Large Occluded Human Image Completion via Image-Prior Cooperating

Hengrun Zhao*[1], Yu Zeng*[2], Huchuan Lu†[1], Lijun Wang[1],

[1]Dalian University of Technology
[2]Johns Hopkins University
zhaohengrun@mail.dlut.edu.cn, yzeng22@jhu.edu, lhchuan@dlut.edu.cn, ljwang@dlut.edu.cn

Abstract

The completion of large occluded human body images poses a unique challenge for general image completion methods. The complex shape variations of human bodies make it difficult to establish a consistent understanding of their structures. Furthermore, as human vision is highly sensitive to human bodies, even slight artifacts can significantly compromise image fidelity. To address these challenges, we propose a large occluded human image completion (LOHC) model based on a novel image-prior cooperative completion strategy. Our model leverages human segmentation maps as a prior, and completes the image and prior simultaneously. Compared to the widely adopted prior-then-image completion strategy for object completion, this cooperative completion process fosters more effective interaction between the prior and image information. Our model consists of two stages. The first stage is a transformer-based auto-regressive network that predicts the overall structure of the missing area by generating a coarse completed image at a lower resolution. The second stage is a convolutional network that refines the coarse images. As the coarse result may not always be accurate, we propose a Dynamic Fusion Module (DFM) to selectively fuses the useful features from the coarse image with the original input at spatial and channel levels. Through extensive experiments, we demonstrate our method's superior performance compared to state-of-the-art methods.

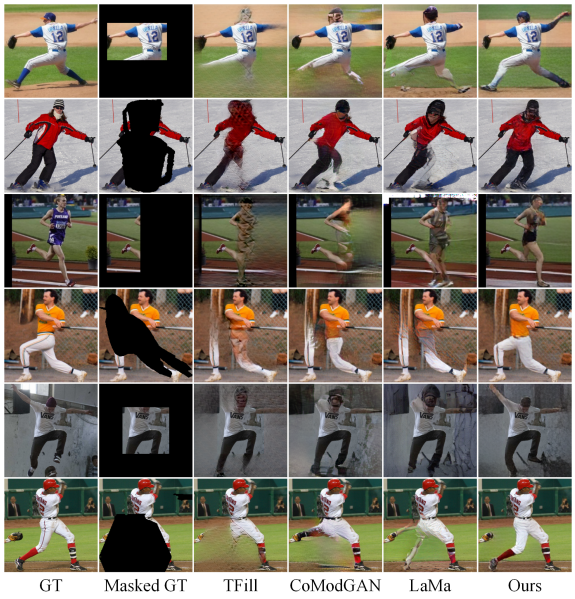

Figure 1: A demonstration of our method and several representative image completion methods (Zheng et al. 2022a; Zhao et al. 2021a; Suvorov et al. 2022) for completing images under various occlusions.

Introduction

Image completion (a.k.a. image inpainting) refers to the task of reconstructing the missing part of partially visible images based on the information of visible parts in the image. It has been an active research topic in the past decades. Traditional methods (Efros and Freeman 2001; Kwatra et al. 2005) and earlier deep learning-based methods (Yu et al. 2018, 2019) are mainly focused on the background inpainting problem, *i.e.* completing the background part in an image, which achieved superior performance and have been incorporated in many practical applications. Recent research has started paying more attention to a more difficult object inpainting problem, *i.e.* completing missing or partially missing objects in an image (Zhao et al. 2021b; Zeng, Lin, and Patel 2022; Xie et al. 2023). Compared to background inpainting, object inpainting is a much harder problem and requires a higher-level semantic understanding of image data. Although recent advance in deep generative models (Suvorov et al. 2022; Zhao et al. 2021a; Zheng et al. 2022a) have shown great promise, object inpainting remains a significant and challenging problem within the field of computer vision.

