3, 45], and image super-resolution [22, 28, 41, 44, 51].

2.2 Image-to-image translation

Image-to-image translation is a typical conditional generative adversarial network [30]. In the original conditional GAN, the condition is a vector which represents category or something else while an input image is regarded as the condition in image translation tasks. The image translation network is somewhat similar to the function in mathematics. For an input, there is a unique output corresponding to it after passing through a mapping function, such a mapping is what our network expects to learn. Pix2pix [15] is a bi-directional method that can learn a mapping from input to output or vice versa. Pix2pixHD [39] attempts to generate high-resolution street view pictures conditioned on semantic segmentation images. Both [15] and [39] need to be trained in a supervised way with the costly paired images. This is a consensus for us that image pairs are difficult to obtain. Unlike [15, 39] and [22, 41], CycleGAN [52], UINT [26], MUINT [14], and UGATIT [20] can realize cross-domain transformation in an unsupervised manner, e.g., style transfer [17, 52]. Other work like SRGAN [22], ESRGAN [41] translates the image in low resolution domain into high resolution domain in a supervised way too.

2.3 Attention-based method

After the attention mechanism was introduced into deep learning, it has been favored by researchers. Attention mechanism improves the performance of networks by mimicking the way humans observe things, selectively focusing on more informative things. The non-local network [40] and self-attention GAN [45] employed in computer vision guide the network to concentrate on the regions of interest by simulating long-range dependency. [12, 50] improve the learning of useful features by learning

the importance of different channel features, and suppress the less valuable features of tasks. The former [40, 45] is often referred to as spatial attention, while the latter [12, 50] is referred to as channel attention. There are also some networks that combine spatial attention and channel attention, for instance, [4, 7, 43], which use the above two modules in series or in parallel to acquire regions of interest. Attention mechanism has been prevalently visualized in classification [12, 43], localization [45], semantic segmentation [7], image generation [45], object detection [43], image caption [4], and other tasks of computer vision [5, 6, 18].

3 Image generation

The proposed method is decomposed into two stages and each stage has different missions. Detailed network architecture will be introduced in Sects. 3.1 and 3.2.

3.1 Coarse image generation

3.1.1 Network architecture

On the whole, all image-to-image translation problems can be summarized as learning a mapping from input image x to output image y , namely $y = f(x)$, where $f(\cdot)$ represents a mapping function. Like most other methods, the generator's overall framework in the first stage is the prevalently adopted network: the encoder-decoder structure, as shown in Fig. 1. The model image is first encoded by the encoder into a latent representation (in this paper, a vector with 512-dimensional), and then, the latent representation is decoded by the decoder into the desired result. But different from traditional auto-encoders, we add skip connections between the encoder and the decoder, which is also termed as a 'U-net'-based [34] structure. In the most image translation tasks, there is a large quantity of shared low-level information between input and output such as color, shape, etc. In our mission, the clothes on the model in the input picture are the results we interested in, so shuttling appearance information which was shared by the input and the output between the mirror layer of the encoder and the decoder can improve the learning ability of the network. From some practical results, a simple skip connection does demonstrate its superiority for some images in good poses and the model pictures (as shown in Fig. 6: in the fifth column at the top and the fourth column at bottom) that have little difference in perspective from the output picture (front view). However, we observe that the network performs unsatisfactorily when the model image was in a side view. This is mainly due to the lack

