Prior Based Human Completion

Zibo Zhao[1] Wen Liu[1] Yanyu Xu[2] Xianing Chen[1] Weixin Luo[1] Lei Jin[1]
Bohui Zhu[4] Tong Liu[5] Binqiang Zhao[4] Shenghua Gao[†1,3]

[1]ShanghaiTech University
[2]Institute of High Performance Computing, A*STAR
[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[4]Alibaba Group [5] Taobao

{zhaozb,liuwen,chenxn1,luowx,jinlei,gaoshh}@shanghaitech.edu.cn yingmu@taobao.com
xu_yanyu@ihpc.a-star.edu.sg {bhiui.zbh,binqiang.zhao}@alibaba-inc.com

Abstract

We study a very challenging task, human image completion, which tries to recover the human body part with a reasonable human shape from the corrupted region. Since each human body part is unique, it is infeasible to restore the missing part by borrowing textures from other visible regions. Thus, we propose two types of learned priors to compensate for the damaged region. One is a structure prior, it uses a human parsing map to represent the human body structure. The other is a structure-texture correlation prior. It learns a structure and a texture memory bank, which encodes the common body structures and texture patterns, respectively. With the aid of these memory banks, the model could utilize the visible pattern to query and fetch a similar structure and texture pattern to introduce additional reasonable structures and textures for the corrupted region. Besides, since multiple potential human shapes are underlying the corrupted region, we propose multi-scale structure discriminators to further restore a plausible topological structure. Experiments on various large-scale benchmarks demonstrate the effectiveness of our proposed method.

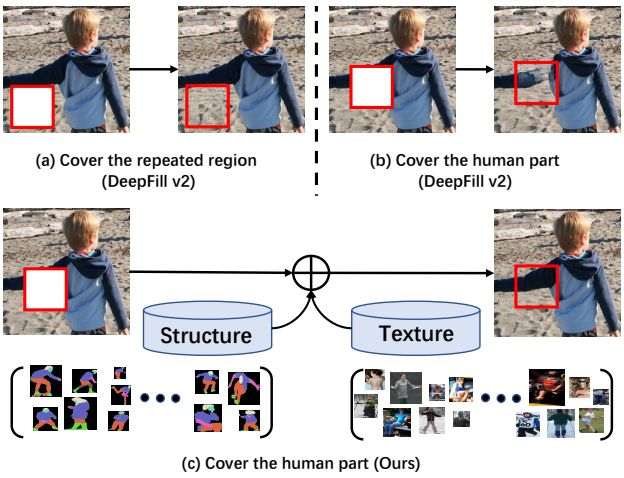

Figure 1. Illustration of the human body completion. (a) Existing image inpainting methods [48] usually works well to recover these repeated regions, (b) while they might fail to recover the unique human body region. (c) We mainly focus on how to recover the corrupted unique human body in the image via borrowing structure and texture information of other images from the learned memory. With the help of additional prior information, our proposed method could generate a more realistic and plausible image.

1. Introduction

Human image completion aims to repair human body parts in the corrupted human image. It has various potential applications, including the restoration of old human photographs and films, human image re-composition, and fashion clothing re-editing.

By far, most image inpainting works [27, 45, 33, 44] make the best use of the repeated textures in visible regions, such as the attention mechanism [47], to restore a photo-realistic image from the damaged part. Since there are both human body and background in corrupted regions, and they usually have the different style of textures, it is ambiguous to place the repeated textures into the right position. Thus, following works [11, 54, 10, 32] take an additional human mask as guidance to ease the texture ambiguities between foreground and background. However, since each human body part is unique (only one head), it is impracticable to borrow the repeated textures from other parts. Besides, a human body is non-rigid with diverse poses, and thereby multiple potential human shapes exist for the damaged re-

† Corresponding author

gion. This ambiguity could further make the inpaintor produce a terrible result with unreasonable human body structure. As shown in Figure 1 (a), the existing method usually works well to repair the corrupted area from the visible repeated texture, such as the sand. While when dealing with the unique human body part, such as the arm in Figure 1 (b), it produces a result with an implausible human body structure.

Fortunately, the human body is highly structured. We could make the full use of the structure information of a human body to guide the network to generate decent results with reasonable human shapes. There are several types of human structures, including a binary foreground/background mask [3, 20], 2D pose [6, 34], human parsing [38, 39], densepose [1, 25], and 3D body mesh [2, 13, 21]. From left to right, they have a more powerful capability in representing the structure in detail but with more difficulties for their corresponding structure estimator to infer an accuracy structure from a corrupted image. We use the human parsing as a human structure prior, considering the trade-off between the representative ability and the accuracy of its estimator in a corrupted image [30].