Among all objects, the completion of the human in an image presents unique challenges due to both its intrinsic difficulties and the elevated scrutiny from human viewers. As human bodies have distinct features from the surrounding environment, the traditional inpainting principle based on borrowing features from the background is no longer effective. Due to the presence of clothes, different parts of the human body can possess unique features, which brings additional challenges for modeling. In addition, compared to other objects, human vision is more adapted to perceive the

*These authors contributed equally.
†Corresponding author

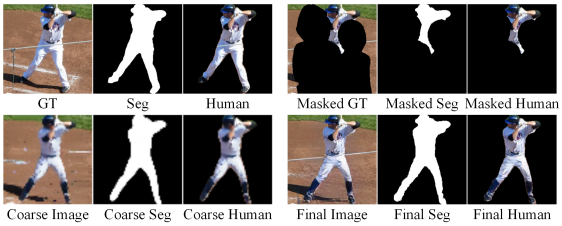

Figure 2: Example results. We show a collection of images and corresponding human segmentation maps and human areas in image, comprising the original image, the masked image, and the images that were completed by the coarse and refinement networks.

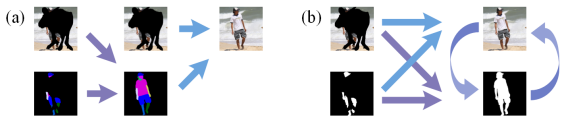

Figure 3: Illustration of prior guided completion process. (a) Classical prior guided methods such as (Wu et al. 2019; Zhao et al. 2021b; Han et al. 2019) usually complete the prior information first, and then complete the image. (b) Our proposed LOHC completes both prior and image simultaneously, enabling a bidirectional interaction between the prior and image information.

human body. Therefore, even small distortions and artifacts can lead to very unpleasant results.

Fortunately, while human bodies can make a variety of postures and actions, they still have a roughly fixed form and topology, allowing the utilization of prior information such as segmentation maps (Wu et al. 2019; Zhao et al. 2021b; Han et al. 2019) and posture (Lassner, Pons-Moll, and Gehler 2017; Balakrishnan et al. 2018; Men et al. 2020; Grigorev et al. 2019). Most previous approaches for human image inpainting first complete a partially occluded human parsing map and then use it as a prior to guide the completion of the images. While these approaches have demonstrated promising results, their effectiveness is limited, especially in cases where a large area of the human body is occluded, as shown in Fig. 1. This is because completing a parsing map with a large area missing is not significantly easier for the model than completing the corresponding image. Over-reliance on inpainted parsing maps can lead to worse image inpainting results when the inpainted parsing maps are inaccurate. Furthermore, the domain gap between the parsing map and the image makes it difficult for the model to exploit the information in the parsing map to guide the image completion process, which limits the guidance effect of the parsing map.

To tackle these challenges, we propose a two-stage deep learning network with an image-prior cooperative completion strategy for human image completion. The example of completion results is shown in Fig. 2. Different from traditional inpainting methods that rely solely on the guidance of pre-completed prior information(Wu et al. 2019; Zhao et al. 2021b; Nazeri et al. 2019; Yang and Guo 2020), the proposed method completes the image and human body segmentation map simultaneously and uses the completed segmentation map to provide additional supervision on the human body area in the image. As shown in Fig. 3, compared to the unilateral guidance of the segmentation map for image completion, our cooperative completion strategy allows the information of the segmentation map to always interact with the image information in the process of the simultaneous completion of the segmentation map and the image. The cooperative completion strategy enables the network to better learn and understand the relationship between people and the environment in the image with the help of the segmentation map, which ultimately leads to better human completion results. In addition, We found that the existing discriminators cannot provide satisfactory guidance on both the local texture details and the human body structure. Therefore, we use two discriminators that focus on the texture details and the global human body structure in the image, respectively, ensuring that the generated human image has both a reasonable structure and realistic details.

We summarize our contributions as follows:

- We proposed an innovative human image completion strategy based on the prior of the human body segmentation map, allowing the network to fully leverage the prior information to complete the image more accurately.
- We introduce a human image completion network that can realistically complete the human body in an image, even in cases where large areas are occluded. Code is available at https://github.com/ZhaoHengrun/LOHC.
- We develop a set of training strategies for human image completion that effectively incorporate both human and environmental factors, improving the overall quality of the completed images.