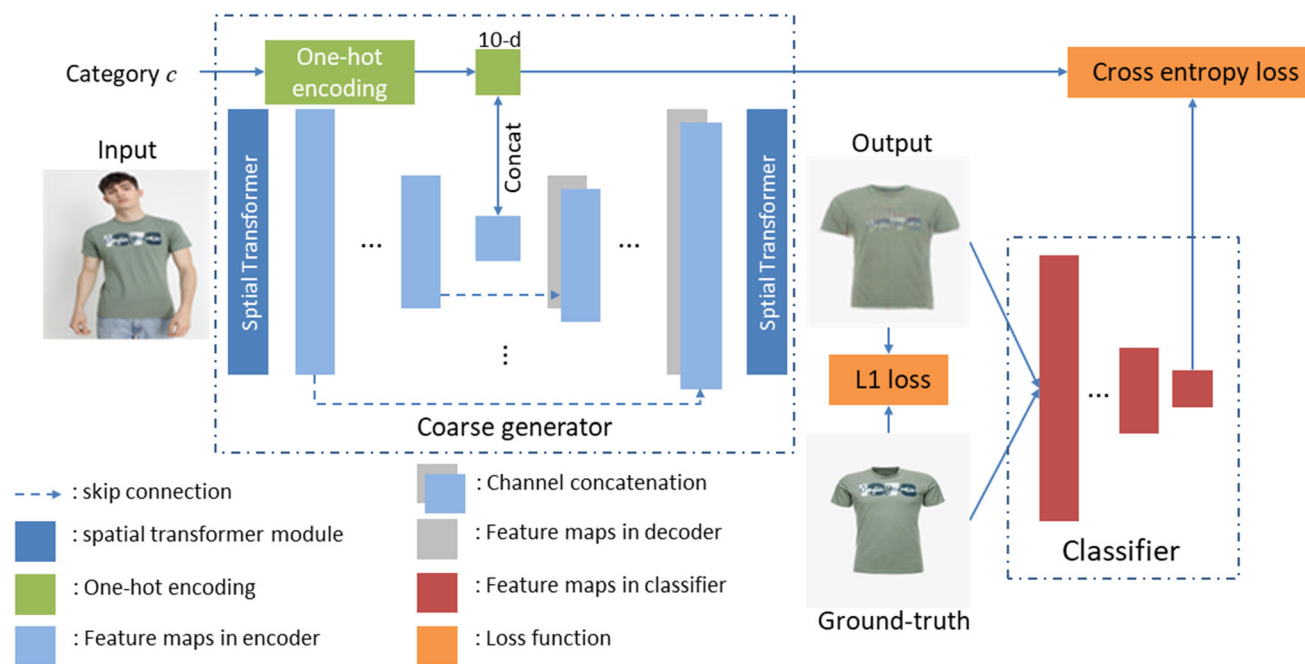


Fig. 1 Architecture of network in the first stage

ability of being spatially invariant [16] to the model image in convolutional neural networks.

In order to tackle this problem, we propose to introduce a learnable spatial transformer module into the network, which can deal with the deformation or migration problem during image translation in a way of a relatively low parameters and computational cost. The architecture of the spatial transformer module is visualized in Fig. 2. It mainly consists of three parts: a localization network, a grid generator, and a bilinear sampler. The localization network learns the parameter θ according to the input data (in our case, θ is a 6-dimensional vector), and the parameter is transmitted to the next part to generate a transformation grid. The last bilinear sampler is conditioned on generated grid and the input data, and the output is obtained by bilinear interpolation. This module not only transforms images but also feature maps. What’s more, inserting this

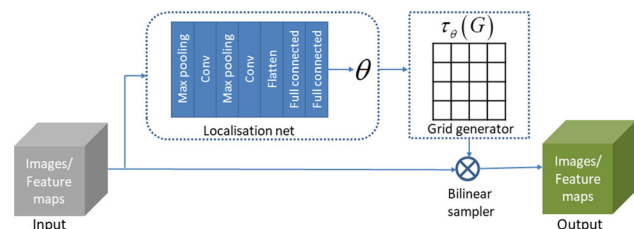


Fig. 2 An overview of spatial transformer module. θ is the learnable transformation parameters. Input and output are images or feature maps

module into any networks does not affect the gradient backpropagation because the bilinear sampler is differentiable and does not require an additional loss to supervise the transformation parameter θ . In this paper, we insert a spatial transformer module in the first and last layers of the network, as shown in the blue rectangle in Fig. 1.

Not only that, in our data set, there are not only the upper body or lower body images of only one class of clothing (as shown in Fig. 6: the third and fourth column at the top), but also the pictures of the whole body of the model (as shown in Fig. 6: the first column at the top and the third column at the bottom), which means there are two or more types of clothing. For this situation, the network is usually confused, and does not know which type of clothing image to generate.

Therefore, as for the lack of supervision of the generated clothing category, we propose that the input of the network not only needs to include the model image, but also the clothing category we expect. Add an additional supervisory signal to the generator to encourage the generator to synthesize the target in a guided way. Therefore, we consider converting each type of clothing into an n -dimensional vector in the form of one-hot coding (the i -th class: the i -th dimension is 1, other dimensions are all 0, where $0 < i < n$). We concatenate the category coding before the encoded latent vector is sent to the decoder to direct the decoder to decode the particular category of clothing.

3.1.2 Objective function

For image translation problems, traditional metrics that measure the discrepancy between pixel values, such as the Manhattan distance (L_1 distance) and the Euclidean distance (L_2 distance), can fully satisfy our needs for generating rough images in the first stage. After comparison, the L_1 distance, which is less susceptible to blurring, is employed to constrain the deviation between pixels. Besides, in order to generate the correct category of clothing, the cross-entropy loss that is widely adopted in multi-label classification tasks is introduced into the entire loss function of the first stage. Suppose that the generator of in the first stage is G_1 , classifier is C , the input image is denoted as x , the category condition is denoted as c , and the ground-truth is expressed as y , the result of the first stage y_{co} can be expressed as:

$$y_{co} = G_1(x, c). \quad (1)$$

Therefore, the content difference $L_{content1}$ between the generated result and the target picture can be formulated as:

$$L_{content1} = \|y - y_{co}\|. \quad (2)$$

Classification loss L_{cls} can be calculated as:

$$L_{cls} = \sum_{i=1}^n \left[c_{co}^{(i)} \log c^{(i)} + (1 - c_{co}^{(i)}) \log (1 - c^{(i)}) \right]. \quad (3)$$

where n is the number of categories in the multi-label tasks, in this paper, is predefined as 10, and $c^{(i)}$ and $c_{co}^{(i)}$ is one-hot coding and predicted category vector by classifier of the i -th sample. Our objective function can be expressed as:

$$G_1^* = \operatorname{argmin}_{G_1} [\mu L_{content1}(G_1) + L_{cls}(G_1, C)], \quad (4)$$

where μ is the weight of the content loss. By optimizing the loss function described above, it is possible to obtain a coarse image of the desired clothing with the category we are interested in, as the appearance condition information for the next stage guiding the network to generate fine images.

