DESIGN-GAN: CROSS-CATEGORY FASHION TRANSLATION DRIVEN BY LANDMARK ATTENTION

Yining Lang[1], *Yuan He*[1], *Jianfeng Dong*[2,3,†], *Fan Yang*[1], *Hui Xue*[1]

[1] Alibaba Group [2] Zhejiang Gongshang Univresity
[3] Alibaba-Zhejiang University Joint Institute of Frontier Technologies
louis.lyn, heyuan.hy, xizhuo.yf, hui.xueh@alibaba-inc.com
[†] dongjf24@gmail.com

ABSTRACT

The rise of generative adversarial networks has boosted a vast interest in the field of fashion image-to-image translation. However, previous methods do not perform well in cross-category translation tasks, e.g., translating jeans to skirts in fashion images. The translated skirts are easier to lose the detail texture of the jeans, and the generated legs or arms often look unnatural. In this paper, we propose a novel approach, called DesignGAN, that utilizes the landmark guided attention and a similarity constraint mechanism to achieve fashion cross-category translation. Moreover, we can achieve texture editing on any customized input, which can even be used as an effective way to empower fashion designers. Experiments on fashion datasets verify that DesignGAN is superior to other image-to-image translation methods.

Index Terms— DesignGAN, Fashion Translation, Landmark Attention, Texture Editing.

1. INTRODUCTION

Generative Adversarial Networks (GAN) [1] have shown impressive results in image synthesis [2, 3, 4, 5] and video generation [6, 7, 8]. More recently, the image-to-image translation which aims to map an image from a source domain to a target domain received considerable attention in the computer vision community. However, it requires paired images from each of the source and target domains, which limits the training. Unsupervised approaches [9, 10, 11] overcome this problem with cyclic losses which encourage the translated domain to be faithfully reconstructed when mapped back to the original domain. Due to the reduction of restrictions, the performance in practical applications like semantic manipulation [12, 13, 14], super-resolution [15, 16, 17], and domain adaptation[18, 19] has enhanced by a large margin.

Previous methods [13, 20, 9, 21, 22] mainly involve the translation of textures (e.g., from summer to winter) or shapes [23]. We found that they often fail on cross-category translation tasks, more specifically, the task involves significant changes in shapes. Recently, Mo et al. [24] proposed the InstaGAN that can facilitate the shape transformation (e.g., from jeans to skirts), which greatly promoted this research filed. But the resulting skirts are easier to lose the texture details of jeans, and on the other hand, the synthesized texture of legs or arms often looks unnatural, as shown in Fig.1.

To this end, we propose a novel approach that incorporates the landmark guided attention and a similarity constraint mechanism to improve cross-category fashion translation, based on the framework of CycleGAN [9]. Inspired by the previous human image generation works [25, 26], we utilize the predicted landmarks of both human skeleton and clothes as the attention to enhance the discrimination ability. To the best of our knowledge, we are the first ones who use the predicted landmarks of both human skeleton and clothes as the attention to guide the fashion-related generation. Also, different from the InstaGAN which requires a ground-truth mask, our approach learns joint-mappings between two domains based on the landmark guided regions.

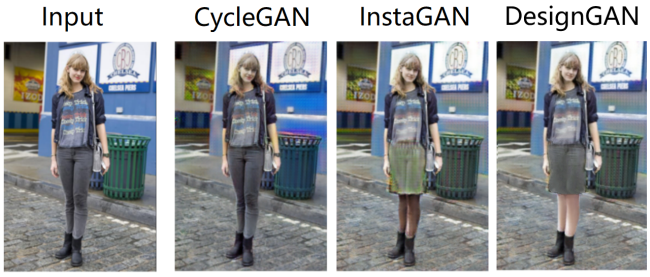

Fig. 1. Translation results of the prior work (CycleGAN[9], InstaGAN [24]), and our proposed approach, DesignGAN.

On the other hand, we introduce a similarity constraint mechanism to control the texture similarity of both clothes and skin between the generated images and the input samples within the translated region. Benefiting from the landmark guided attention and the similarity constraints mechanism, our method shows more impressive results for clothes category translation tasks compared to previous methods (e.g., CycleGAN[9], InstaGAN [24]).

