
Fashion Editing with Multi-scale Attention Normalization

**Haoye Dong¹, Xiaodan Liang¹, Yixuan Zhang², Xujie Zhang¹,
Zhenyu Xie¹, Bowen Wu¹, Ziqi Zhang¹, Xiaohui Shen³, Jian Yin¹**

¹Sun Yat-sen University, ²Petuum Inc, ³ByteDance AI Lab

donghy7@mail2.sysu.edu.cn, xdliang328@gmail.com

Abstract

Interactive fashion image manipulation, which enables users to edit images with sketches and color strokes, is an interesting research problem with great application value. Existing works often treat it as a general inpainting task and do not fully leverage the semantic structural information in fashion images. Moreover, they directly utilize conventional convolution and normalization layers to restore the incomplete image, which tends to wash away the sketch and color information. In this paper, we propose a novel Fashion Editing Generative Adversarial Network (FE-GAN), which is capable of manipulating fashion images by free-form sketches and sparse color strokes. FE-GAN consists of two modules: 1) a free-form parsing network that learns to control the human parsing generation by manipulating sketch and color; 2) a parsing-aware inpainting network that renders detailed textures with semantic guidance from the human parsing map. A new attention normalization layer is further applied at multiple scales in the decoder of the inpainting network to enhance the quality of the synthesized image. Extensive experiments on high-resolution fashion image datasets demonstrate that the proposed method significantly outperforms the state-of-the-art methods on image manipulation.

1 Introduction

Fashion image manipulation aims to generate high-resolution realistic fashion images with user-provided sketches and color strokes. It has huge potential values in various applications. For example, a fashion designer can easily edit clothing designs with different styles; filmmakers can design characters by controlling the facial expression, hairstyle, and body shape of the actor or actress. In this paper, we propose FE-GAN, a fashion image manipulation network that enables flexible and efficient user interactions such as simple sketches and a few sparse color strokes. Some interactive manipulation results of FE-GAN are shown in Figure 1, which indicates that it can generate realistic images with convincing and desired details by controlling the sketch and color strokes.

In general, image manipulation has made great progress due to the significant improvement of neural network techniques [2, 6, 7, 14, 17, 21, 35]. However, previous methods often treat it as an end-to-end one-stage image completion problem without flexible user interactions [12, 16, 19, 20, 25, 32, 33]. Those methods usually do not explicitly estimate and then leverage the semantic structural information in the image. Furthermore, they excessively use the conventional convolutional layers and batch normalization, which significantly dissolve the sketch and color information from the input during propagation. As a result, the generated images usually contain unrealistic artifacts and undesired textures.

To address the above challenges, we propose a novel Fashion Editing Generative Adversarial Network (FE-GAN), which consists of a free-form parsing network and a parsing-aware inpainting network with multi-scale attention normalization layers. Different from the previous methods, we do not



Figure 1: Some interactive results of our FE-GAN. The input contains free-form mask, sketch, and sparse color strokes. Zoom in for details.

directly generate the complete image in one stage. Instead, we first generate a complete parsing map from incomplete inputs, and then render detailed textures on the layout induced from the generated parsing map. Specifically, in the training stage, given an incomplete parsing map obtained from the image, a sketch, sparse color strokes, a binary mask, and a noise sampled from the Gaussian distribution, the free-form parsing network learns to reconstruct a complete human parsing map guided by the sketch and color. A parsing-aware inpainting network then takes the generated parsing map, the incomplete image, and composed masks as the input of encoders, and synthesizes the final edited image. To better capture the sketch and color information, we design an attention normalization layer, which is able to learn an attention map to select more effective features conditioned on the sketch and color. The attention normalization layer is inserted at multiple scales in the decoder of the inpainting network. Moreover, we develop a foreground-based partial convolutional encoder for the inpainting network that is only conditioned on the valid pixels of the foreground, to enable more accurate and efficient feature encoding from the image.

We conduct experiments on our newly collected fashion dataset, named FashionE, and two challenging datasets: DeepFashion [36] and MPV [4]. The results demonstrate that incorporating the multi-scale attention normalization layers and the free-form parsing network can help our FE-GAN significantly outperforms the state-of-the-art methods on image manipulation, both qualitatively and quantitatively. The main contributions are summarized as follows: 1) We propose a free-form parsing network that enables users to control parsing generation flexibly by manipulating the sketch and color. 2) We develop a newly attention normalization for extracting features effectively based on a learned attention map. 3) We design a parsing-aware inpainting network with foreground-aware partial convolutional layers and multi-scale attention normalization layers, which can generate high-resolution realistic edited fashion images.

2 Related Work

Image Manipulation. Image manipulation with Generative Adversarial Networks (GANs) [6] is a popular topic in computer vision, which includes image translation, image completion, image editing, etc. Based on conditional GANs [18], Pix2Pix [11] is proposed for image-to-image translation. Targeting at synthesizing high-resolution photo-realistic image, Pix2PixHD [27] comes up with a novel framework with coarse-to-fine generators and multi-scale discriminators. [22, 33] design frameworks to restore low-resolution images with an original (square) mask, which generate some artifacts when facing the free-form mask and do not allow image editing. To make up for these deficiencies, Deepfillv2 [12] utilizes a user’s sketch as input and introduces a free-form mask to replace the original mask. On top of Deepfillv2, Xiong et al. [30] further investigate a foreground-aware image inpainting approach that disentangles structure inference and content completion explicitly. Faceshop [25] is a face editing system that takes sketch and color as input. However, the synthesized image would have blurry edges on the restored region, and it would obtain undesirable result if too much area erased. Recently, another face editing system SC-FEGAN [32] is proposed, which generates high-quality images when users provide the free-form as input. However, SC-FEGAN

is designed for face editing. In this paper, we propose a novel fashion editing system conditioned on the sketch and sparse color, utilizing feature involved in the parsing map, which is usually ignored by previous methods. Besides, we introduce a novel multi-scale attention normalization to extract more significant features conditioned on the sketch and color.

Normalization Layers. Normalization layers have become an indispensable component in modern deep neural networks. Batch Normalization (BN) used in Inception-v2 network [9], making the training of deep neural networks easier. Other popular normalization layers, including Instance Normalization (IN) [3], Layer Normalization (LN) [13], Weight Normalization (WN) [24], Group Normalization (GN) [34], are classified as unconditional normalization layers because no external data is utilized during normalization. In contrast to the above normalization techniques, conditional normalization layers require external data. Specifically, layer activations are first normalized to zero mean and unit deviation. Then a learned affine transformation is inferred from external data, which is utilized to modulate the activation to denormalized the normalized activations. The affine transformations are various among different tasks. For style transfer tasks [26, 31], affine parameters are spatially-invariant since they only control the global style of the output images. As for semantic image synthesis tasks, SPADE [23] applies a spatially-varying affine transformation to preserve the semantic information. In this paper, we propose a novel normalization technique named attention normalization. Instead of learning the affine transformation directly, attention normalization learns an attention map to extract significant information from the normalization activations. What’s more, compared to the SPADE ResBlk in SPADE [23], attention normalization has a more compact structure and occupies less computation resource.