Related Work

General Image Completion

Traditional image completion methods, such as patch-based and color diffusion-based methods(Efros and Freeman 2001; Kwatra et al. 2005; Barnes et al. 2009; Ballester et al. 2001; Chan and Shen 2001; Criminisi, Pérez, and Toyama 2004), often copy or propagate existing image content to fill in occluded areas. These methods may produce blurry and unrealistic results when applied to complex visual scenes. In the past few years, most of the mainstream works have focused on deep learning-based solutions that incorporate higher-level image understanding (Suvorov et al. 2022; Zheng et al. 2022b; Zhao et al. 2021a; Iizuka, Simo-Serra, and Ishikawa 2017; Park et al. 2020; Pathak et al. 2016; Yi et al. 2020; Yu et al. 2018, 2019; Zeng et al. 2021). These methods leverage generative adversarial networks (GANs) (Goodfellow et al. 2020) to generate complex structures and high-resolution details that are perceptually indistinguishable from real images. By combining advanced visual features with semantic understanding, GAN-based methods significantly improve the effectiveness of image completion, achieving high-quality results even in the presence

of large occluded areas and complex structures. Diffusion models (Sohl-Dickstein et al. 2015) have recently demonstrated amazing image generation capabilities and have also achieved excellent results in the field of image completion (Rombach et al. 2022; Lugmayr et al. 2022). Nevertheless, these methods often demand substantial computational resources, thereby constraining their research and practical application.

Object Image Completion

Completing objects in images is a more challenging task compared to completing general content. Xiong et al. (Xiong et al. 2019) propose a foreground-aware image inpainting method that utilizes explicit contour guidance for image completion. Ke et al. (Ke, Tai, and Tang 2021) propose a video object inpainting network that models the shape and boundary of the object separately in the frame sequence to achieve both shape completion and texture generation of the object. Zeng et al. (Zeng, Lin, and Patel 2022) proposed a Contextual Object Generator (CogNet), which innovatively completes the occluded area by generating an object based on contextual content and the shape of the occluded mask. Xie et al. (Xie et al. 2023) proposed SmartBrush for completing a missing region in an image with an object using both text and shape guidance.

The human body is a highly intricate object, making its completion even more challenging. In recognition of this, several researchers have focused on the human image completion task. Han et al. (Han et al. 2019) proposed Fashion Inpainting Networks (FiNet), which can reconstruct missing clothing parts in fashion portrait images based on partially missing parsing maps. Wu et al. (Wu et al. 2019) propose a two-stage deep learning framework for portrait image completion that extracts a complete human body structure using a human parsing network in the first stage and fills the unknown area in the image using an image completion network in the second stage. Zhao et al. (Zhao et al. 2021b) propose a prior-based human image completion method (PBHC) that uses structure and structure-texture correlation priors to recover a reasonable human shape and compensate for occluded texture.

However, their methods are limited by the effectiveness of human parsing map completion. Due to the lack of involvement of advanced semantic information, the completion of the human parsing map often fails to produce satisfactory results when there are large occluded areas or the character has complex postures or self-occlusions, resulting in poor quality of the final generated image.

Method

Image-Prior Cooperative Completion

Given an image I with a binary mask M indicating the area to inpaint, a image completion model G typically generates a completed image I_g based on the masked image $I_m = I \odot M$ and the mask M:

$$I_g = G(I_m, M), \tag{1}$$

where $\odot$ represents element-wise multiplication. In object inpainting, object prior (such as segmentation map) is often used as guidance. These prior-based object completion methods typically complete the segmentation prior S_m first and then proceed to complete the image based on the completed prior using separate models G_S and G_I:

$$S_g = G_S(I_m, S_m, M) \tag{2}$$

$$I_g = G_I(I_m, S_g, M) \tag{3}$$

Taking inspiration from recent studies (Wang et al. 2023a,b; Ye and Xu 2022; Jain et al. 2023; Xi et al. 2022) that highlight the benefits of multitask joint learning compared to separate single-task learning in unified perceptual models, our work delves into the image-prior cooperative completion strategy for human image completion. Instead of performing a prior pre-completion process using a separate network, we model the joint completion of a masked image and prior using an image-prior co-completion process.

Given an incomplete image I_m and the corresponding incomplete segmentation map S_m, our co-completion process aims to produce the competed image I_g and segmentation prior S_g simultaneously with a unified model G:

$$I_g, S_g = G(I_m, S_m, M) \tag{4}$$

The image-prior cooperative completion strategy allows for the prior information to be incorporated throughout the entire joint completion process and combined more effectively with image features, leading to a stronger understanding of human body representation information and ultimately improving the quality of completed human images.