3.2 Fine image generation

3.2.1 Network architecture

Since the coarse picture in the first stage only has a rough shape and appearance, the second stage refines the rough picture. The goal is to synthesize clear, photo-realistic clothing images, and it is a preferable choice to achieve this goal in a technique that is adversarial learning. In many different practical applications, [15] has demonstrated its

feasibility in dealing with image translation problems. In the second stage, we use the pix2pix framework as baseline, and make some modifications to it. The detailed structure is shown in Fig. 3.

In our method, there are two images that are fed to the fine generator, one is the coarse image y_{co} generated in the first stage containing the appearance information and the expected category, and the other is the original input x , which is utilized to provide details that y_{co} does not have, and then to refine y_{co} . We propose a dual path attention-based generator which contains three sub-modules: input encoder, coarse encoder, and decoder. The structure of coarse encoder is the same as that of input encoder.

However, skip connections between coarse encoder and decoder are different from input encoder and decoder. If input encoder transmit the information to the image decoding layer directly like the first stage, which inevitably contains a large amount of redundant information (useless information like the human body in our mission). When redundant information and the desired information such as the high frequency information we want are transmitted to the mirror layer with equal importance, this will unavoidably impair the performance of the network greatly. Therefore, so as to make the network more focused on the regions of interest, we introduce a channel attention module instead of the original skip connections between input encoder and decoder while coarse encoder is directly connected to decoder. The channel attention module [50] assigns different weights to different channels of feature maps according to the learned coefficient. The module learns that assigning higher weights to those informative channels to represent the information of that channel is of higher significance. The structure of the channel attention module is shown in Fig. 4.

Furthermore, most of the previous GAN-based methods are of relatively small kernel size (no more than 5) for lower parameter and higher efficiency. In this way, it is hard to learn the relationship between the parts that are far away. Learning the global dependencies of images enables the network to know what to generate. Therefore, we insert a self-attention module into decoder to learn the relationship between any two pixels. The structure of the self-attention module is shown in Fig. 5.

For the discriminator, we use the same Markovian discriminator as [15] whose input is the concatenation of input x and fine output or ground-truth. It does not judge real and fake of the whole picture like the conventional discriminator. Instead, it divides the picture into many small patches, and judges the real and fake of each patch. The realness of the whole picture depends on the average result of all the patches, so it is also known as PatchGAN.

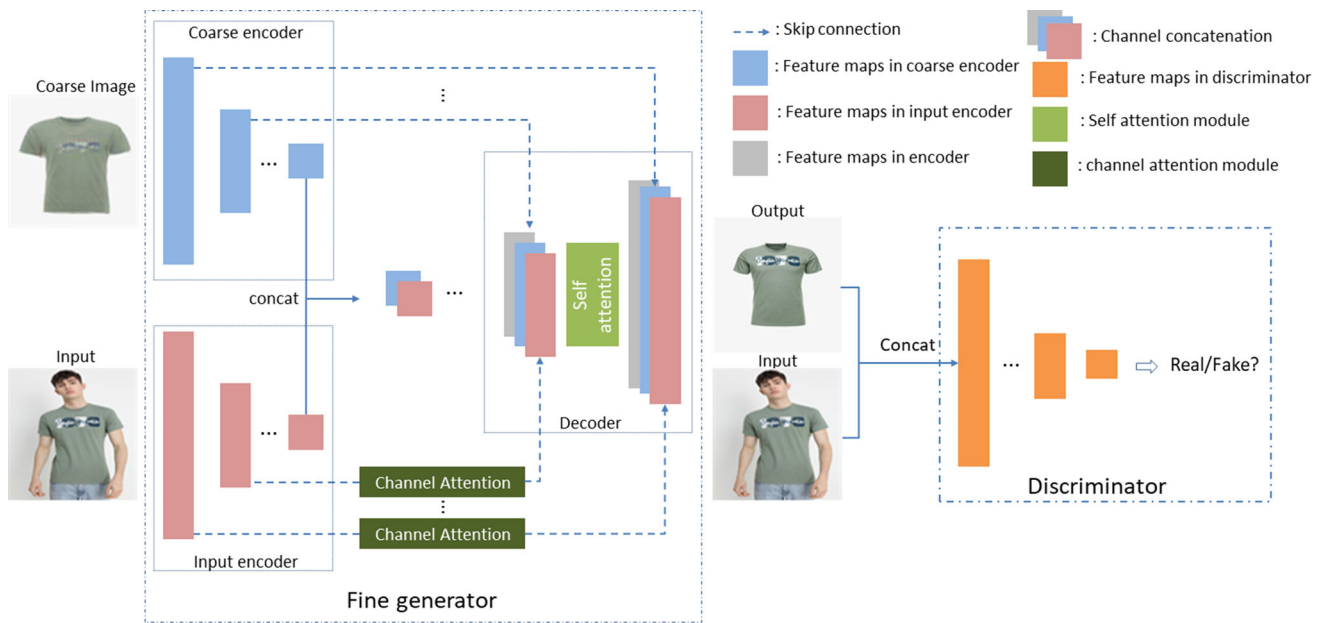


Fig. 3 An overview architecture of our proposed dual path attention-based generator and Markovian discriminator in the second stage. The generator incorporates three sub-modules: coarse encoder, input encoder, and decoder