3 Fashion Editing

We propose a novel method for editing fashion image, allowing users to edit images with a few sketches and sparse color strokes on an interested region. The overview of our FE-GAN is shown in Figure 2. The main components of our FE-GAN include a free-form parsing network and a parsing-aware inpainting network with the multi-scale attention normalization layers. We first discuss the free-form parsing network in Section 3.1. It can manipulate human parsing guided by free-form sketch and color, and is crucial to help the parsing-aware inpainting network produce convincing interactive results, which is described in Section 3.2. Then, in Section 3.3, we describe the attention normalization layers inserted at multiple scales in the inpainting decoder that can selectively extract effective features and enhance visual quality. Finally, in Section 3.4, we give a detailed description of the learning objective function used in our FE-GAN.

3.1 Free-form Parsing Network

Compared to directly restoring an incomplete image, predicting a parsing map from an incomplete parsing map is more feasible since there are fewer details in the parsing map. Meanwhile, the semantic information in the parsing map can be a guidance for rendering detail textures in each part of an image precisely. To this end, we propose a free-form parsing network to synthesize a complete parsing map when giving an incomplete parsing map and arbitrary sketch and color strokes.

The architecture of the free-form parsing network is illustrated in the upper left part of Figure 2. It is based on the encoder-decoder architecture like U-net [21]. The encoder receives five inputs: an incomplete parsing map, a binary sketch that describes the structure of the removed region, a noise sampled from the Gaussian distribution, sparse color strokes and a mask. More details about the input data will be discussed in Section 4.2. It is worth noting that given the same incomplete parsing map and various sketch and color strokes, the free-form parsing network can synthesize different parsing map, which indicates that our parsing generation model is controllable. It is significant for our fashion editing system since different parsing maps guide to render different contents in the edited image.

3.2 Parsing-aware Inpainting Network

The architecture of parsing-aware inpainting network is illustrated on the bottom of Figure 2. Inspired by [16], we introduce a partial convolution encoder to extract feature from the valid region in incomplete images. Our proposed partial convolution in partial convolution encoder is a bit different from the original version. Instead of using the mask directly, we utilize the composed mask to make

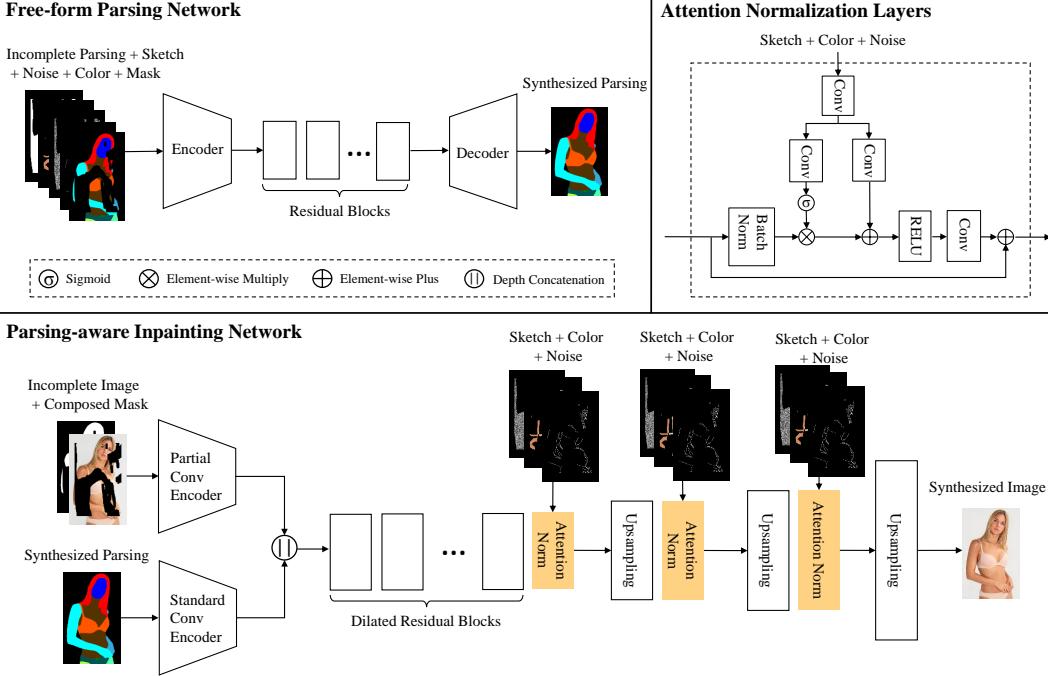


Figure 2: The overview of our FE-GAN. We first feed the incomplete human parsing, sketch, noise, color, and mask into free-form parsing network to obtain complete synthesized parsing. Then, incomplete image, composed mask, and synthesized parsing are fed into parsing-aware inpainting network for manipulating the image by using the sketch and color.

the network focus only on the foreground region. The composed mask can be expressed as:

$$M' = (\mathbf{1} - M) \odot M_f, \quad (1)$$

where M' , M and M_f are the composed mask, original mask and foreground mask respectively. \odot denotes element-wise multiply. Besides the partial convolution encoder, we introduce a standard convolution encoder to extract semantics feature from the synthesized parsing map. The human parsing map has semantics and location information that will guide the inpainting, since the content in a region with the same semantics should be similar. Given the semantic features, the network can render textures on the particular region more precisely. Two encoded feature maps are concatenated together in a channel-wise manner. Then the concatenated feature map undergoes several dilated residual blocks. During the upsampling process, well-designed multi-scale attention normalization layers are introduced to obtain attention maps, which are conditioned on sketch and color strokes. Unlike SC-FEGAN, the learned attention maps are helpful to select more effective feature in the forward activations. We explain the details in the next section.

3.3 Attention Normalization Layers

Attention Normalization Layers (ANLs) are similar to SPADE [23] to some extent and can be regarded as a variant of conditional normalization. However, instead of inferring an affine transformation from external data directly, ANLs learn an attention map which is used to extract the significant information in the earlier normalized activation. The upper right part of Figure 2 illustrates the design of ANLs. The details of ANLs are shown below.