Overall Architecture

We model the image-prior cooperative completion process with a large occluded human image completion (LOHC) network. It consists of a coarse network and a refinement network, as depicted in Fig. 4. The coarse network aims to complete the overall structure of the missing area at a lower resolution. It takes the downsampled occluded image I_m, segmentation map S_m and mask M as input and generates a low-resolution complete image I_c and a segmentation map Sc. Inspired by the excellent contextual interaction capabilities of Masked Autoencoder (MAE) (He et al. 2022), we use an auto-regressive transformer architecture that operates at image patches. Subsequently, the refinement network generates a full-resolution complete image I_g and human body segmentation map S_g based on the I_c and S_c.

Coarse Network

The coarse network is a transformer-based encoder-decoder network. In order to enable the network to encode the human body area more effectively, we utilize two encoders (Encoder I and H in Fig. 4) to capture the information of the human body area and the environment separately. Then the two sets of encoder features are concatenated and decoded by a single large decoder into image patches. To reduce the computing cost and focus on the overall structure of the image, we scale I_m, S_m, and M to 64×64 and split them into non-overlapping 4×4 patches to be processed by the coarse network. We remove all of the patches that have been partially or completely occluded by M from the encoders' input, retaining only the patches that are completely unoccluded. The decoder then learns to predict the full set of patches from the ground-truth unoccluded image and segmentation map.

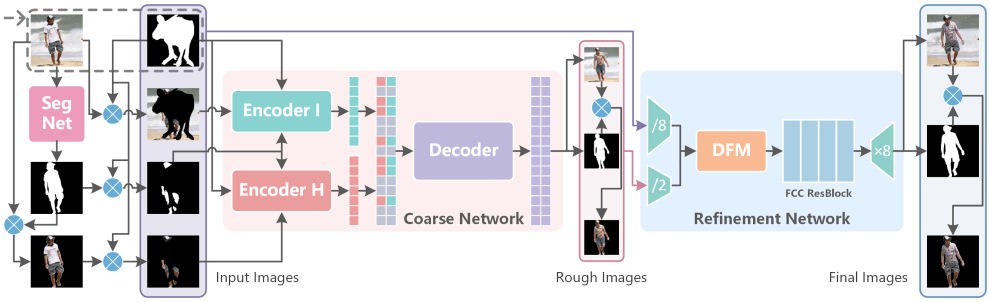

Figure 4: The pipeline of our proposed human image completion network. Given an incomplete image and the corresponding incomplete segmentation map, the coarse network generates a coarse completed image and segmentation map simultaneously at low resolution. The refinement network then completes the final high-resolution image and segmentation map based on the coarse result and the original incomplete image.

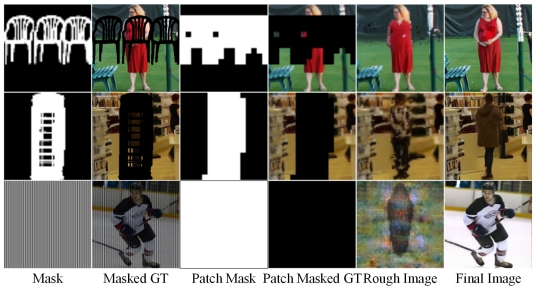

Figure 5: Mask example. In cases where the mask is small and dense, the patch mask caused by MAE may lead to significant expansion of the mask. Nevertheless, thanks to the dynamic fusion module, the refinement network can still selectively utilize effective information to generate realistic content.

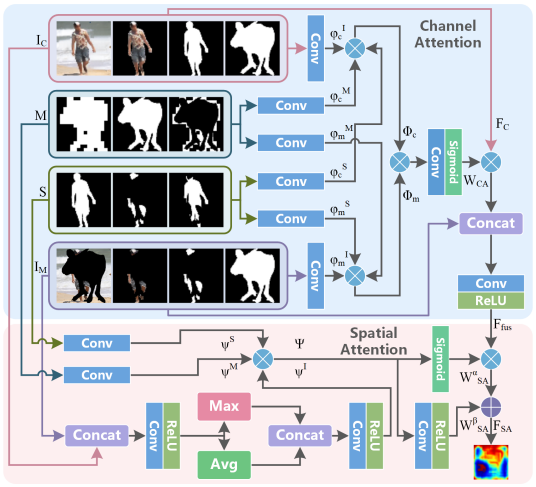

Figure 6: Architecture of our proposed dynamic fusion module (DFM), which includes two parts: channel attention (blue area) and spatial attention (pink area).