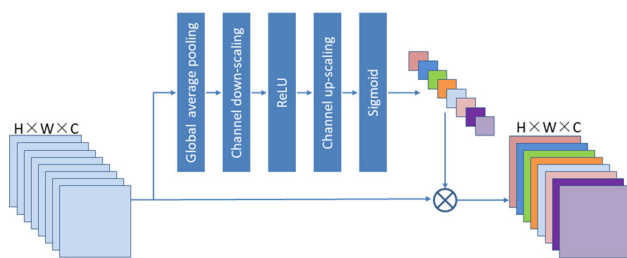


Fig. 4 Architecture of channel attention module, where input is feature maps within input encoder and output feature maps are fed into mirrored layer within decoder, \otimes indicates matrix multiplication

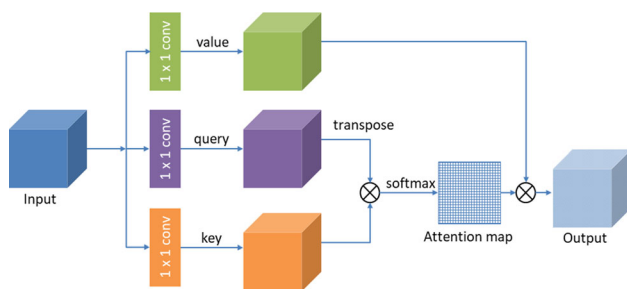


Fig. 5 Architecture of self-attention module, where input is feature maps within decoder with length and width of 64, \otimes represents matrix multiplication. Attention map is activated with softmax on each row

3.2.2 Objective function

Assume that the generator of in the second stage is G_2 , discriminator is D , the result of the second stage is y_{fi} , which can be expressed as:

$$y_{fi} = G_2(x, y_{co}). \tag{5}$$

In addition to the original GAN loss function, like the first phase, l_1 norm is also part of the overall loss function. The loss function of GAN can be expressed as:

$$L_{GAN} = \log D(x, y_{fi}) + \log(1 - D(x, y_{fi})). \tag{6}$$

Content loss is represented as:

$$L_{content2} = \|y - y_{fi}\|. \tag{7}$$

Furthermore, the perceptual loss [17, 22] L_{pe} is added to overcome the ambiguity caused by l_1 loss on the generated results. The perceptual loss measures the difference between the output of the network and ground-truth in the high-dimensional feature space and optimizing perceptual loss encourages the network to generate more high-frequency details. We use the pre-trained VGG-19 [36] model on the ImageNet as a high-dimensional feature space extractor to calculate the absolute difference (in L_1 sense) between the two extracted feature maps. So, the perceptual loss L_{pe} can be expressed as:

$$L_{pe} = \left\| \text{VGG}(y)_{i,j} - \text{VGG}(y_{fi})_{i,j} \right\|, \tag{8}$$

where $\text{VGG}(\cdot)_{i,j}$ represents feature maps obtained from the j -th convolution layer after rectified linear unit before the i -th pooling layer within VGG-19 network. The whole loss function can be expressed as:

$$L_{total} = \alpha L_{GAN} + \beta L_{content2} + L_{pe}, \tag{9}$$

where α and β are hyper parameter, representing the weight of each loss.

4 Simulations

4.1 Data set description and simulation settings

4.1.1 Data set description

To our best knowledge, the existing open clothing data set such as the comprehensive data set: DeepFashion [27] and the multi-view clothing data set: MVC [23] are unable to meet our requirements: paired model pictures and corresponding tiled clothes on the model. Therefore, for the purpose of implementing our experiment, we have to surf the Internet to collect garment images through a web crawler program and build a new data set by our own. After comparing the pictures of large-scale e-commerce platforms at home and abroad such as Zalando, Taobao, Jingdong, we selected the pictures on the German e-commerce platform as our data set because of its high quality and variety. Then, those images that satisfy our experimental needs were screened out and the corresponding category tags are labeled manually. After the above process, our final data set consists of 34,762 pairs of images, of which 32,747 pairs are employed in training and the remaining image pairs are regarded as testing set. In order to enhance the robustness of the network, we select 10 types of garment in this newly built data set, namely T-shirts, shirts, jackets, coats, blouses, jumpsuits, jeans, trousers, skirts, and dresses. At the time, we have the quantity of each kind of clothing balanced to obviate the negative impact that samples imbalance will bring about on the experimental results, so the amount of each type of sample pairs is controlled artificially more than 1000. The number of each category and some sample pairs are shown in Fig. 6.

4.1.2 Implementation

The training and testing of this paper is implemented on Tensorflow with a NVIDIA TITAN V GPU. Similar with other GANs, Adam optimizer, the prevalently used optimizer in GANs is employed to optimize loss functions with a learning rate of 0.0002 and a momentum of 0.5. All weight parameters are initialized with normal distribution whose mean value is 0 and standard deviation is 0.02. Batch size is set as 1 to ensure a one-to-one correspondence between input and output. All image pairs in training set are trained for 200 epochs. And in order to avoid overfitting, we utilize random clipping to 256×256 for data

augmentation after resizing to 286×286 and dropout on the last few layers of the decoder in both two stages. We choose feature maps in VGG19(\cdot)_{5,4}. Hyper parameter μ in the first stage is defined as 100. Different weight of each loss α and β are set as 1 and 100. The self-attention module is located between the second to last and the third to last decoding layers within decoder in the fine generator, with the input feature map size of 64×64 .