Our approach can even customize the input texture when complete a translation. Giving an original image and a specific texture input, our approach can generate a translated clothes with target texture (e.g., cobble, dotted), which can not achieve by previous works (e.g., InstaGAN [24]). It can even be used as an effective way to empower fashion designers.

Our main contribution is three-fold: First, we propose a novel network DesignGAN, which can translate the fashion category guided by landmark attention. Second, we solve the problem that losing the details of clothes texture and showing the unnatural legs or arms after translation by introducing the texture similarity constraint mechanism. Finally, we can achieve texture editing on any customized input, which can strengthen its applicability up to a higher level.

† Corresponding Author: Jianfeng Dong (dongjf24@gmail.com).

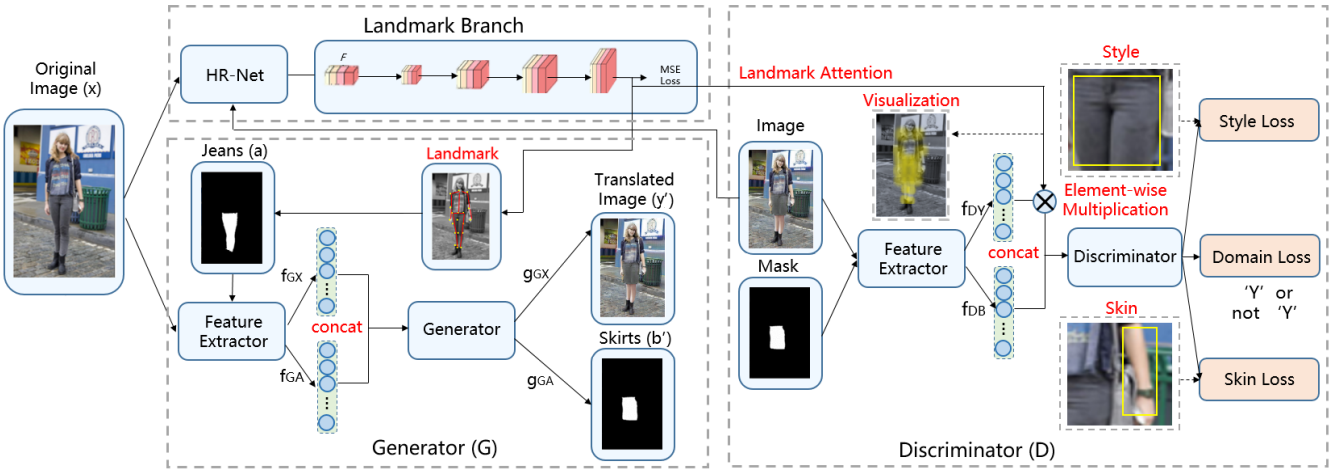

Fig. 2. The framework of the proposed DesignGAN. We adopted the network architectures of CycleGAN [9] as the building blocks. During the generation process, we added the landmark branch to predict the landmarks of both human skeleton and clothes in the input image x, which can guide the segmentation of the original category region (**a**). As for the discriminator D, we introduce a similarity mechanism to ensure the translation effect of both clothes and skin, achieved by the style loss and skin texture loss. Moreover, the landmark attention is introduced to strengthen the discrimination power of the model.

2. TECHNICAL APPROACH

2.1. Problem Formulation

The goal of unsupervised translation we tackle is to learn mappings across two image domains (X and Y) with unpaired samples. In particular, our approach learns joint-mappings between category-attached spaces ($X \times A$ and $Y \times B$) based on the landmark guided regions. More specifically, translating the concated feature maps of the whole image ($x \in X$) and the category region ($\mathbf{a} \in A$) (e.g., the region of jeans), as illustrated in Fig.2. A and B represent the original and target category regions, respectively. On the other hand, the predicted landmarks are applied as an attention mechanism, which can further enhance its discrimination ability and conversely, improve generation quality.

To solve the problems of unsimilar texture of clothes and unnatural texture of the skin, we design a style loss and a skin loss to constrain the similarity within the corresponding regions. Also, the similarity constraint can be used to translate a customized input texture.