Let x^i denotes the activations of the layer i in the deep neural network. Let N denotes the number of samples in one batch. Let C^i denotes the number of channels of x^i . Let H^i and W^i represent the height and width of activation map in layer i respectively. When the activations x^i passing through ANLs, they are first normalized in a channel-wise manner. Then the normalized activations are modulated by the learned attention map and bias. Finally, the modulated activations pass through a rectified linear unit (RELU) and a convolution layer and concatenate with the original normalized

activations. The activations value before the final concatenation at position ($n \in N, c \in C^i, h \in H^i, w \in W^i$) is signed as:

$$f(\alpha_{c,h,w}^i(\mathbf{d}) \frac{x_{n,c,h,w}^i - \mu_c^i}{\sigma_c^i} + \beta_{c,h,w}^i(\mathbf{d})), \quad (2)$$

where $f(x)$ denotes RELU and convolution operations, $x_{n,c,h,w}^i$ is the activation value at particular position before normalization, μ_c^i and σ_c^i are the mean and standard deviation of activation in channel c . As the same of BN [9], we formulate them as:

$$\mu_c^i = \frac{1}{NH^iW^i} \sum_{n,h,w} x_{n,c,h,w}^i \quad (3)$$

$$\sigma_c^i = \sqrt{\frac{1}{NH^iW^i} \sum_{n,h,w} (x_{n,c,h,w}^i)^2 - (\mu_c^i)^2} \quad (4)$$

The $\alpha_{c,h,w}^i(\mathbf{d})$ and $\beta_{c,h,w}^i(\mathbf{d})$ are learned attention map and bias for modulating the normalization layer, which are conditioned on the external data \mathbf{d} , namely, the sketch and color strokes and noise in this paper. Our implementations of $\alpha_{n,h,w}^i$ and $\beta_{n,h,w}^i$ are straightforward. The external data is first projected into an embedding space through a convolution layer. Then the bias is produced by another convolution layer, and the attention map is generated by a convolution layer and a sigmoid operation, which limits the range of feature map values between zero and one, and ensures the output to be an attention map. The effectiveness of ANLs is due to their inherent characteristics. Similar to SPADE [23], ANLs can avoid washing away semantic information in activations, since the attention map and bias are spatially-varying. Moreover, the multi-scale ANLs can not only adapt the various scales of activations during upsampling but also extract coarse-to-fine semantic information from external data, which guide the fashion editing more precisely.

3.4 Learning Objective Function

Due to the complex textures of the incomplete image and the variety of sketch and color strokes, the training of the free-form parsing network and parsing-aware inpainting network is a challenging task. To address these problems, we apply several losses to make the training easier and more stable in different aspects. Specifically, we apply adversarial loss \mathcal{L}_{adv} [6], perceptual loss $\mathcal{L}_{\text{perceptual}}$ [14], style loss $\mathcal{L}_{\text{style}}$ [14], parsing loss $\mathcal{L}_{\text{parsing}}$ [5], multi-scale feature loss $\mathcal{L}_{\text{feat}}$ [27], and total variation loss \mathcal{L}_{TV} [14] to regularize the training. We define a face TV loss to remove the artifacts of the face by using \mathcal{L}_{TV} on face region. We define a mask loss by using the L1 norm on the mask area, let I_{gen} be generated image, let I_{real} be ground truth, and let M be the mask, which is computed as:

$$\mathcal{L}_{\text{mask}} = \|I_{\text{gen}} \odot M - I_{\text{real}} \odot M\|_1, \quad (5)$$

we also define a foreground loss to enhance the foreground quality. Let $M_{\text{foreground}}$ be the mask of foreground part, then $\mathcal{L}_{\text{foreground}}$ can be formally computed as

$$\mathcal{L}_{\text{foreground}} = \|I_{\text{gen}} \odot M_{\text{foreground}} - I_{\text{real}} \odot M_{\text{foreground}}\|_1, \quad (6)$$

similar to $\mathcal{L}_{\text{foreground}}$, we formulate a face loss $\mathcal{L}_{\text{face}}$ to improve the quality of face region.

The overall objective function $\mathcal{L}_{\text{free-form-parser}}$ for free-form parsing network is formulated as:

$$\mathcal{L}_{\text{free-form-parser}} = \gamma_1 \mathcal{L}_{\text{parsing}} + \gamma_2 \mathcal{L}_{\text{feat}} + \gamma_3 \mathcal{L}_{\text{adv}}, \quad (7)$$

where hyper-parameters γ_1, γ_2 and γ_3 are weights of each loss.

The overall objective function $\mathcal{L}_{\text{inpainter}}$ for parsing-aware inpainting network written as:

$$\mathcal{L}_{\text{inpainter}} = \lambda_1 \mathcal{L}_{\text{mask}} + \lambda_2 \mathcal{L}_{\text{foreground}} + \lambda_3 \mathcal{L}_{\text{face}} + \lambda_4 \mathcal{L}_{\text{faceTV}} + \lambda_5 \mathcal{L}_{\text{perceptual}} + \lambda_6 \mathcal{L}_{\text{style}} + \lambda_7 \mathcal{L}_{\text{adv}}, \quad (8)$$

where hyper-parameters λ_i , ($i = 1, 2, 3, 4, 5, 6, 7$) are the weights of each loss.



Figure 3: Qualitative comparisons with Deepfill v1 [33], Partial Conv [16], and Edge-connect [19] on DeepFashion [36], MPV [4], and FashionE, respectively.

4 Experiments

4.1 Datasets and Metrics

We conduct our experiments on **DeepFashion** [36] from Fashion Image Synthesis track. It contains 38,237 images which are split into a train set and a test set, 29,958 and 8,279 images respectively. **MPV** [4] contains 35,687 images which are split into a train set and a test set, 29,469 and 6,218 samples. For better contributing to the fashion editing community, we collected a new fashion dataset, named **FashionE**. It contains 7,559 images with the size of 320×512 . In our experiment, we split it into a train set of 6,106 images and a test set of 1,453 images. The dataset will be released upon the publication of this work. The size of the image is 320×512 across all datasets.

We utilize the **Irregular Mask Dataset** provided by [16] in our experiments. The original dataset contains 55,116 masks for training and 24,866 masks for testing. We randomly select 12,000 images, splitting it into one train set of 9,600 masks and one test set of 2,400 masks. To mimic the free-form color stroke, we utilize one irregular mask dataset from [10] as **Irregular Strokes Dataset**. The mask region stands for stroke in our experiment. In our experiment, we split it into a train set of 50,000 masks and a test set of 10,000 masks. In our experiment, all the masks are resized to 320×512 .

Metrics. We evaluate our proposed method, as well as compared approaches on three metrics, PSNR (Peak Signal Noise Ratio), SSIM (Structural Similarity index) [28], and FID (Fréchet Inception Distance) [8]. We apply the Amazon Mechanical Turk (AMT) for evaluating the qualitative results.