4.1.3 Evaluation metrics

Assessing the performance a generative model such as GANs is an intractable mission. There is no standard evaluation metric to measure the performance of GANs since GAN proposed. The majority of initial evaluations are based on subjective feelings. For traditional GAN models, the quality and diversity of the generated images is equally important. The subsequent Inception Score (IS) [35] and Fréchet Inception Distance (FID) [11] are two widely used indicators. FID is obtained by calculating the Fréchet distance between two Gaussian distributions simulated by output and ground-truth mean and covariance. In contrast to Inception Score, Fréchet Inception Distance is more sensitive to mode collapse and is more robust to noise. Therefore, FID is regarded as one of evaluation metrics for its ascendancy which can be formulated as:

$$FID(y, y_{fi}) = \|\mu_y - \mu_{y_{fi}}\|_2^2 + \text{Tr} \left[\mathbf{C}_y + \mathbf{C}_{y_{fi}} + 2(\mathbf{C}_y \mathbf{C}_{y_{fi}})^{\frac{1}{2}} \right] \tag{10}$$

where $\mu_{(\cdot)}$ represents mean value, $\mathbf{C}_{(\cdot)}$ represents covariance, Tr means trace in linear algebra.











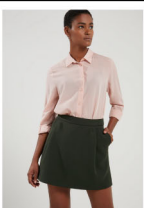





However, for conditional GAN, not only the quality of the generated image is required to be high, but also the conditions are met. Our objective is to acquire garment images that are as similar as possible to ground-truth. Consequently, we employ another evaluation metric structural similarity (SSIM) [42] which measures image similarity from the aspects of brightness, contrast, and structure, and can better express the subjective feelings of individuals. It can be expressed as:

$$\begin{aligned} L(y, y_{fi}) &= \frac{2\mu_y \mu_{y_{fi}} + C_1}{\mu_y^2 + \mu_{y_{fi}}^2 + C_1} \\ C(y, y_{fi}) &= \frac{2\sigma_y \sigma_{y_{fi}} + C_2}{\sigma_y^2 + \sigma_{y_{fi}}^2 + C_2} \\ S(y, y_{fi}) &= \frac{\sigma_{yy_{fi}} + C_3}{\sigma_y \sigma_{y_{fi}} + C_3} \end{aligned} \tag{11}$$

$$SSIM(y, y_{fi}) = L(y, y_{fi}) \cdot C(y, y_{fi}) \cdot S(y, y_{fi})$$

where $\mu_y, \mu_{y_{fi}}, \sigma_y, \sigma_{y_{fi}}, \sigma_{yy_{fi}}$ indicate the mean of ground-truth and output, the variance of ground-truth and output, the covariance of ground-truth and output, respectively, $C_1,$

Fig. 6 The number of different kinds of clothing in our data set. The quantity of different categories of clothing is controlled as much as possible

Cat.	T-shirts #3982	Shirts #3008	Jackets #4053	Coats #3793	Blouses #4181
Model					
Clothing					
Cat.	Jeans #3639	Trousers #3421	Skirts #3528	Jumpsuits #1315	Dresses #1659
Model					
Clothing					

C_2 and C_3 are predefined constant. In our experiments, hyper parameters C_1 , C_2 and C_3 are defined as 6.5025, 58.5225, and 29.26125, respectively. And $L(y, y_{fi})$, $C(y, y_{fi})$ and $S(y, y_{fi})$ indicate brightness, contrast, and structure similarity. SSIM, with values between 0 and 1, is positively correlated with image similarity, which means that the higher the similarity, the closer SSIM is to 1.

4.2 Comparisons

For the purpose of demonstrating the superiority of our proposed framework, we conduct some additional experiments on the same data set with other alternative methods. We compare our results with auto-encoder, pix2pix, conditional GAN, CatGAN [47], and pix2pixHD qualitatively and quantitatively.

4.2.1 Qualitative results

The comparison results of different generative models can be visualized in Fig. 7. As is shown in Fig. 7, auto-encoder (Column 3) can only produce blurry images similar to the approximate shape of the target clothing for the reason that it doesn't how to render the details of apparel. Our method in first stage (Column 8) modifies an encoder-encoder structure by adding a spatial transformer module, and it is not difficult to observe that our coarse generator can not only obtain the rough shape, what's more, but also produce more details. Approaches with adversarial learning generate shaper edges and create more details, but also suffer defective artifacts, which lead to unnatural pictures. Pix2-pix (Column 4) produces garment images with more sharp details compared to auto-encoder while generating clothing of the wrong category sometimes, as is visualized in row 6. Results of methods such as conditional GAN (Column 5)

Cat.	Input	GT	AE	cGAN	pix2pix	CatGAN	pix2pixHD	Stage1	Stage2
T-shirts									
Shirts									
Jackets									
Coats									
Blouses									
Jeans									
Trousers									
Skirts									
Jumpsuits									
Dresses									

Fig. 7 Qualitative comparison results of different methods, where AE represents auto-encoder, GT denotes ground-truth, HD denotes pix2pixHD. The column of stage2 is our final results

and CatGAN (Column 6) manifest that though those means generate garment items in correct category, large amounts of texture on clothes are distorted or even lost. Pix2pixHD (Column 7) is a leading method which synthesizes high resolution images while retaining abundant undistorted texture. However, pix2pixHD does not input any category supervision information which may result in generating clothing in an incorrect category. In contrast to precedent methods, our newly proposed model (last column) generates detailed tiled results correctly and preserves regular texture approximate to the target, demonstrating the effectiveness of our novel network.