By leveraging the structure prior, we first learn a structure and a texture memory bank, encoding the common body structure and texture patterns, respectively. To learn more representative features for both the structure and texture, the memory prior encoding network takes the concatenation of structure and texture as the input. Then, we propose to estimate the human parsing map of the corrupted image and recover a reasonable complete human parsing map with the help of the structure memory bank. To further eliminate the shape ambiguities underlying in corrupted regions and produce an image with a more plausible topological structure, we also use multi-scale structure discriminators to regularize the connections at different scales of topological levels to be reasonable, such as the global level (the whole body), the middle level (upper body), and the local level (arm-with-torso). In this way, we could recover a semantic plausibly structure map.

Conditioned on the recovered complete structure map, we propose a texture completion module to restore the human body textures in the corrupted regions. Since the texture of each human body part is unique, it is infeasible for the incomplete area to borrow the repeated textures from the visible regions directly or with the advanced attentional mechanism [47]. We further propose a structure-texture correlation prior to introduce additional plausible textures for the corrupt region. Specifically, the texture completion model could retrieve a reasonable texture pattern from the texture memory bank as a compensation for the incomplete region corresponding to its complete structure. Also, to preserve the identity of the visible part, we use skip-connections between encoding features and memory features. As shown in Figure 1 (c), our method could recover the human body part from the corrupted regions with a plausible human body structure.

We summarize our contributions as follows: (i) Different from previous image inpainting works, we focus on recovering the unique human body part from the incomplete region with a plausible human body structure; (ii) To further produce an image with a reasonable topological structure in the human body, we propose to use multi-scale structure discriminators to regularize the connection among all the topological levels; (iii) We propose a structure and texture memory bank to introduce more additional priors as compensation for the corrupted region. Extensive experiments demonstrate the effectiveness of our proposed method.

2. Related Work

2.1. Image Inpainting

Benefiting from the growth of deep generative models [9, 51], learning-based image inpainting methods [7, 29, 22, 50, 31, 53, 19, 46, 52] achieved impressive performance by learning high-level information from large scale datasets. Following the assumption that missing pixels could be found from visible parts in images, early methods [27, 45, 12, 40, 35], train the network to map corrupted images to complete images and force the model to reconstruct complete images. [12] introduces global and local discriminators. The former critics the texture of the whole image, and the latter penalizes the patch of generated contents. However, it still produces results with artifacts because the model can not capture the long-term correlation.

To better utilize the information of visible parts and capture the long-term correlation between uncorrupted and corrupted regions, [47] devises contextual attention, which computes the similarity score between the visible part and generated part. Then it always borrows the most relevant patch to fill up the content. Befitting from the attention module, the model could capture the relation between feasible regions and generated regions. [16] uses an iterative method to utilize visible parts in the image to a greatest extent. However, because they process valid pixels and invalid pixels in the same way, the results are still generated with artifacts. [17, 48] solve the problem by designing new convolution operators. Based on the contextual attention module, [48] further proposes using gated convolution to learn the validate information automatically. The model produces smooth results by using the correlation between each pixel and reduce the influence from invalid pixels. However, results are always structurally unreasonable because it ignores the high-level structure.

Except for attention-based methods, [24, 43, 15] propose to use edge maps as the condition for networks to guide the model extract high-level features. These methods produce

results with reasonable structure on easy scenes but fail in the complex scenes because the environmental lighting and texture on objects would make the edge-map ambiguous. Thus, the edge map may mislead the model to generate wired results on the wild data.

2.2. Human Parsing

Previous works [42, 8] address the human semantic segmentation task as a multi-task problem, and they perform pose-estimation and human semantic segmentation together. Later, [30] utilizes edge maps of images as conditions to facilitate the prediction. Their method achieves high performance. Recent methods [38, 39] propose first to predict each part of the human and then merge them up in a bottom-up strategy.

2.3. Fashion Editing

Fashion editing is driving more attention in the community and industry due to its great commercial value. [37] proposes to evaluate the compatibility between different cloth with traditional computer vision methods. Later, [54, 41, 4, 23] utilize human parsing maps as conditions to change the cloth on human bodies in single person images. Unlike the above methods, [11, 5] propose solving the fashion editing task with inpainting methods. Since the style of clothes might obey a specific distribution, [11] uses a network to learn a fashion distribution to predict the human parsing maps and images. However, they can only predict one cloth of the whole human body, and the mask is fixed. [5] utilizes an inpainting network to recover the human parsing and then produce images. Although this method can handle more kinds of masks, the model needs some guidance in the input to facilitate the result and may fail in practical applications.

3. Methodology

Our proposed method consists of three stages, including prior encoding, segmentation completion, and texture completion, as illustrated in Figure 2.

3.1. Problem Formulation

Given a corrupted single person image, human completion aims to fill up its missing regions. We denote the input corrupted image $I^c \in \mathbb{R}^{3\times H\times W}$ as $I_{gt} \odot (1-M)$, where $I_{gt} \in \mathbb{R}^{3\times H\times W}$ denotes the ground-truth image and $M \in \mathbb{R}^{1\times H\times W}$ denotes the images mask (1 for the lost pixels). The predicted image is denoted with $I_{pred} \in \mathbb{R}^{3\times H\times W}$. The existing image completion methods usually train a network G to recover the missing content, as formulated:

$$I_{pred} = G(I^c). \tag{1}$$

However, corrupted human parts in images are always unique, leading to failure cases of the existing methods via borrowing information within the image. Thus, we introduce structure and texture correlation prior to our designed memory bank module to encode additional information from other images. We also propose utilizing the structure prior by leveraging the human segmentation map as input to the network to encode the formal human structure better. To this end, human body completion can formulate as:

$$I_{pred} = G(I^c, S, E), \tag{2}$$

where E represents the learned memory bank module, and S represents the estimated corrupted human semantic map.