4.2 Implementation Details

Training Procedure. The training procedure is two-stage. The first stage is to train free-form parsing network. We use $\gamma_1 = 10$, $\gamma_2 = 10$, $\gamma_3 = 1$ in the loss function. The second stage is to train parsing-aware inpainting network. We use $\lambda_1 = 5.0$, $\lambda_2 = 50$, $\lambda_3 = 1.0$, $\lambda_4 = 0.1$, $\lambda_5 = 0.05$, $\lambda_6 = 200$, $\lambda_7 = 0.001$ in the loss function. For both training stages, we use Adam [15] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and learning rate is 0.0002. The batch sizes of stage 1 is 20, and stage 2 is 8. In each training cycle, we train one step for the generator and one step for the discriminator. All the experiments are conducted on 4 Nvidia 1080 Ti GPUs.

Sketch & Color Domain. The way of extracting sketch and color domain from images is similar to SC-FEGAN. Instead of using HED [29], we generated sketches by Canny Edge Detector [1]. Relying on the result of human parsing, we use the median color of each segmented area to represent the color of that area. More details are presented in the supplementary material.

Discriminators. The discriminator, used in free-form parsing network, has a similar structure as the multi-scale discriminator in Pixel2PixelHD [27], which has two PatchGAN discriminators. The discriminator, used in parsing-aware inpainting network, has a similar structure as inpainting discriminator in Edge-connect [19], with five convolutions and spectral norm blocks.

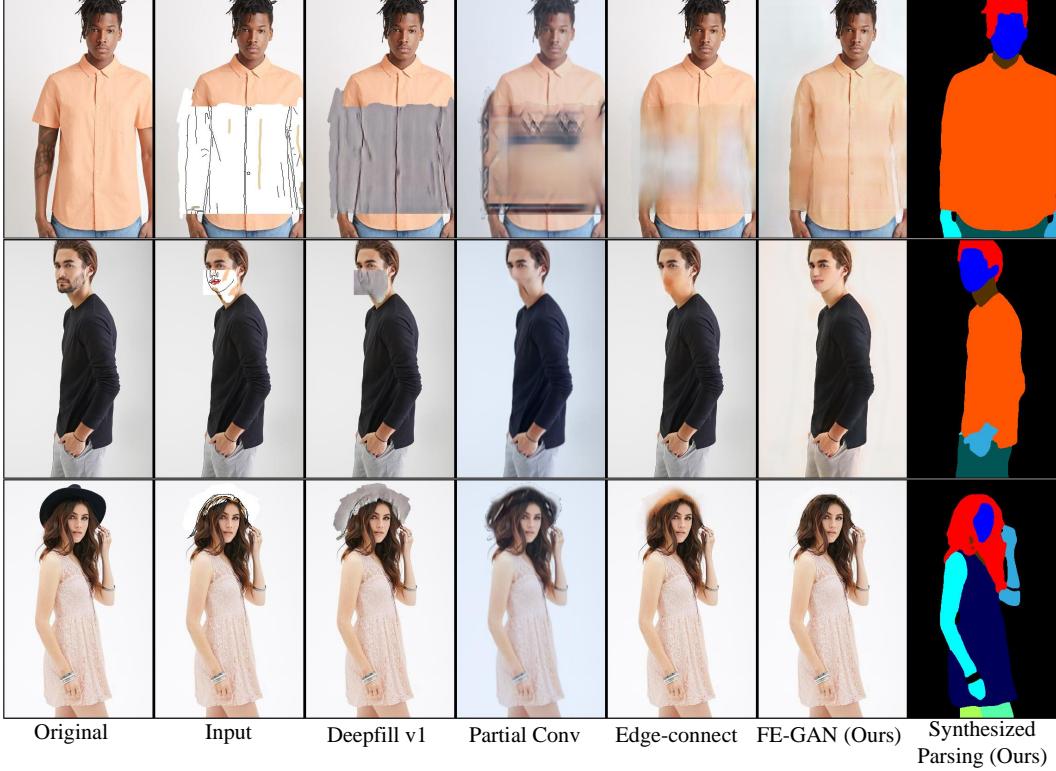


Figure 4: Some interactive comparisons with Deepfill v1 [33], Partial Conv [16], and Edge-connect [19] on DeepFashion [36], MPV [4], and FashionE, respectively.

Table 1: Quantitative comparisons on DeepFashion [36], MPV [4], and FashionE datasets.

Model	DeepFashion [36]			MPV [4]			FashionE		
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID
Deepfill v1 [33]	16.885	0.781	60.994	18.450	0.808	58.742	19.170	0.814	56.738
Partial Conv [16]	19.103	0.827	17.728	20.408	0.850	22.751	20.635	0.848	20.148
Edge-connect [19]	26.236	0.901	12.633	27.557	0.924	7.888	29.154	0.926	5.182
FE-GAN (Ours)	29.552	0.928	3.700	30.602	0.944	3.796	30.974	0.938	3.246

Compared Approaches. To make a comprehensive evaluation of our proposed method, we conduct three comparison experiments based on the recent state of the art approaches at image inpainting [16, 19, 33]. It comprises of an edge generator and an image completion module. The re-implementations followed the source codes provided by authors. To make a fair comparison, all inputs consist of incomplete images, masks, sketch, color domain, and noise across all comparison experiments.

4.3 Quantitative Results

PSNR computes the peak signal-to-noise ratio between images. SSIM measures the similarity between two images. Higher value of PSNR and SSIM mean better results. FID is tended to replace Inception Score as one of the most significant metrics measuring the quality of generated images. It computes the Fréchet distance between two multivariate Gaussians, the smaller the better. As mentioned in [28], there is no good numerical metric in image inpainting. Furthermore, our focus is even beyond the regular inpainting. We can observe from Table 1, our FE-GAN achieves the best PSNR, SSIM, and FID scores and outperforms all other methods among three datasets.

4.4 Qualitative Results

Beyond numerical evaluation, we present visual comparisons for image completion task among three datasets and four methods, shown in Figure 3. Three rows, from top to bottom, are results from

DeepFashion, MPV, and FashionE. The interactive results for those methods are shown in Figure 4. The last column of the Figure 4, are the results of the free-form parsing network. We can observe that the free-form parsing network can obtain promising parsing results by manipulating the sketch and color. Thanks to the multi-scale attention normalization layers and the synthesized parsing result from the free-form parsing network, our FE-GAN outperforms all other baselines on visual comparisons.

4.5 Human Evaluation

To further demonstrate the robustness of our proposed FE-GAN, we conduct the human evaluation deployed on the Amazon Mechanical Turk platform on the DeepFashion [36], MPV [4], and FashionE. In each test, we provide two images, one from compared methods, the other from our proposed method. Workers are asked to choose the more realistic image out of two. During the evaluation, K images from each dataset are chosen, and n workers will only evaluate these K images. In our case, $K = 100$ and $n = 10$. We can observe from Table 2, our proposed method has a superb performance over the other baselines. This confirms the effectiveness of our FE-GAN comprised of a free-form parsing network and a parsing-aware network, which generates more realistic fashion images.

Table 2: Human evaluation results of pairwise comparison with other methods.