4.2.2 Quantitative results

Our experimental results are also quantitatively compared using the metrics FID and SSIM. As is shown in Table 1, our method is significantly ahead of other methods with FID and SSIM but lags marginally behind pix2pixHD with SSIM. Compared with pix2pixHD, however, our training time is less than three quarters of its time with similar performance. This also indicates the efficiency of our approach.

4.2.3 Generating specific category clothing

We check whether this generative model can really output clothing according to the category in model images with multiple categories. As visualized in Fig. 8, as for hard examples with multiple categories of clothing, our method is capable of generating clothing based on the category, but multiple types of clothing also inescapably have negative consequences for our results to some extent. Consequently, it is obviously inspected that the most generated clothing pictures have recognizable shape but only few details and texture.

4.2.4 Ablation study

In this part, we explore how the absence of spatial transformer module and channel attention module in our network will affect the performance of our method. Also, we attempt to train our method in an end-to-end manner for convenience, and compare the results of the two training methods. All comparison results are summarized in Table 2.

w/o spatial transformer module (STM) In this experiment, the spatial transformer module is not taken into consideration to investigate the function of spatial transformer module.

w/o channel attention module (CAM) In addition, this portion investigates the impact of channel attention module on our method by shuttling information directly through skip connections.

w/o self-attention module (SAM) Besides, this experiment explores the impact of self-attention module on our method. We compare the gap between our model with SAM and without SAM.

one stage versus two stages We integrate this two-stage method into a one-stage way to analyze the performance of single-stage and multi-stage methods or whether the single-stage approach will have a negative impact on the network.

As shown in Table 2, the absence of any of the three modules has somewhat damage to the generated results. In contrast with the absence of SAM and CAM in the second stage, the disappearance of STM in the first stage has less negative impact on the experimental results. However, there is a significant decline in quantitative results of our framework without all three aforementioned modules, no matter in FID or SSIM. In view that the cumbersome training manner of multi-stage techniques, we integrate three modules into a single-stage method empirically and make an extra survey. The experimental result shows that mode collapse occurs in the integrated one-stage method, which directly leads to excessive FID value. These ablation studies confirm the novelty and effectiveness of our approach.

4.2.5 Discussion on different categories

By observing the experimental results, we find that the generation performance of different types of clothing on our model is uneven. In order to explore the differences in the results test on data set between different types of clothing in our method, the test set is divided into 10 portions according to their category. The quantitative discrepancy between different categories of clothing is shown in Table 3 and plotted in Fig. 9.

From those quantitative statistics, we can draw a conclusion that jeans and trousers, which have less features,

Table 1 Quantitative comparison results between other methods and our method

Methods	AE	cGAN	pix2pix	CatGAN	HD	Stage1	Stage2
FID	1.438	0.341	0.280	0.273	0.257	0.365	0.207
SSIM	0.503	0.581	0.683	0.680	0.774	0.707	0.771

HD represents pix2pixHD. Bold indicates the best result



Fig. 8 Results of the same input with different category. Upper and lower are general terms used to simplify the drawing. In our experiment, the input category is one of all 10 categories

Table 2 Ablation studies on spatial transformer module, self-attention module, and channel attention module

	FID	SSIM
Ours w/o STM CAM SAM	0.312	0.691
Ours w/o STM	0.223	0.757
Ours w/o CAM	0.270	0.743
Ours w/o SAM	0.248	0.750
One stage	2.496	0.585
Ours	0.207	0.771

In addition, the comparison between single stage and multi-stage of our approach is visualized in the fifth row and the sixth row. Bold indicates the best result

most of them are in the same shape with few patterns; perform well on both metrics, while other categories perform relatively not so well as jeans and trousers, which may be due to their different patterns on clothes that are intractable to learn. Blouses perform unsatisfactory on both metrics for the reason that most blouses not only have various patterns on it, but also range in diverse shapes, which undermines the performance.

Table 3 SSIM and FID measured on different kinds of clothing in our data set

	T-shirts	Shirts	Jackets	Coats	Blouses	Jeans	Trousers	Skirts	Jumpsuits	Dresses
SSIM	0.710	0.736	0.777	0.817	0.683	0.810	0.789	0.824	0.796	0.809
FID	0.334	0.347	0.418	0.490	0.513	0.134	0.361	0.409	0.346	0.399

Bold indicates the best result of different categories

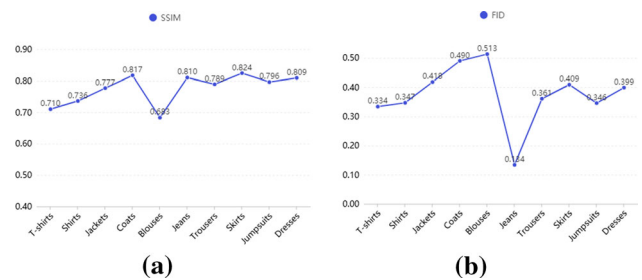


Fig. 9 SSIM and FID line chart of different types of clothing

4.3 Application

Virtual try-on is an inevitable choice for the future development of the fashion industry, which not only enables consumers who purchase apparel online to try-on clothing, but also allows customers to know the appearance after trying on without undressing in the real store, significantly improving customers’ experience. Virtual try-on can be considered as an extension of our work, the state-of-the-art virtual dressing method called cp-vton [38] is employed to clarify the superiority and necessity of our method, and our method is of high value in practical environment.