2.2. DesignGAN Architecture

Fig.2 illustrates the framework of our approach based on the CycleGAN [9] architecture. We first extract individual features from the original image x and landmark guided category region **a** (e.g., the region of jeans) using image feature extractor f_{GX} and category feature extractor f_{GA}, respectively. Then, we concatenate the image feature and category features guided by landmarks. The concated feature map is feed to the generator. We train two coupled generators $G_{XY}\colon X \times A \to Y \times B$ and $G_{YX}\colon Y \times B \to X \times A$, where G_{XY} translates the original data$(x, \mathbf{a})$ to the target domain $(y', \mathbf{b}')$, vice versa for G_{YX}.

On the other hand, our approach encodes both y and **b** (or y' and $\mathbf{b}'$), and determines whether the pair is from the domain or not (i.e., is skirt or not in the example) with adversarial discriminators. We also have two coupled discriminators $D_Y\colon Y \times B \to \{Y, \text{not } Y\}$ and $D_X\colon X \times A \to \{X, \text{not } X\}$, where D_Y determines if the data (original $(y, \mathbf{b})$ or translated $(y', \mathbf{b}')$) is in the target domain $Y \times B$ or not, vice versa for D_X.

The various parts of the architecture (e.g., extractors, generators) are not mandatory, which can be replaced by other networks that have the same effect.

2.3. Landmark Attention

The landmark branch predicts the landmarks of both human and clothes within the image simultaneously based on the HR-Net [29] backbone, as illustrated in Fig.2. The prediction ability for the human skeleton is trained on COCO [30] dataset, while the ability for clothes landmark is trained on Deepfashion2 [31] dataset. Note that they use the same type of backbone, but they do not share the model.

We transform the landmark estimation task to predicting k heatmaps, where each heatmap indicates the location confidence of the k-th landmark. After extracting the feature map by the HR-Net backbone, we use several groups of transposed convolution to produce a high-resolution landmark heatmap with the same scale as the input image and utilize a regressor to estimate the heatmaps where the landmark positions are chosen. The mean squared error loss function is applied for comparing the heatmaps between the ground-truth and the predicted results.

We use the predicted human skeleton and clothes landmarks to guide the segmentation of the category region (e.g., the jeans in Fig.2), which is different from InstaGAN [24] that requires a ground-truth mask as input. For instance, the collar region is surrounded (i.e., segmented) by several human and clothes landmarks around the neck. What's more, the predicted landmarks can be applied to the discrimination process as an attention mechanism to further enhance the training effect.

We introduce the landmark attention during the discrimination process by making $f'_{DY} = f_{DY} \circ f_{Att}$, where f_{DY} represents the extracted feature map of image y or y', f_{Att} represents the attention map, and $\circ$ stands for Hadamard product. By multiplicating with the attention map, the critical features are strengthened, while irrelevant features are filtered out, via different weighted elements. For instance, the landmarks around critical areas like the waistline and ankle can guide the extraction of features, which makes these key features have more possibility to retain. Thus, the discriminatory power of the model can be strengthened.

Fig. 3. (a) Translation results on multi-human parsing (MHP) [27] dataset. (b-d) Results on clothing co-parsing (CCP) [28] dataset.

2.4. Loss Function

The goal of our work is cross-category fashion translation while keeping the original contexts (e.g., the style of the clothes, the skin of the legs). To this end, we both consider the domain loss, which makes the generated outputs to follow the overall 'look' of the target domain, and the **style/skin** loss, which make the outputs keep the original texture of **clothes/skin**.

Following a similar scheme with our baseline model, CycleGAN [9], we use the LSGAN [32] loss as the adversarial loss, and consider both the cycle-consistency loss [33] and the identity mapping loss [34] as rough constraints to keep the overall content of the original image. We apply adversarial losses to both mapping functions. For the mapping function G_{XY} and its discriminator D_Y , we express the objective as:

$$\begin{aligned}\mathcal{L}_{LSGAN} = \mathbb{E}_{(y,\mathbf{b})\sim p_{data}}[(D_Y(y,\mathbf{b})-1)^2]+\\ \mathbb{E}_{(x,\mathbf{a})\sim p_{data}}[D_Y(G_{XY}(x,\mathbf{a}))^2],\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{cyc} = \mathbb{E}_{(x,\mathbf{a})\sim p_{data}}[||G_{YX}(G_{XY}(x,\mathbf{a}))-(x,\mathbf{a})||_1]+\\ \mathbb{E}_{(y,\mathbf{b})\sim p_{data}}[||G_{XY}(G_{YX}(y,\mathbf{b}))-(y,\mathbf{b})||_1],\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{idt} = \mathbb{E}_{(y,\mathbf{b})\sim p_{data}}[||G_{XY}(y,\mathbf{b})-(y,\mathbf{b})||_1]+\\ \mathbb{E}_{(x,\mathbf{a})\sim p_{data}}[||G_{YX}(x,\mathbf{a})-(x,\mathbf{a})||_1].\end{aligned} \tag{3}$$