Comparison Method Pair	DeepFashion [36]	MPV [4]	FashionE
Ours vs Deepfill v1 [33]	0.849 vs 0.151	0.845 vs 0.155	0.857 vs 0.143
Ours vs Partial Conv [16]	0.917 vs 0.083	0.864 vs 0.136	0.799 vs 0.201
Ours vs Edge-connect [19]	0.790 vs 0.210	0.691 vs 0.309	0.656 vs 0.344

5 Ablation Study

To evaluate the impact of the proposed component of our FE-GAN, we conduct an ablation study on FashionE with using the model of 20 epochs. As shown in Table 3 and Figure 5, we report the results of the different versions of our FE-GAN. We first compare the results using attention normalization to the results without using it. We can learn that incorporating the attention normalization layers into the decoder of the inpainting module significantly improves the performance of image completion. We then verify the effectiveness of the proposed free-from parsing network. From Table 3 and Figure 5, we observe that the performance drops dramatically without using parsing, which can depict the human layouts for guiding image manipulation with higher-level structure constraints. The results report that the main improved performance achieved by the attention normalization and human parsing. We also explore the impact of our designed objective function that each of the losses can substantially improve the results.

Method	PSRN	SSIM	FID
Full	30.035	0.932	4.092
w/o attention norm	29.185	0.920	5.191
w/o parsing	29.109	0.923	5.355
w/o $\mathcal{L}_{\text{mask}}$	28.813	0.921	4.773
w/o $\mathcal{L}_{\text{foreground}}$	29.848	0.927	5.030

Table 3: Ablation studies on FashionE.



Figure 5: Ablation studies on FashionE. (a1)(b1): Ours(Full); (a2): w/o attention norm; (b2): w/o parsing.

6 Conclusions

We propose a novel Fashion Editing Generative Adversarial Network (FE-GAN), which enables users to manipulate the fashion image with an arbitrary sketch and a few sparse color strokes. The FE-GAN incorporates a free-form parsing network to predict the complete human parsing map to guide fashion image manipulation. Moreover, we develop a foreground-based partial convolutional encoder and design an attention normalization layer which used in the multiple scales layers of the decoder for the inpainting network. The experiments on fashion datasets demonstrate that our FE-GAN outperforms the state-of-the-art methods and achieves high-quality performance with convincing details.

References

- [1] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8:679–698, 1986.
- [2] 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.
- [3] Victor Lempitsky Dmitry Ulyanov, Andrea Vedaldi. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [4] Haoye Dong, Xiaodan Liang, Bochao Wang, Hanjiang Lai, Jia Zhu, and Jian Yin. Towards multi-pose guided virtual try-on network. *arXiv preprint arXiv:1902.11026*, 2019.
- [5] Ke Gong, Xiaodan Liang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, pages 6757–6765, 2017.
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.
- [7] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. *arXiv preprint arXiv:1711.08447*, 2017.
- [8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.
- [9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.
- [10] Kazizat T. Iskakov. Semi-parametric image inpainting. *arXiv preprint arXiv:1807.02855*, 2018.
- [11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [12] Jimei Yang-Xiaohui Shen Xin Lu Thomas S. Huang Jiahui Yu, Zhe Lin. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.
- [13] Geoffrey E. Hinton Jimmy Lei Ba, Jamie Ryan Kiros. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [16] Guilin Liu, Fitzsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018.
- [17] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017.
- [18] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [19] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mohammad Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.
- [20] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.
- [22] Hiroshi Ishikawa Satoshi Iizuka, Edgar Simo-Serra. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*,36(4):107, 2017.
- [23] Ting-Chun Wang Jun-Yan Zhu Taesung Park, Ming-Yu Liu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.
- [24] Diederik P. Kingma Tim Salimans. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NIPS*, 2016.
- [25] Attila Szabó-Siavash Arjomand Bigdeli-Paolo Favaro Matthias Zwicker Tiziano Portenier, Qiyang Hu. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*,37(4):99, 2018.
- [26] Manjunath Kudlur Vincent Dumoulin, Jonathon Shlens. A learned representation for artistic style. In *ICLR*, 2016.
- [27] 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. *arXiv preprint arXiv:1711.11585*, 2017.
- [28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.
- [29] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 5505–5514, 2015.
- [30] Wei Xiong, Jiahui Yu, Zhe L. Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. *arXiv preprint arXiv:1901.05945*, 2019.
- [31] Serge Belongie Xun Huang. Arbitrary style transfer in realtime with adaptive instance normalization. In *ICCV*, 2017.
- [32] Jongyoul Park Youngjoo Jo. Sc-fegan: Face editing generative adversarial network with user’s sketch and color. *arXiv preprint arXiv:1902.06838*, 2019.
- [33] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018.
- [34] Kaiming He Yuxin Wu. Group normalization. In *ECCV*, 2018.
- [35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.
- [36] Shi Qiu-Xiaogang Wang Ziwei Liu, Ping Luo and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

Appendix

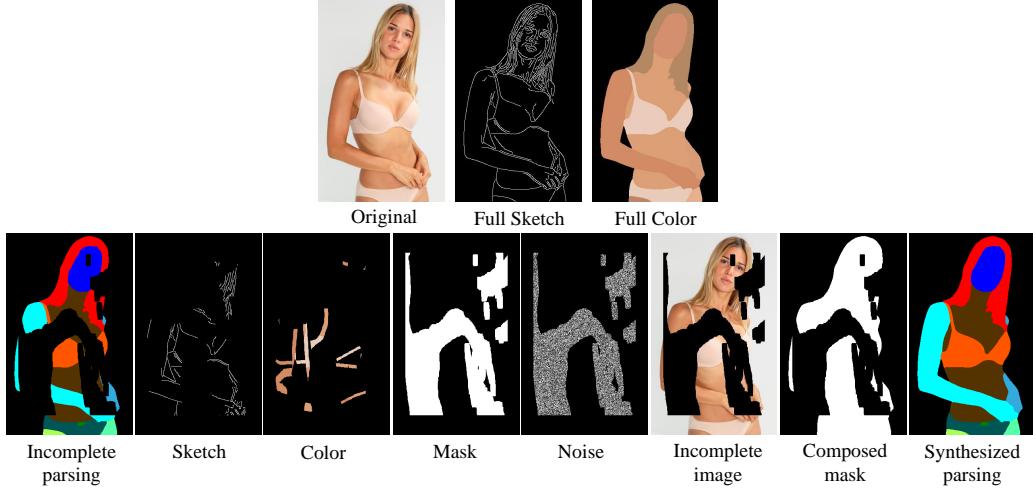


Figure 6: Example of model inputs shown in the second row. The inputs of the free-form parsing network consist of incomplete parsing, sketch, color, mask, and noise; the inputs of parsing-aware inpainting network contain incomplete image, composed mask and synthesized parsing. The inputs of attention normalization layers are a sketch, color, and noise. We first generate the sketches by using Canny [1] shown in the second column of the first row. Then, we use a human parser [5] to extract the median color of each part of the person, shown in the last column of the first row.