The cp-vton is a two-stage virtual dressing method, the first stage is to learn a mapping from the target garment

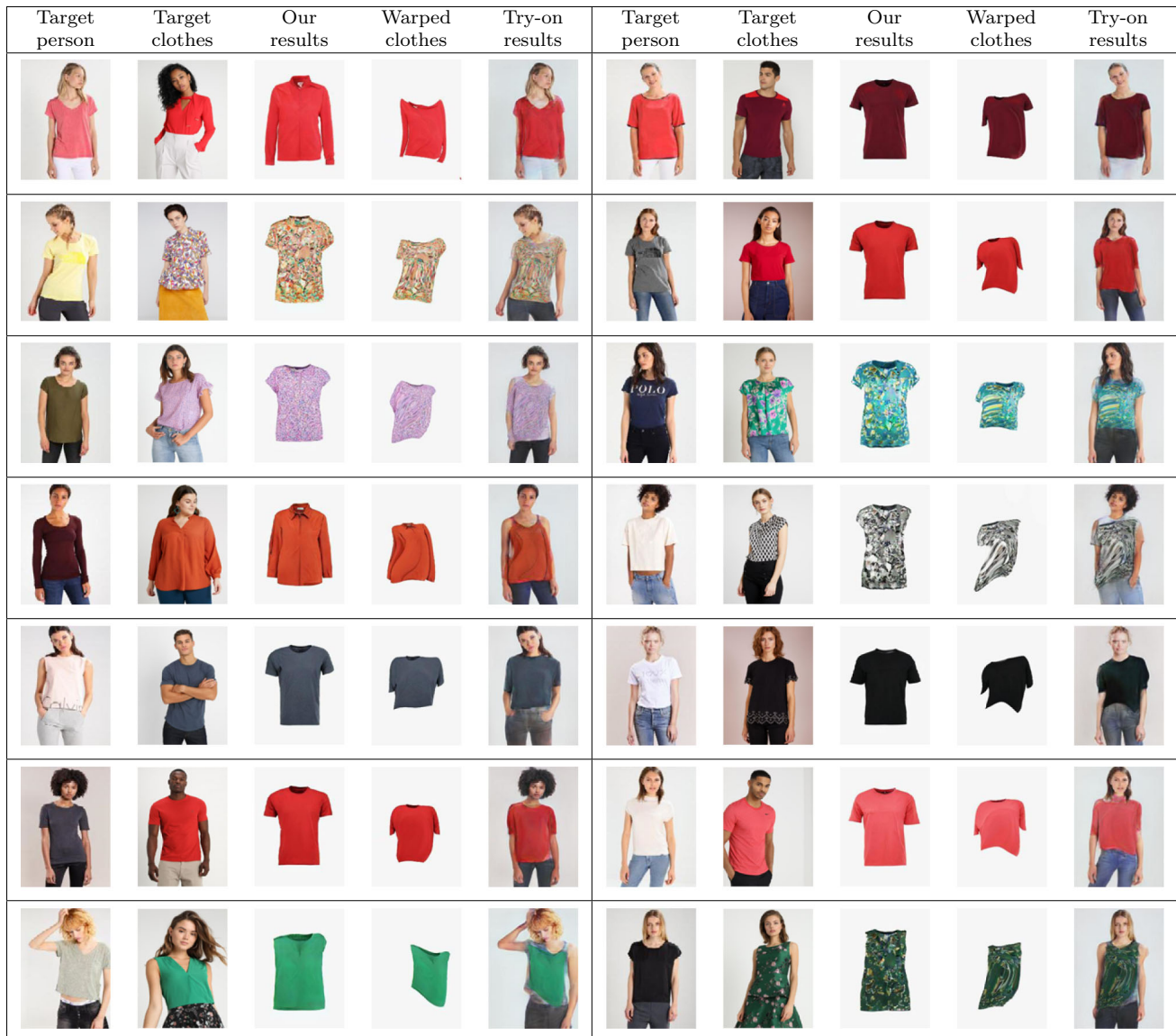


Fig. 10 Results of the same input with different categories. Upper and lower are general terms used to simplify the drawing. In our experiment, the input category is one of all 10 categories

picture to a warped one, which makes the target image has the same shape with clothing on person, and then, utilizing synthetic warped clothing and person representation as input to refine at the second stage. The result images after virtual try-on is displayed in Fig. 10. The fourth and eighth columns are try-on results test on our generated tiled clothes, which further proves the feasibility of our method.

5 Conclusion

This paper explores a novel two-stage solution toward image-to-image translation from model to tiled clothing. The first stage introduces a spatial transformer module and

a classifier, which manages to generate coarse results in specified input category and preserve as much appearance information as possible. At the second stage, channel attention modules and the self-attention module are inserted into fine generator which enables fine generator to concentrate on informative parts and employing the adversarial learning fashion generates sharper details. Comprehensive experiments conducted on our newly built data set demonstrate the overall framework accurately synthesizes high-fidelity garment images that conserve texture of input without much distortion. Our approach is capable of achieving similar performance to state-of-the-art supervised image-to-image translation method but takes less training time. In conclusion, our network achieves

excellent performance both quantitatively and qualitatively on our data set and is an effective and efficient scheme.