3.2. Prior Encoding

We design a **memory bank module** to encode the common information of human bodies in both structure and texture. Besides, we also leverage the multi-scale architecture to capture local and global information on different scales for further improvements. Specifically, each memory bank is a 64×512 dictionary that stores the learned latent vectors. Each latent vector corresponds to the stored feature. This stage illustrates in Figure 2 (a).

We maintain two types of memory bank: a structure memory bank E^s and a texture memory bank E^t, which design to encode and store the human structure and texture information in the whole training dataset, respectively. Each type of memory bank has bi-level memories; the low-level memory bank stores the local detail of images, and the high-level memory bank stores the global information of images.

The Structure Memory Bank. We use an auto-encoder embedded with bi-level memories to extract the common information in numerous images. Using concatenation of the complete single person image I_{gt} and the complete human segmentation map S_{gt} as input, the auto-encoder is constrained to reconstruct the complete human segmentation map S_{pred}, as shown in Figure 2 (a). The encoder extracts the low-level feature map $f^s_{low} \in \mathcal{R}^{C\times 64\times 64}$ and high-level feature map $f^s_{high} \in \mathcal{R}^{C\times 32\times 32}$. We first use high-level features to query in the high-level memory bank E^s_{high}. Specifically, for each feature $f^s_j \in f^s_{high}$, we use an L-2 distance to measure the similarity between f^s_j and e_i, and search its most similar latent vector $e_i \in E^s_{high}$, which is formulated as:

$$f^s_j = e_i, i = \arg\min_k \|f^s_j - e_k\|^2_2, \tag{3}$$

where $k \in \{0, 1, ..., 512\}$. We gain the new high-level feature map $\hat{f}^s_{high}$ with retrieved features in E^s_{high}. The decoder up-samples f^s_{high} into scale 64×64. Similarly, we get the new low-level feature map $\hat{f}^s_{low}$ from the low-level memory bank E^s_{low}, and the decoder further reconstructs the human segmentation map S_{pred}. To encode the common structure information of other images, the auto-encoder is