We call the three loss functions above domain losses, as illustrated in figure 2. They achieve the overall structure of the translated clothes, but lack of texture details. Thus, we design two pixel-wise L2 loss functions: the style loss and the skin loss to penalizes the L2 difference between the RGB channels of the generated result and that of the input texture.

Since the size of the category region changes during the translation (e.g., jeans→skirt), the original region (**a**) is resized to the same size as the translated one (**b**′). The style loss is defined as follow:

$$\mathcal{L}_{style} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{c=1}^{3}||p(\mathbf{a})_{i,j}-p(\mathbf{b}')_{i,j}||_2\circ w_{style}, \tag{4}$$

where N represents the number of pixels within the translated region $\mathbf{b}'$ (e.g., the region of skirt), i and j indicate the coordination of each pixel. $p(\mathbf{a})$ and $p(\mathbf{b}')$ represents the value of channel c within the original region **a** and translated region $\mathbf{b}'$, respectively. $\circ$ stands for Hadamard product. w_{style} represents the weight map of style loss, in which the weights gradually increases from the center of the region $\mathbf{b}'$ to the edge, because the edge area has more detail than the center area.

In order to make the skin texture of the legs or arms in the translated image y' look more natural, we define a skin loss to constrain the similarity between the generated skin region $\mathbf{b}'_s$ and the original skin region $\mathbf{a}_s$. For the generated skin of legs, our approach utilizes the landmarks around the arms to obtain the original skin texture. (vice versa for the generation of arms). For instance, in the translation from jeans to skirt, the newly generated leg texture refers to the arm texture in the original image. Since the texture area obtained from the landmarks is small and irregular in shape, we match the closest one of the five pre-prepared regular skin textures, as the skin texture region $\mathbf{a}_s$ of the original image. The original skin region $\mathbf{a}_s$ is also resized to the same size as the output one $\mathbf{b}'_s$. The pixel-wise L2 loss is defined as follow:

$$\mathcal{L}_{skin} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{c=1}^{3}||p(\mathbf{a}_s)_{i,j}-p(\mathbf{b}'_s)_{i,j}||_2\circ w_{skin}, \tag{5}$$

where N represents the number of pixels within the generated skin region $\mathbf{b}'_s$ (e.g., the skin texture of legs), i and j indicate the coordination of each pixel. $p(\mathbf{a}_s)$ or $p(\mathbf{b}'_s)$ represents the value of channel c within the original skin region $\mathbf{a}_s$ or generated skin region $\mathbf{b}'_s$. w_{skin} represents the weight map of skin loss, in which the weights increase horizontally from the center of the human skeleton (i.e., the skeleton of legs or arms) to the edge. In this way, the texture of skin looks more stereoscopic in the generated image.

Finally, the total loss of DesignGAN becomes:

$$\begin{aligned}\mathcal{L}_{All} = \mathcal{L}_{LSGAN}+\lambda_{cyc}\mathcal{L}_{cyc}+\lambda_{idt}\mathcal{L}_{idt}+\\ \lambda_{style}\mathcal{L}_{style}+\lambda_{skin}\mathcal{L}_{skin},\end{aligned} \tag{6}$$

where $\lambda_{cyc}, \lambda_{idt}, \lambda_{style}, \lambda_{skin}$ indicate the weight for each term, which are set as (1,3,5,5) by default.

2.5. Customized Texture

In addition, we expand the previous task to customized texture input, which provides meaningful complement to the existing fashion translation field. Instead of the texture of original clothes, we constraint the similarity between the customized texture and the translated one.