Figure 7: Some interactive comparisons of Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN (Ours). The results of our FE-GAN are shown in the last column. Zoom in for details.



Figure 8: Some interactive results of our FE-GAN, shown in the third column. The input contains free-form mask, sketch, and sparse color strokes. The results of our free-form parsing network shown in the last column. Zoom in for details.



Figure 9: Some interactive results of our FE-GAN, shown in the third column. The input contains free-form mask, sketch, and sparse color strokes. The results of our free-form parsing network shown in the last column. Zoom in for details.



Figure 10: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on FashionE.

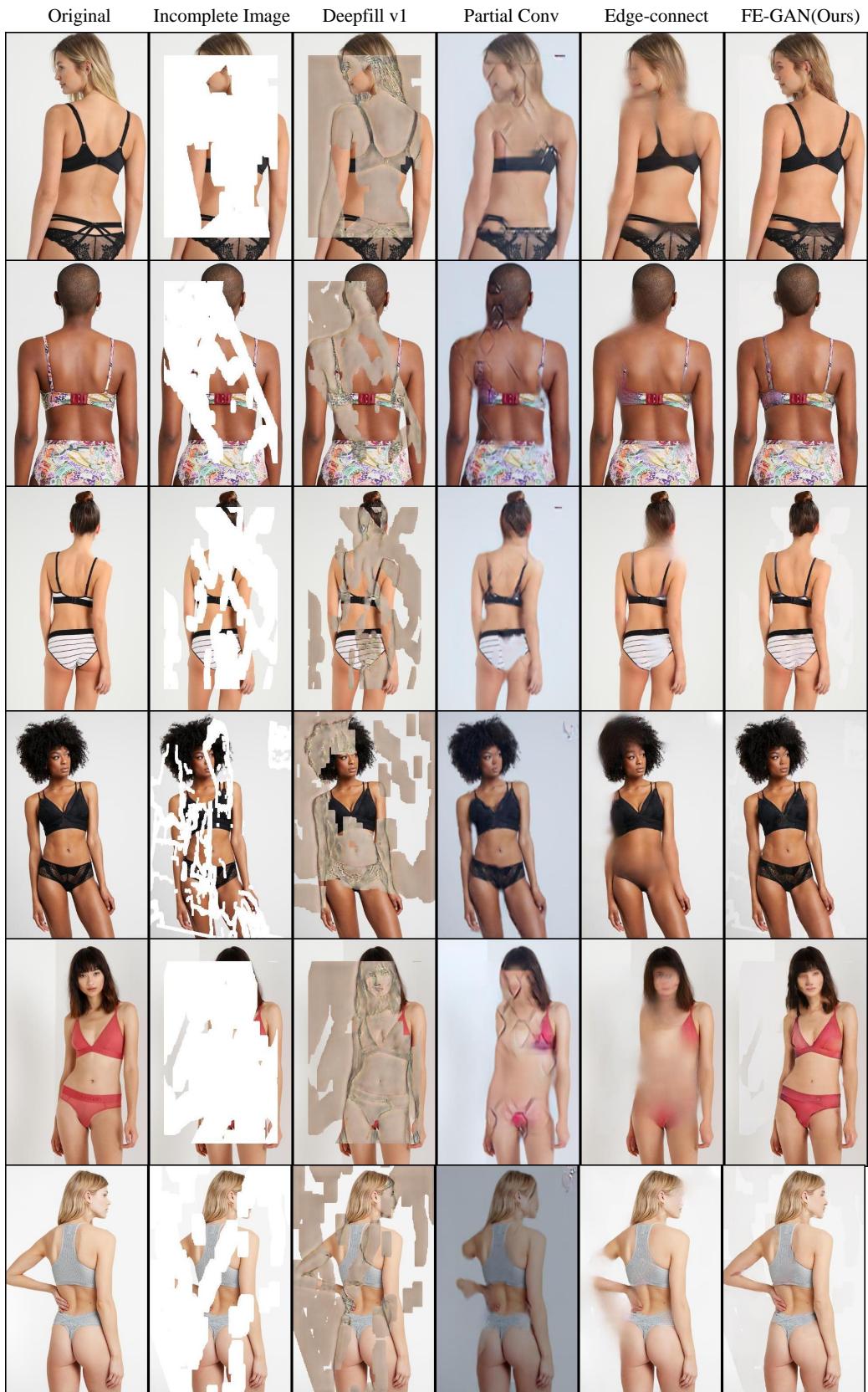


Figure 11: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on FashionE.



Figure 12: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on FashionE.