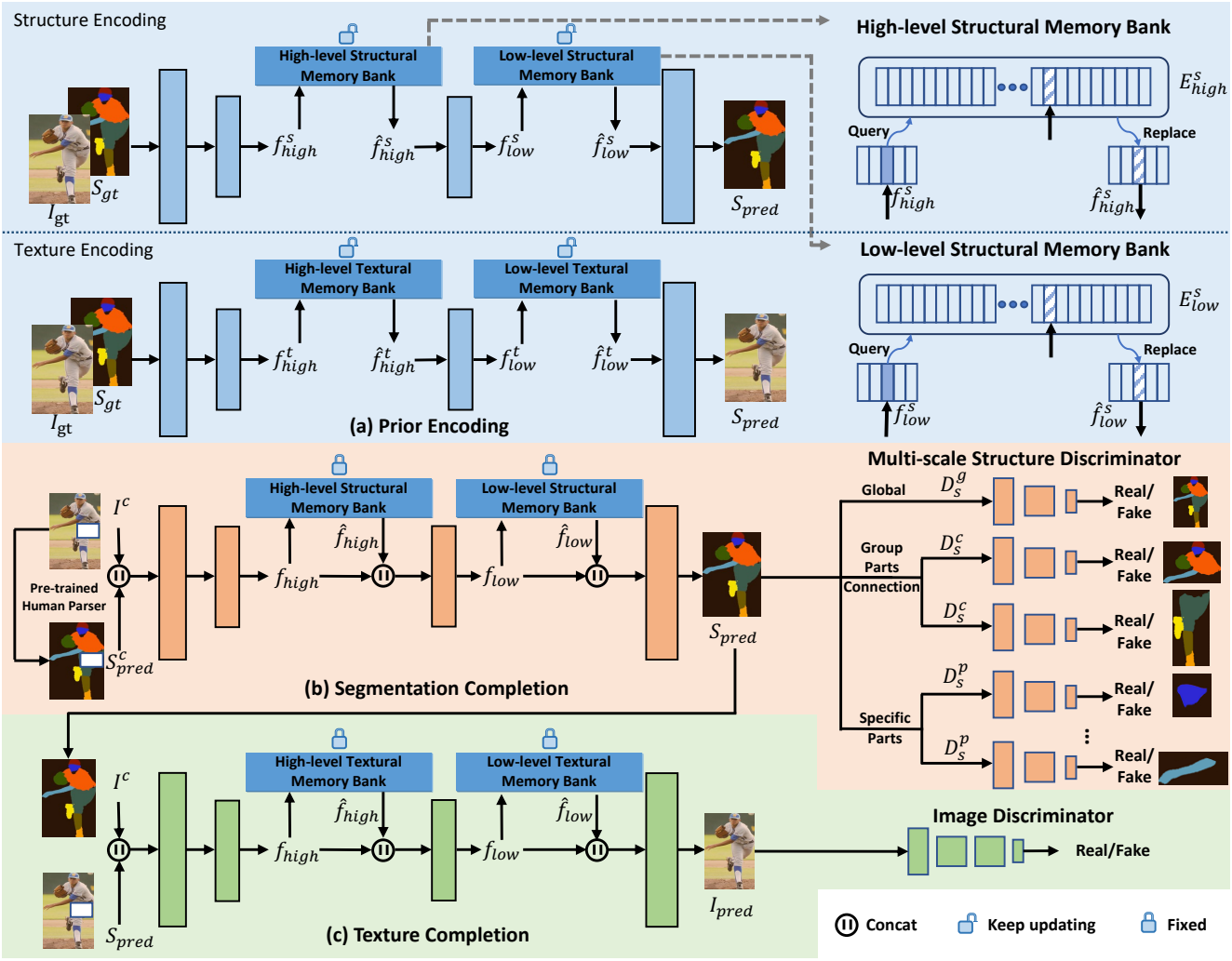

Figure 2. The overview of our proposed method. The model consists of three stages, including (a) Prior Encoding, (b) Segmentation Completion, and (c) Texture Completion. Note that two auto-encoders in (a) have different parameters.

optimized by minimizing the cross-entropy loss L_c:

$$L_c(S_{gt}, S_{pred}) = -\frac{1}{HW}\sum_{m=1}^{HW}\sum_{c=1}^{C} S_{gt}\log\left(S_{pred}\right). \quad (4)$$

The Texture Memory Bank. Following a similar strategy, we force the auto-encoder to maintain a low-level texture memory E^t_{low} and a high-level texture memory E^t_{high} via reconstructing the complete image I_{gt}. The model takes a complete image I_{gt} and the ground-truth segmentation map S_{gt} as input to reconstruct the complete image I_{pred}. We use an L-1 loss to constraint the model. During the reconstruction, texture memory bank could encode outside information of textures from other images.

Unlike the VQ-VAE [36, 28]: 1) we use the concatenation of images and segmentation maps as input to both auto-encoder for better representing the external information; 2) we maintain two types of dictionaries to make the memory bank module focus on structure information and texture information, respectively.

3.3. The Segmentation Completion

The segmentation completion module designs to infer the structure information via a learned structure memory bank and our proposed multi-scale structure discriminators. This module utilizes the corrupted image I^c and its estimated corrupted human segmentation map S^c_{pred} by the CE2P model [30] as a condition to reconstruct the complete structure of the human body S_{pred}. The prediction process can be written as:

$$S_{pred} = G_s(I^c, S^c_{pred}, E^s), \quad (5)$$

where G_s represents the segmentation completion module and E^s means the learned structure memory.

This module consists of an encoder, the learned bi-level structure memories, and a decoder, as shown in Fig-

ure 2 (b). The encoder takes the concatenation of I^c and S^c_{pred} as input and extracts the low-level feature map $f^s_{low} \in \mathcal{R}^{C\times 32\times 32}$ and the high-level feature map $f^s_{high} \in \mathcal{R}^{C\times 64\times 64}$. We use features from extracted feature maps f_{low} and f_{high} to retrieve the most similar features from two learned structure memories E^s_{low} and E^s_{high}, respectively. Further, they are reformed into the new low-level feature map $\hat{f}^s_{low}$ and high-level feature map $\hat{f}^s_{high}$. With the aid of structure memory banks, the module could borrow the valid information from visible regions or querying from other images to recover the missing similar structure. Also, we employ two skip-connections to concatenate the extracted feature map and the replaced feature map in two levels to maintain the identification of the original image and ensure that feature maps not worse than the original feature map. Finally, the decoder predicts the complete human segmentation map S_{pred}, which is a crucial to the next stage. The model is optimized by minimizing the cross-entropy loss L_c between S_{pred} and S_{gt}.

Multi-scale Structure Discriminators. We observe that although human body parts are flexible, the topological connections between these parts are relatively fixed. For example, the waist must connect the upper body and lower body, and arms connect to the upper body. Additionally, due to insufficient priors and unbalanced data, it is unlikely to identify the generated human segmentation map has a reasonable structure by a single discriminator.