More specifically, giving an original image x, the customized texture (e.g., a cobble texture) is pasted into the corresponding category region **a** (e.g., the region of jeans) guided by landmarks. Then, the feature extractor encodes the entire original image x and the jeans region **a** with the custom texture attached to it. The concated feature map is feed to the generator. After that, the translation of the clothes is also achieved by the domain loss, and the similarity of

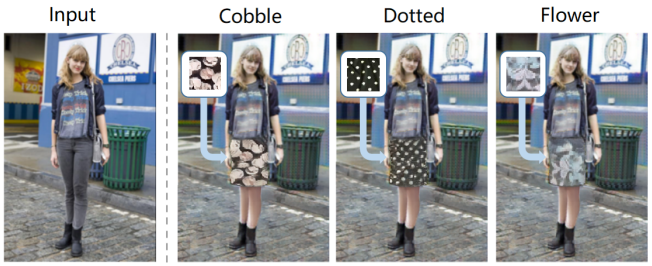

Fig. 4. The translation results with texture editing based on cobble, dotted, and flower inputs.

the texture is constrained by style loss and skin loss between corresponding regions. Finally, we can obtain an image y' with a custom input texture in the translation region $\mathbf{b}'$ (i.e., the region of skirt).

3. EXPERIMENT

3.1. Implementations Details

For different groups of translation, we sample two classes of fashion images from clothing co-parsing (CCP) [28] or multi-human parsing (MHP) [27]**dataset** which are also used by previous method. In our experiment, we evaluate our approach on four groups of translation, including jeans→skirt, skirt→jeans, shorts→pants, and pants→shorts. For each set of transitions, we selected 1500 images on average as training data. We set the classification score as our **evaluation metrics**, which is also used by previous works (e.g., InstaGAN [24]). More specifically, the classification score is defined as the ratio of images predicted as the target class by a pretrained ImageNet [35] classifier for each domain.

For the training setting of DesignGAN, we followed the InstaGAN [24] that resizing the input images to 300×200 (height×width) for CCP [28] dataset and 240×160 for MHP [27] dataset, respectively. We used Adam optimizer with the batchsize of 2. The training is completed with 8 GPUs in parallel. The initial learning rates of generator and discriminator are set as 0.0005 and 0.0002 for the first 100 epochs, and linearly decayed to zero for the next 100 epochs.

3.2. Cross-Category Translation Results

We compare our model with InstaGAN [24], and CycleGAN [9], as presented in Fig.3. CycleGAN method hardly changes the shape of the clothes and fails in all samples. InstaGAN method can generate reasonable shapes of the target categories. However, the translated clothes are easier to lose the details of texture. This issue is especially obvious in the case of skirt→jeans, where colored skirts are converted into black jeans. In addition, benefiting from our landmark attention mechanism, the generation effect around the key points (e.g., the waistline) of the clothes is better than the result of InstaGAN without any focus.

On the other hand, for the CycleGAN and InstaGAN, the resulting texture of legs or arms often looks unnatural, because they do not take into account the similarity constraints of the skin. Through the landmark branch, we can obtain the skin texture around the human arms or legs area, and utilize the skin loss as a constraint for the discriminator, to make the texture approximate the natural skin.

We quantitatively evaluate the performance of our approach, CycleGAN baseline, and InstaGAN. Table 1 shows the evaluation results for different translation categories. Our approach outperforms CycleGAN and InstaGAN in all experiments, which demonstrates better effectiveness of the proposed model.

Table 1. Evaluation results for different translations on CCP [28] dataset, including jeans→skirt, skirt→jeans, shorts→pants, and pants→shorts. We set the classification score as the **metrics**.

Category	skirt	jeans	pants	shorts
Real	0.888	0.946	0.984	0.720
CycleGAN	0.371	0.483	0.524	0.085
InstaGAN	0.600	0.540	0.768	0.232
DesignGAN	**0.653**	**0.587**	**0.821**	**0.313**

Table 2. Results for our ablation study on CCP [28] dataset.