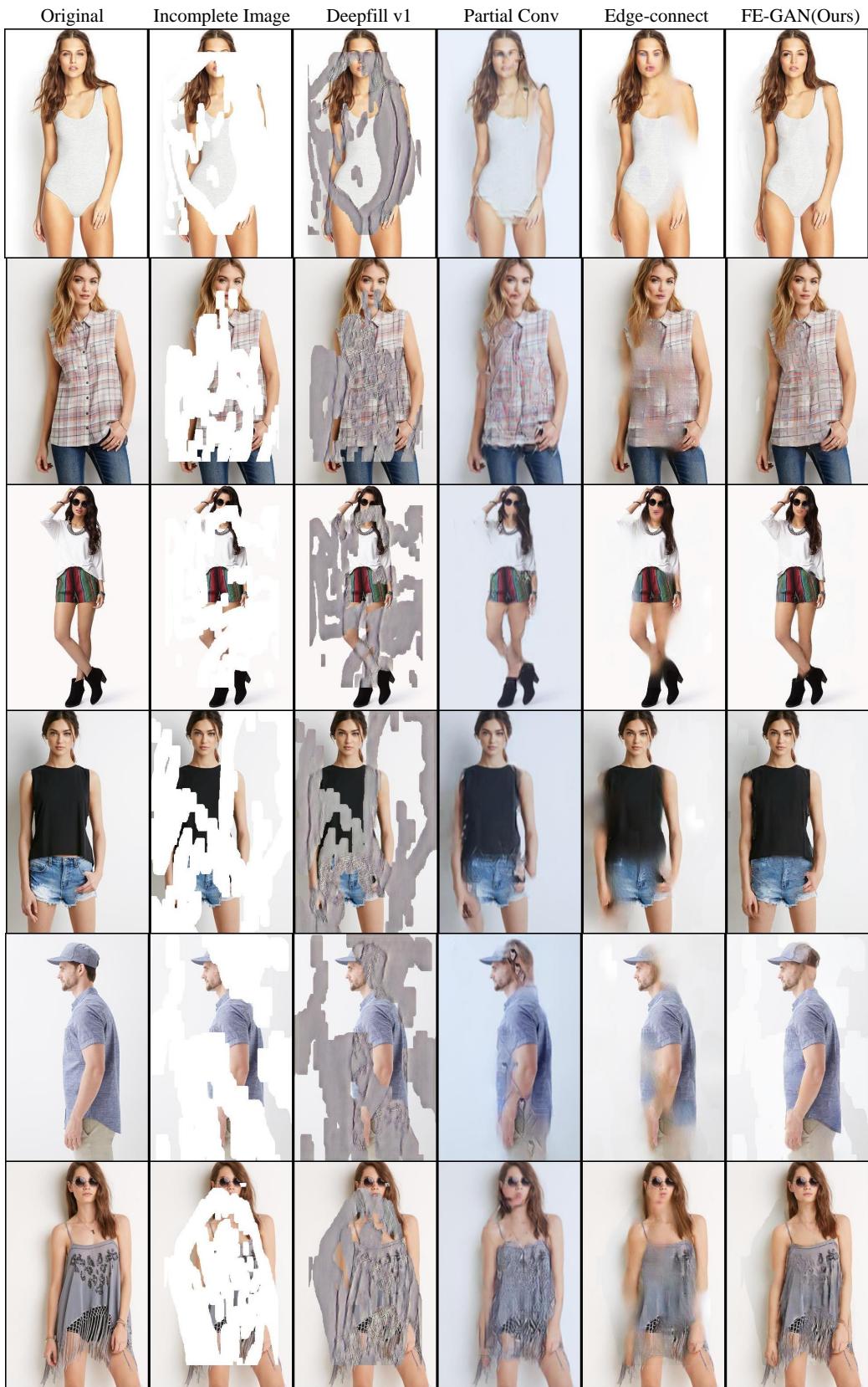


Figure 13: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on DeepFashion [36]

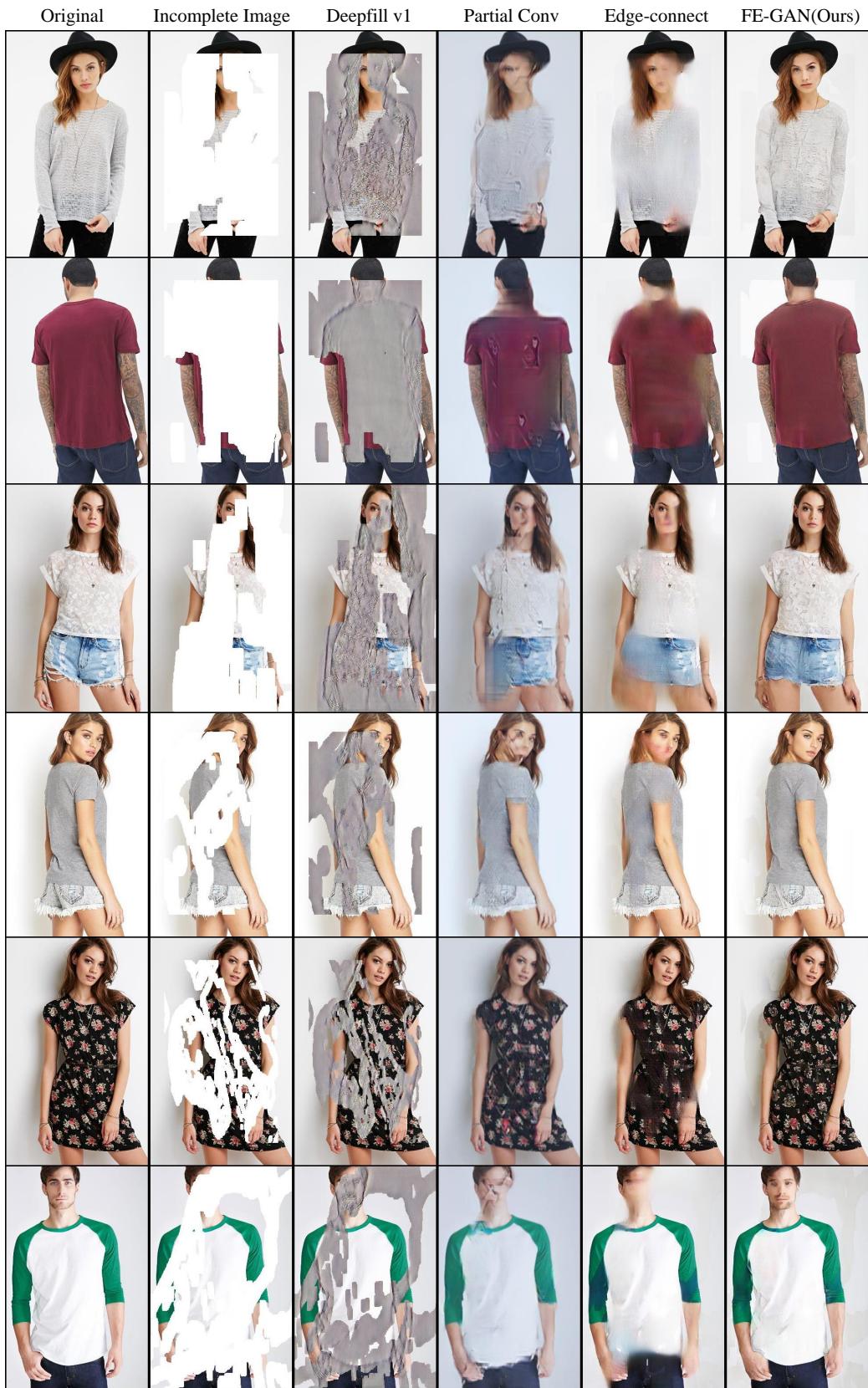


Figure 14: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on DeepFashion [36]