Therefore, we design multi-scale structure discriminators as a regularizer to constrain the model generates segmentation maps with reasonable topological structure. The regularizer defines as $D_s = \{D^g_s, D^c_s, D^p_s\}$, where D^g_s is a global discriminator, D^c_s is a group parts connection discriminators, D^p_s is a group of specific part discriminators. The specific part discriminator D^p_s critics whether the corresponding part exists in the result, it is defined as $D^p_s = \{D^{\text{head}}_s, D^{\text{torso}}_s, D^{\text{arms}}_s, D^{\text{hands}}_s, D^{\text{legs}}_s, D^{\text{feet}}_s\}$. The group parts connection discriminator D^c_s critics topological relations between different parts in the predicted segmentation map to penalize abnormal combinations of human parts like feet appear in the upper body. Here, $D^c_s = \{D^{\text{upper}}_s, D^{\text{lower}}_s\}$. The global discriminator D^g_s critics the structural rationality of generated semantic segmentation maps.

We use the adversarial loss to optimize multi-scale structure discriminators. For example, the adversarial loss leveraged on the upper body discriminator can be written as:

$$L^u_{adv} = \mathbb{E}[\log D^u_s(S^{\text{upper}}_{gt})] + \mathbb{E}[1 - \log D^u_s(S^{\text{upper}}_{pred})], \quad (6)$$

where S^{upper}_{gt} and S^{upper}_{pred} denote the upper body in ground-truth segmentation maps and predicted segmentation maps.

The full objective function is as follows:

$$L_2 = \lambda^s_1 L_s + \lambda^s_2 L^g_{adv} + \lambda^s_3 \sum_i^B L^i_{adv}, \quad (7)$$

where L^g_{adv} is the loss of global discriminator, B = {head, torso, arms, hands, legs, feet, upper body, lower body}, and we set the hyper-parameters as $\lambda^s_1 = 6$, $\lambda^s_2 = 0.15$ and $\lambda^s_3 = 0.1$, respectively.

3.4. The Texture Completion

The texture completion module utilizes the predicted human segmentation map S_{pred} as conditions to further recover the appearance of a single person image. The process of texture completion to produce plausible images is

$$I_{pred} = G_t(I^c, S_{pred}, E^t), \quad (8)$$

where I_{pred} is the final result, and E^t is the texture memory.

The texture completion module consists of a generator, a bi-level texture memories, and a discriminator. Segmentation maps could guide the network to extract high-level information and further ensure the network would at least produce images with reasonable structure. Benefitting from the correlation prior, when the lost content is unlikely to borrow from the visible region, the model could query from the memory bank with the surroundings of the missing part to retrieve the detail. These priors guarantee the model to generate images with plausible textures.

To produce decent results, we employ L_1 loss, perceptual loss, style loss and feature matching loss over I_{pred} and I_{gt} to optimize G_t. We use a pre-trained VGG-16 network to compute the perceptual loss, and the loss can writes as:

$$L_p = \sum_{i=1}^{N} \frac{1}{H_i W_i C_i} \|\mathbb{P}^i(I_{gt}) - \mathbb{P}^i(I_{pred})\|_1, \quad (9)$$

where $\mathbb{P}^i(\cdot)$ denotes the feature maps from the i^{th} pooling layer of VGG-16. H_i, W_i and C_i corresponds to the height, weight and channel of the feature map from the i^{th} layer. Similarly, the style loss is defined as:

$$\begin{aligned} L_s = \sum_{i=1}^{N} \frac{1}{C_i \times C_i} \frac{1}{H_i W_i C_i} \|\mathbb{P}^i(I_{gt})(\mathbb{P}^i(I_{gt}))^T \\ - \mathbb{P}^i(I_{pred})(\mathbb{P}^i(I_{pred}))^T\|_1. \end{aligned} \quad (10)$$

Feature matching loss measure the distance between I and I_{pred}. Denoting the feature map from the i^{th} layer in D_t as $D^i_t(\cdot)$, the loss is as follow:

$$L_{fm} = \sum_{i=1}^{N} \|D^i_t(I_{gt}) - D^i_t(I_{pred})\|_1. \quad (11)$$

We also leverage an adversarial loss to train the model, which can be written as:

$$L_{adv} = \mathbb{E}[\log D_t(I_{gt})] + \mathbb{E}[1 - \log D_t(I_{pred})]. \quad (12)$$

We use hyper-parameters with value $\lambda^t_1 = 1$, $\lambda^t_2 = 0.1$, $\lambda^t_3 = 250$, $\lambda^t_4 = 10$ and $\lambda^t_5 = 10$ to balance the different

loss functions and we arrive at the full objective loss functions in stage 3 as:

$$L_3 = \lambda_1^t L_1 + \lambda_2^t L_p + \lambda_3^t L_{style} + \lambda_4^t L_{fm} + \lambda_5^t L_{adv}. \quad (13)$$

3.5. The Training Detail

In implementations, we separately train each module. First, we train the prior encoding module to maintain structure and texture memories. Then, these learned memories are fixed and plugged into the corresponding completion modules. Second, we train the semantic segmentation completion module to recover the human segmentation map, and finally, we train the texture completion module to generate the inpainted result.