Acknowledgements This work is partially supported by National Key Research and Development Program of China (2019YFC1521300), supported by National Natural Science Foundation of China (61971121, 61672365), supported by the Fundamental Research Funds for the Central Universities of China and DHU Distinguished Young Professor Program, and also supported by the Fundamental Research Funds for the Central Universities of China (JZ2019HGPA0102).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Arjovsky M, Chintala S, Bottou L (2017) Wasserstein gan. arXiv preprint [arXiv:1701.07875](https://arxiv.org/abs/1701.07875)
- Berthelot D, Schumm T, Metz L (2017) Began: Boundary equilibrium generative adversarial networks. arXiv preprint [arXiv:1703.10717](https://arxiv.org/abs/1703.10717)
- Brock A, Donahue J, Simonyan K (2018) Large scale gan training for high fidelity natural image synthesis. arXiv preprint [arXiv:1809.11096](https://arxiv.org/abs/1809.11096)
- Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS (2017) Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5659–5667
- Fan J, Chow TW (2019) Exactly robust kernel principal component analysis. *IEEE Trans Neural Netw Learn Syst*
- Fan J, Udell M (2019) Online high rank matrix completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8690–8698
- Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3146–3154
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
- Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. In: Advances in neural information processing systems, pp 5767–5777
- Han X, Wu Z, Wu Z, Yu R, Davis LS (2018) Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7543–7552
- Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637
- Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
- Huang X, Li Y, Poursaeed O, Hopcroft J, Belongie S (2017) Stacked generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5077–5086
- Huang X, Liu MY, Belongie S, Kautz J (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 172–189
- Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
- Jaderberg M, Simonyan K, Zisserman A, et al (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
- Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. Springer, Berlin, pp 694–711
- Kang Z, Pan H, Hoi SC, Xu Z (2019) Robust graph learning from noisy data. *IEEE Trans Cybern*
- Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint [arXiv:1710.10196](https://arxiv.org/abs/1710.10196)
- Kim J, Kim M, Kang H, Lee K (2019) U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint [arXiv:1907.10830](https://arxiv.org/abs/1907.10830)
- Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114)
- Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
- Liu KH, Chen TY, Chen CS (2016) Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval, pp 313–316. ACM
- Liu L, Zhang H, Ji Y, Wu QJ (2019) Toward ai fashion design: an attribute-gan model for clothing match. *Neurocomputing* 341:156–167
- Liu L, Zhang H, Xu X, Zhang Z, Yan S (2019) Collocating clothes with generative adversarial networks cosupervised by categories and attributes: a multidiscriminator framework. *IEEE Trans Neural Netw Learn Syst*
- Liu MY, Breuel T, Kautz J (2017) Unsupervised image-to-image translation networks. In: Advances in neural information processing systems, pp 700–708
- Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1096–1104
- Ma J, Zhang H, Yi P, Wang ZY (2019) SCSCN: a separated channel-spatial convolution net with attention for single-view reconstruction. *IEEE Trans Ind Electron*
- Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802
- Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint [arXiv:1411.1784](https://arxiv.org/abs/1411.1784)
- Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th international conference on machine learning, vol 70, pp 2642–2651. JMLR. org
- Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434)
- Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. arXiv preprint [arXiv:1605.05396](https://arxiv.org/abs/1605.05396)

34. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 234–241
35. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford, A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems, pp 2234–2242
36. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv: 1409.1556](https://arxiv.org/abs/1409.1556)
37. Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems, pp 3483–3491
38. Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M (2018) Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV), pp 589–604
39. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B (2018) High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8798–8807
40. Wang X, Girshick RB, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
41. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C (2018) Esrgan: enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV)
42. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP et al (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process.* 13(4):600–612
43. Woo S, Park J, Lee JY, So Kweon I (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
44. Xu H, Liang P, Yu W, Jiang J, Ma J (2019) Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In: Proceedings of international joint conference artificial intelligence, pp 3954–3960
45. Zhang H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks. arXiv preprint [arXiv: 1805.08318](https://arxiv.org/abs/1805.08318)
46. Zhang H, Ji Y, Huang W, Liu L (2019) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. *Neural Comput Appl* 31(11):7361–7380
47. Zhang H, Sun Y, Liu L, Wang X, Li L, Liu W (2018) Clothing-out: a category-supervised gan model for clothing segmentation and retrieval. *Neural Comput Appl*, pp 1–12
48. Zhang H, Sun Y, Liu L, Xu X (2019) Cascadegan: a category-supervised cascading generative adversarial network for clothes translation from the human body to tiled images. *Neurocomputing*
49. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2017) Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 5907–5915
50. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
51. Zhu H, Cheng Y, Peng X, Zhou JT, Kang Z, Lu S, Fang Z, Li L, Lim JH (2019) Single-image dehazing via compositional adversarial network. *IEEE Trans Cybern*
52. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
53. Zhu S, Urtasun R, Fidler S, Lin D, Change Loy C (2017) Be your own prada: fashion synthesis with structural coherence. In: Proceedings of the IEEE international conference on computer vision, pp 1680–1688

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.