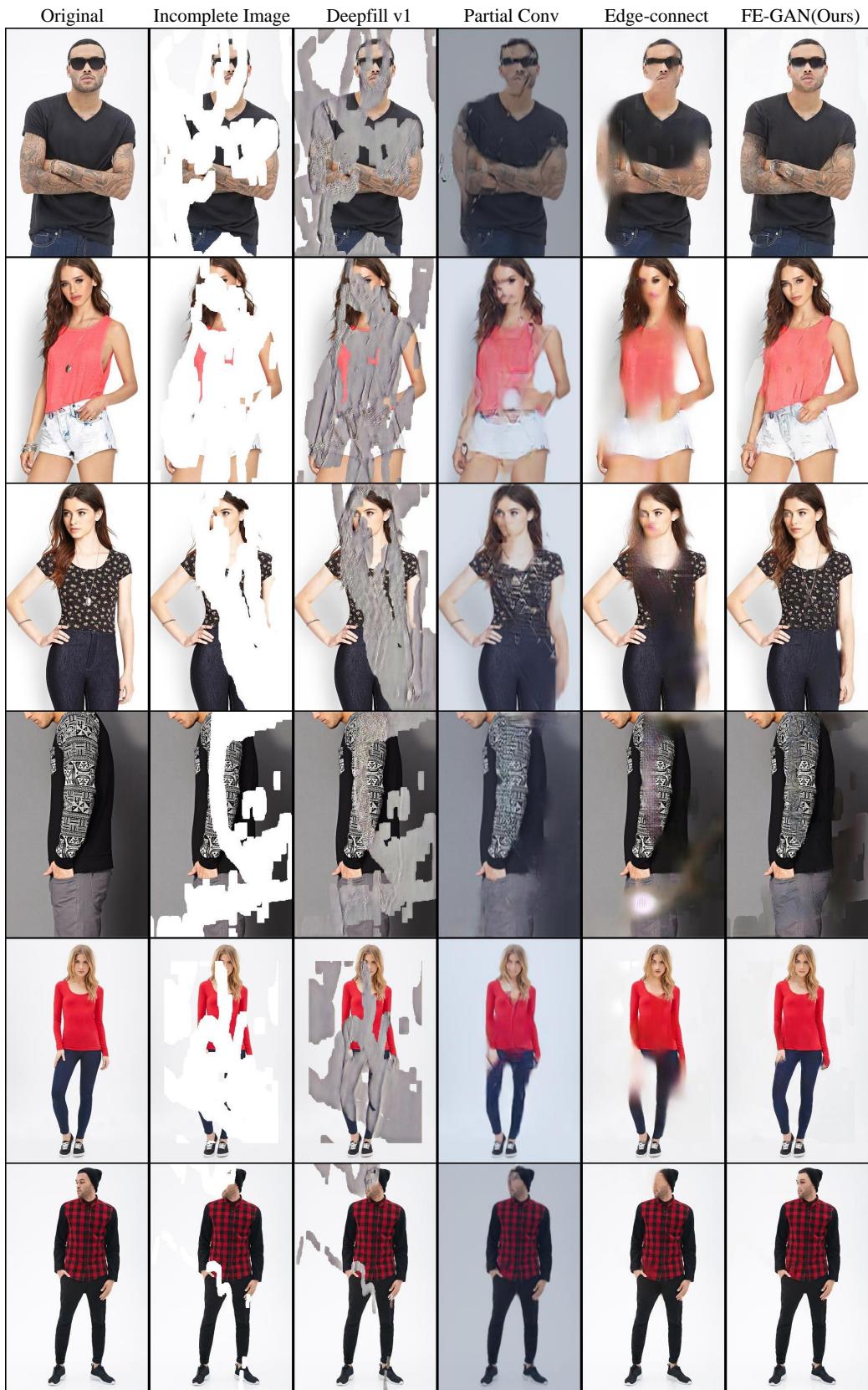


Figure 15: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on DeepFashion [36]

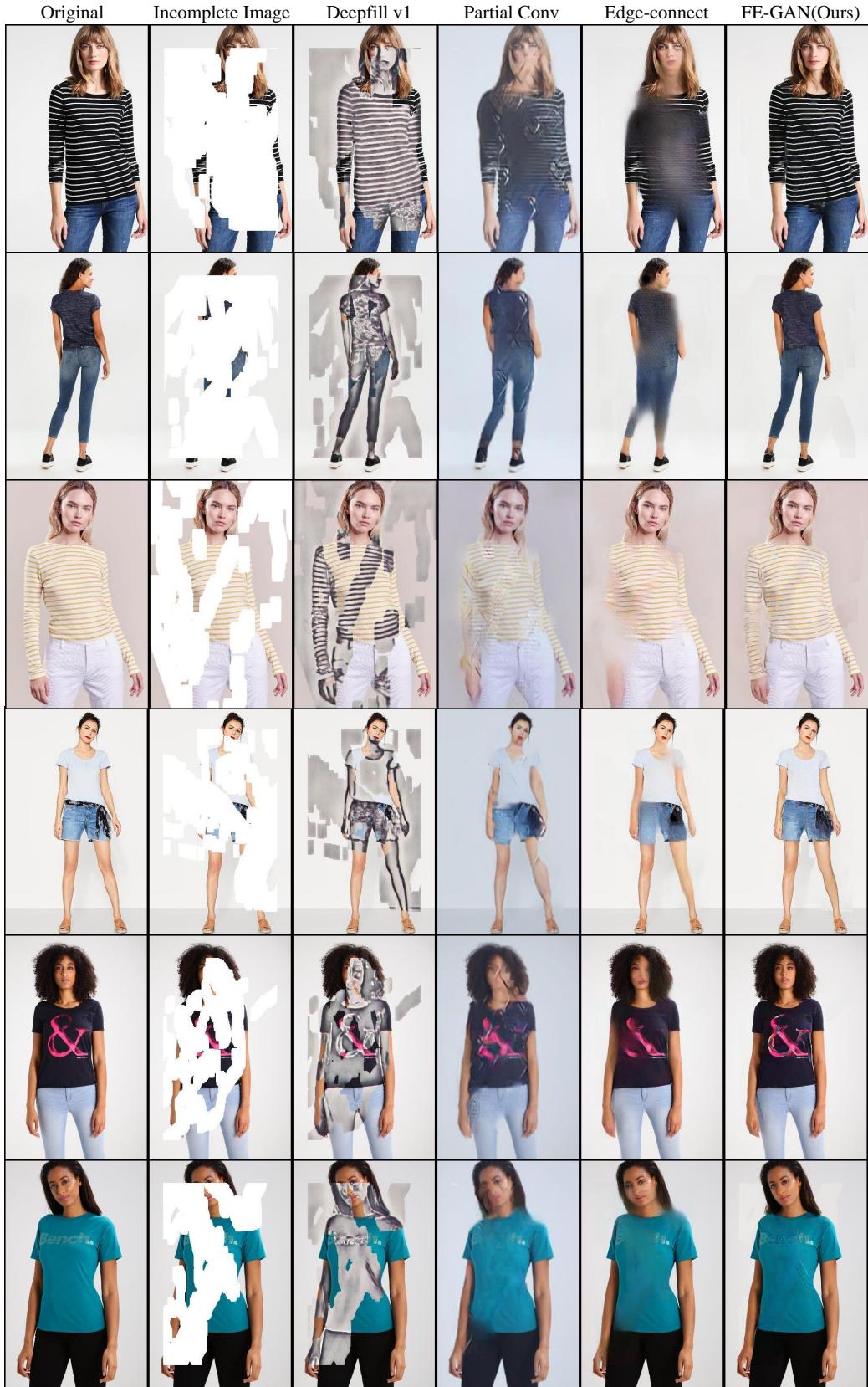


Figure 16: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on MPV [4].



Figure 17: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on MPV [4].



Figure 18: Qualitative comparisons between Deepfill v1 [33], Partial Conv [16], Edge-connect [19], and FE-GAN on MPV [4].