4. Experiments

4.1. Experiments Set-up

Our method is built with Pytorch [26] on a NVIDIA GTX 1080 GPU. We apply the Adam optimizer with $\beta = (0.9, 0.999)$ to train all modules. In the stage of prior encoding, we fix the learning rate as 3e-4. In the following two stages, we set the learning rate as 1e-4 for the generators and 4e-4 for the discriminators.

Datasets. We conduct experiments on two large-scale datasets: the LIP dataset [8] and the Chictopia10K dataset [14]. The LIP dataset contains 50462 images and their corresponding semantic segmentation maps. There are 20 semantic labels, including 19 human parts labels and one background label. It is a challenging dataset, owing to the complex background and various actions in each image. To our best knowledge, none of the existing inpainting methods conduct experiments on the LIP dataset. The Chictopia10K dataset contains 17,706 images, and each image annotates with fine-grained semantic segmentation maps. The annotation has 21 classes, including 20 human part labels and one background label.

Metrics. To evaluate the performance of each method, we use three common metrics: mean Intersection over Union (mIoU), Peak Signal to Noise Ratio (PSNR), and Structural Similarity (SSIM). The mIoU evaluates the accuracy of predicted semantic segmentation maps. The PSNR and SSIM evaluate the quality of the generated images. For all metrics, high values mean better performance.

Baselines. We compare our method with the following baselines, EdgeConnect [24], DeepFill v2 [48], RN [49], RFR [16], and MEDFE [18].

Settings. We design the following two settings for fair comparisons. **1) Structure-based human completion.** It designs to evaluate the completion capability based on a corrupted image and a corrupted human segmentation map. Similar to the segmentation completion stage in our method, these methods first train to predict complete segmentation maps by taking the masked image and masked segmentation map as input. Since these methods are not designed for human body completion, they use the ground-truth human segmentation map as input in texture completion, while our method uses the predicted segmentation map. **2) Vanilla human completion.** It designs to evaluate the inpainting performance only based on a corrupted image.

4.2. The Performance Comparison

To comprehensively evaluate the performance, we compare our method among baselines on both the LIP dataset and the Chictopia10K dataset in the above two settings.

The experimental result under structure-based human completion setting and vanilla human completion setting are shown in Table 1 and Table 2, respectively. The numerical result shows that our method surpasses other methods in both settings. Besides, the modified previous methods under structure-based human completion setting always achieve better performance than previous methods under vanilla human completion setting, which indicates that the structure prior indeed benefits to recover the human body. In Table 2, compared with other modified methods, our

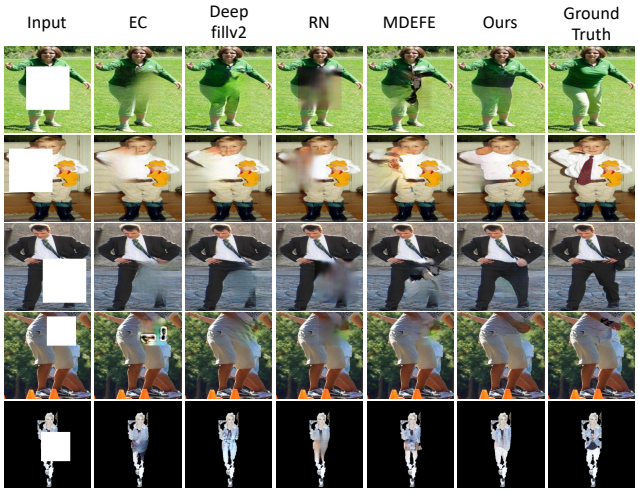

Figure 3. **Qualitative Analysis of the vanilla human completion**. Each column illustrates the input image, output of existing methods, the output of our method, and the ground-truth, respectively. It shows that only our method can produce results with plausible structure and texture. It further proves the effectiveness of our method for human completion. Best viewed with zoom-in.

	LIP		Chictopia10K	
	PSNR	SSIM	PSNR	SSIM
EdgeConnect [24]	19.35	0.8147	28.53	0.9168
DeepFill v2 [48]	23.06	0.8732	31.91	0.9676
RN [49]	20.25	0.8263	19.98	0.8361
RFR [16]	16.96	0.7343	16.00	0.7768
MEDFE [18]	22.54	0.8717	31.17	0.9597
Ours	**25.57**	**0.9139**	**36.58**	**0.9802**

Table 1. The performance comparison under a vanilla human completion setting on the LIP dataset and the Chictopia10K dataset.