Category	skirt	jeans	pants	shorts
No Attention	0.631	0.562	0.798	0.288
No Style Loss	0.627	0.554	0.781	0.261
No Skin Loss	0.642	0.569	0.813	0.308
Complete	**0.653**	**0.587**	**0.821**	**0.313**

3.3. Ablation Study

In this section, we perform an in-depth study of each component in our proposed model, as illustrated in Table 2. Our approach is composed of the CycleGAN architecture, the landmark attention branch, the style loss, and skin loss. We individually remove each component from the complete model. It can be found that the evaluation results drops obviously in every translation task when we remove each component. In particular, when the style loss is removed, the accuracy drops the most. In conclusion, the above ablation study demonstrates that each part plays a key role in the translation of the image, which further proves that the design of our model is reasonable and effective.

3.4. Texture Editing

Another contribution of our approach is the ability to edit the texture during the translation of the fashion category. Through the texture similarity constraint mechanism, we can not only constrain the similarity between the original clothes and the generated ones but also provide the generated clothes with specific textures like the flower, dotted, cobble, etc, as shown in Fig.4. The texture input for each corresponding result is shown in the upper-left box. Compared to the single translation results of the previous method, our method provides users with a richer choice.

Texture editing has a great application potential for fashion-related generation, which can incorporate any custom texture elements into the generated costume. It can even be an effective and convenient way to empower fashion designers.

4. CONCLUSION

Our work introduces a novel approach incorporating the attention guided by landmarks for image-to-image translation. The comparative evaluation demonstrates the effectiveness of our approach on the cross-category tasks. We solved the problem that losing the details of clothes texture after the translation by texture similarity constraint mechanism. Besides, we obtain the skin texture around the arms or legs area via landmarks and utilize the skin loss to make the texture approximate the natural skin. What's more, the predicted landmarks can guide the training process as attention, which can further enhance the discrimination ability. Also, the proposed DesignGAN can achieve texture editing based on a specific customized input, which can strengthen its applicability up to a higher level.

Acknowledgement. This work was supported by National Key R&D Program of China (No.2018YFB1404102), NSFC (No.61902347) and ZJNSF (No.LQ19F020002).

5. REFERENCES

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

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

[3] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros, "Generative visual manipulation on the natural image manifold," in *ECCV*. Springer, 2016.

[4] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy, "Be your own prada: Fashion synthesis with structural coherence," in *ICCV*, 2017.

[5] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo, "Semantic image synthesis via adversarial learning," in *ICCV*, 2017.

[6] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "Video-to-video synthesis," in *NIPS*, 2018.

[7] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh, "Recycle-gan: Unsupervised video retargeting," in *ECCV*, 2018.

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros, "Everybody dance now," *ICCV*, 2019.

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

[10] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, "Unsupervised image-to-image translation networks," in *NIPS*, 2017.

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018.

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

[13] Xiaodan Liang, Hao Zhang, and Eric P Xing, "Generative semantic manipulation with contrasting gan," *arXiv preprint arXiv:1708.00315*, 2017.

[14] Xiaodan Liang, Hao Zhang, Liang Lin, and Eric Xing, "Generative semantic manipulation with mask-contrasting gan," in *ECCV*, 2018.

[15] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.

[16] Adrian Bulat and Georgios Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *CVPR*, 2018.

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

[18] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *CVPR*, 2017.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *ICML*, 2018.

[20] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim, "Unsupervised attention-guided image-to-image translation," in *NIPS*, 2018.

[21] Xiaofeng Mao, Yuefeng Chen, Yuhong Li, Tao Xiong, Yuan He, and Hui Xue, "Bilinear representation for language-based image editing using conditional generative adversarial networks," in *ICASSP*. IEEE, 2019.

[22] Mehmet Günel, Erkut Erdem, and Aykut Erdem, "Language guided fashion image manipulation with feature-wise transformations," *arXiv preprint arXiv:1808.04000*, 2018.

[23] Patrick Esser, Ekaterina Sutter, and Björn Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018.

[24] Sangwoo Mo, Minsu Cho, and Jinwoo Shin, "InstaGAN: Instance-aware image-to-image translation," in *ICLR*, 2019.

[25] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool, "Pose guided person image generation," in *NIPS*, 2017.

[26] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018.

[27] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *MM*. ACM, 2018.

[28] Wei Yang, Ping Luo, and Liang Lin, "Clothing co-parsing by joint image segmentation and labeling," in *CVPR*, 2014.

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," *CVPR*, 2019.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014.

[31] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *CVPR*, 2019.

[32] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[33] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017.

[34] Yaniv Taigman, Adam Polyak, and Lior Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

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