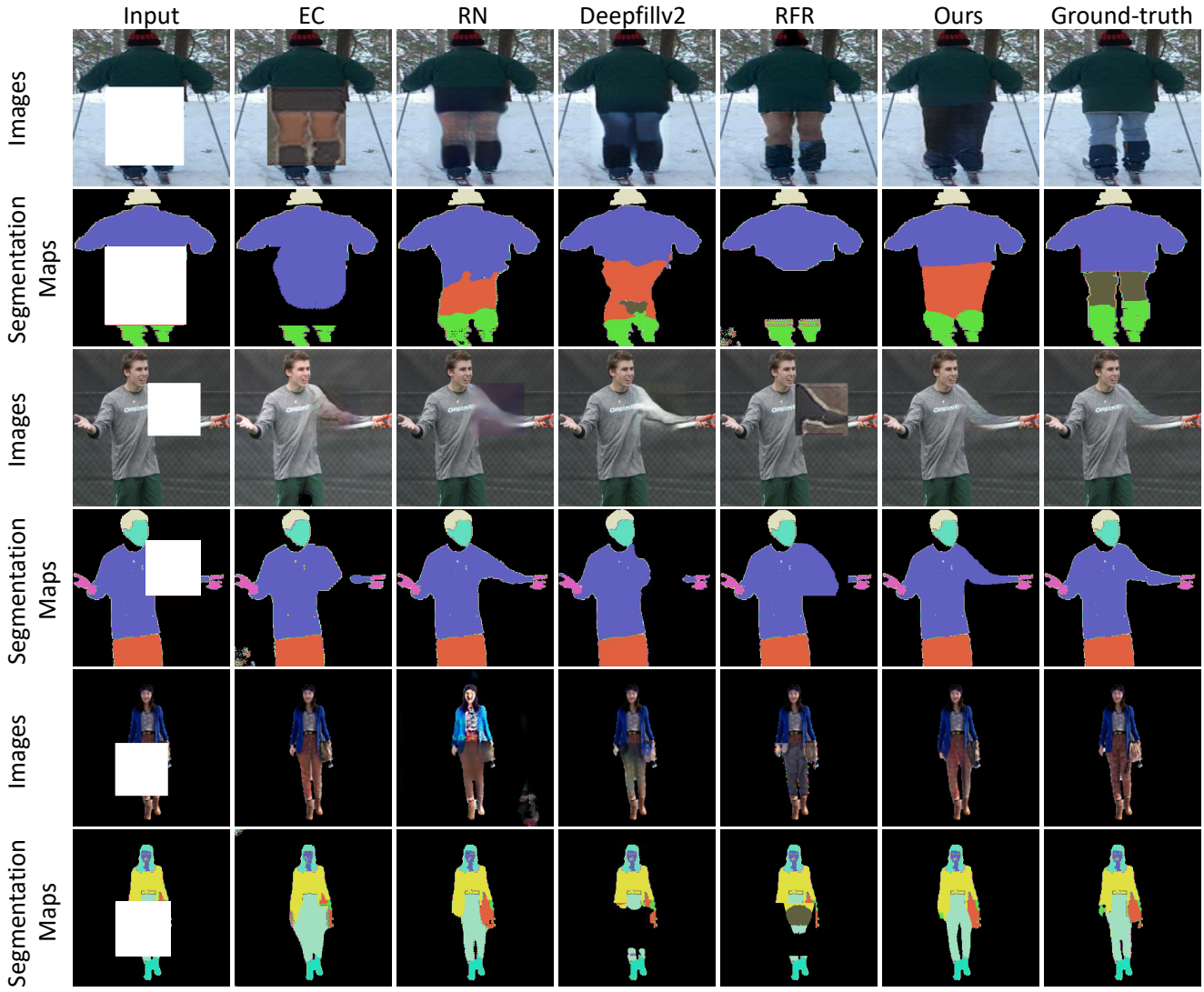

Figure 4. **Qualitative Analysis of the structure-based human completion**. Each row illustrates the input image, outputs of existing methods, the output of our method, and ground truth from left to right. The difference between generated segmentation maps illustrated the effectiveness of our correlation prior and multi-scale structure discriminators since only our model generates plausible segmentation maps. When recovering images, except for our model, the others take complete segmentation maps as the condition. The comparison between generated images shows that our model's performance exceeds the others, which proves the effectiveness of the correlation prior.

	LIP			Chictopia10K		
	mIoU	PSNR	SSIM	mIoU	PSNR	SSIM
EdgeConnect [24]	36.24%	20.53	0.8403	22.78%	28.98	0.9263
DeepFill v2 [48]	37.65%	23.38	0.8532	24.91%	30.01	0.9400
RN [49]	35.95%	21.64	0.8940	37.81%	29.11	0.8734
RFR [16]	10.15%	13.40	0.2672	11.20%	28.87	0.8229
Ours	**56.22%**	**25.57**	**0.9139**	**45.54%**	**36.58**	**0.9802**

Table 2. The performance comparison under a structure-based human completion setting on the LIP and the Chictopia10K dataset.

model could predict better segmentation maps and generate the image with higher quality. It further validates that leveraging structure and texture correlation prior could help the model recover corrupted images better.

Despite the quantitative comparison, we also illustrate some qualitative results of both vanilla and structure-based human completion settings in Figure 3 and Figure 4, respectively. In Figure 3, all existing methods produce images with obvious artifices in structure or texture. While conditioned with human body segmentation maps, these methods could generate plausible images with reasonable structures. This demonstrates that the structure prior indeed benefits to produce decent results, and it works for not only our method but also the others. Although these methods could generate results with reasonable body structure, the texture is still wired and inconsistent on the local body part due to the lack of valid texture information. However, with the help of the texture memory bank, our method could synthesize plausible results in both structure and texture.

Figure 5. **Visualization for Free-form Occlusions**. The first row shows images with free-form occlusions and the second row shows recovered images produced by our model. It shows that when handling free-form masks, our model still generates images with decent structures and nice textures, which indicates that our model could recover corrupted images with free-form occlusions.

Correlation Prior	Regularizer	mIoU	PSNR	SSIM
-	-	51.96%	20.07	0.7969
✓	-	53.78%	22.99	0.8561
-	✓	53.56%	22.63	0.7819
✓	✓	**56.22%**	**25.57**	**0.9139**

Table 3. The experimental result of ablation studies on the LIP dataset. Regularizer denotes multi-scale structure discriminators.

4.3. Ablation Studies

To verify the effectiveness of each component in our proposed method, we conduct ablation studies on the LIP dataset, and experimental results are shown in Table 3.

The effectiveness of the external prior. To explore the impact of the structure and texture correlation prior, we train an auto-encoder as baseline including a segmentation completion stage and a texture completion stage to directly generate complete images, without the correlation prior and multi-scale structure discriminators. Results are shown in Table 3. The gap between the first two rows and that of the last two rows indicates the model could benefits from the structure and texture correlation prior.

The effectiveness of the topological structure prior. We also train the model without the multi-scale structure discriminators as baseline to validate its effectiveness. Results are shown in Table 3. We can see that our method outperforms the baseline, which indicates the regularizer could help the model generate more plausible segmentation maps and further facilitate the model to produce decent images.

Recovering images with free-form occlusions. Except for square occlusions, we also use free-form masks to produce input corrupted images. The visual results are illustrated in Figure 5. The generated images have decent structures and nice textures, which indicates that our model could perform well on free-form occlusions.

Random masks in the testing phase. We leverage many random masks to block different regions in the same image to validate the model's performance. Figure 6 illustrates the visual result. It shows that no matter the whole human part is masked or several human parts are partially masked, our model can produce decent images.

Figure 6. **Visualization for Random Masks**. Using a fixed image, we randomly block different regions on it as inputs. The figure illustrates the results produced by our model. We can observe that whatever the whole human part is masked or several human parts are partially masked, the model can generate plausible results.

5. Conclusion

In this paper, we propose a novel framework to recover corrupted single person images. We found that leveraging human semantic segmentation maps could better guide the model to generate plausible results in both structure and texture. We design a structural and textural memory bank module, which enables the model to infer the missing content with the visible region in the image and query from the outside information. We also design multi-scale structure discriminators to regularize the model to generate a reasonable topological structure of human bodies. Extensive experiments show that our method outperforms others and produce more decent images.

6. Acknowledgements

The work was supported by National Key R&D Program of China (2018AAA0100704), NSFC #61932020, Science and Technology Commission of Shanghai Municipality (Grant No. 20ZR1436000), and Shuguang Program supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2

[3] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018. 2

[4] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9026–9035, 2019. 3

[5] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8120–8128, 2020. 3

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 2

[7] Ruohan Gao and Kristen Grauman. On-demand learning for deep image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1086–1095, 2017. 2

[8] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 3, 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 1

[11] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4491, 2019. 1, 3

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[14] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 853–862, 2017. 6

[15] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5962–5971, 2019. 2

[16] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 2, 6, 7

[17] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 2

[18] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. *Proceedings of the European Conference on Computer Vision*, 2020. 6

[19] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4170–4179, 2019. 2

[20] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 2

[21] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5904–5913, 2019. 2

[22] Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Dailan He, and Aishan Liu. Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In *IJCAI*, pages 3123–3129, 2019. 2

[23] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 3

[24] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2, 6, 7

[25] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 123–138, 2018. 2

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 2

[28] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14866–14876, 2019. 4

[29] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 181–190, 2019. 2

[30] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4814–4821, 2019. 2, 3, 4

[31] Min-cheol Sagong, Yong-goo Shin, Seung-wook Kim, Seung Park, and Sung-jea Ko. Pepsi: Fast image inpainting with parallel decoding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11360–11368, 2019. 2

[32] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 1

[33] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[35] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10521–10530, 2019. 2

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017. 4

[37] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015. 3

[38] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5703–5713, 2019. 2, 3

[39] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8929–8939, 2020. 2, 3

[40] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in neural information processing systems*, pages 331–340, 2018. 2

[41] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 293–301, 2019. 3

[42] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6769–6778, 2017. 3

[43] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5840–5848, 2019. 2

[44] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17, 2018. 1

[45] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6721–6729, 2017. 1, 2

[46] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 2

[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 1, 2

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 2, 6, 7

[49] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region normalization for image inpainting. In *AAAI*, pages 12733–12740, 2020. 6, 7

[50] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 1486–1494, 2019. 2

[51] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 2

[52] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020. 2

[53] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2

[54] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 266–274, 2019. 1, 3