Refinement Network

Our refinement network employs an encoder-decoder structure and utilizes Fast Fourier Convolution (FFC) blocks (Suvorov et al. 2022). The capability to build long-range dependency of transformers makes the coarse network suitable for processing largely occluded images. However, as depicted in Fig. 5, its patch-wise processing scheme ignores useful information in partially occluded patches, resulting in a significant deterioration in the quality of the completed image, especially when the mask is small and dense.

Fortunately, the refinement network based on convolution blocks has inherent advantages in completing small and spotty occluded areas. In order to combine the coarse network's large-area construction capability with the refinement network's nearby pixel completion capability, we selectively fuse the features from I_c and I_m at the pixel level with a Dynamic Fusion Module (DFM) before processing by the refinement network. As shown in Fig. 6, this module comprehensively considers factors such as image content, mask, and segmentation map, and applies weighted aggregation separately to the channels and spatial dimensions.

Specifically, the channel attention mechanism in DFM generates a pixel-wise attention weight for each channel to modulate the features from I_c before the fusion operation:

$$F_c = W_{CA} \odot F_c, \tag{5}$$

where the weight W_{CA} is calculated as:

$$W_{CA} = \Phi_c \odot \Phi_m \tag{6}$$

$$\Phi_i = \phi_i^I \odot \phi_i^S \odot \phi_i^M \tag{7}$$

where $\phi_i^I, \phi_i^S, \phi_i^M$ are computed as follows,

$$\phi_i^I = F(I_i) \tag{8}$$

$$\phi_i^S = F(S_c, S_m, |S_c - S_m|) \tag{9}$$

$$\phi_i^M = F(M_p, M, |M_p - M|) \tag{10}$$

where F represents convolutional blocks, S_c represents the segmentation map completed by the coarse network, S_m represents the original occluded segmentation map, and M_p represents the mask expanded by patching in the coarse network.

The spatial attention is modulated by two pixel-by-pixel weights that are multiplied and added to the features F_{fus} fused through the channel attention, respectively:

$$F_{SA} = W_{SA}^{\alpha} \odot F_{fus} + W_{SA}^{\beta} \tag{11}$$

where W_{SA}^{α} and W_{SA}^{β} are obtained by embedding Ψ, which calculated as:

$$\Psi = \psi^I \odot \psi^S \odot \psi^M \tag{12}$$

Where ψ^S, and ψ^M are mask embedding and segmentation embedding, ψ^I is the spatial attention embedding of the features before fusion, and obtained similar to (Woo et al. 2018).

Loss Functions

Following common practice in the previous studies, we design our loss function by combining an L1-based loss, an adversarial loss and a perceptual loss. We apply L1 loss between the completed image / segmentation map and the ground truth. To provide additional supervision for the human area, we apply an additional L1 loss for the human area. Our L1-based loss term is as follows,

$$\mathcal{L}_{rec} = (\|I - I_g\|_1 + \|S - S_g\|_1 + \tag{13}$$

$$\|H - H_g\|_1) \odot (1 - M) \tag{14}$$

$H = S \odot I$ and $H_g = S_g \odot I_g$ refer to the human area in the ground-truth image and the completed image, respectively. For adversarial loss $\mathcal{L}_D$, we use two PatchGAN discriminators (Isola et al. 2017). One discriminator D^{local} only takes the image as input and focuses on matching the local statistics to the ground-truth image patches. The other discriminator D^{global} takes both the image and the segmented human area and segmentation map as inputs to improve global human body structure. Please refer to (Isola et al. 2017; Goodfellow et al. 2020) for the specific definition of the adversarial loss. We also adopt the feature matching loss proposed (Wang et al. 2018), and the total adversarial loss is computed as follows,

$$\mathcal{L}_{adv} = \mathcal{L}_G^{local} + \mathcal{L}_G^{global} + 10\mathcal{L}_{fm}^{local} + 10\mathcal{L}_{fm}^{global}, \tag{15}$$

where $L_G^{local}, L_{fm}^{local}$ are the vanilla adversarial loss and feature matching loss for D^{local}, and $L_G^{global}, L_{fm}^{global}$ are those for D^{global}. We use an architecture similar to those in (Isola et al. 2017) for the discriminators.

The total loss is defined as follows:

$$\mathcal{L} = 10\mathcal{L}_{rec} + 5\mathcal{L}_{adv} + 60\mathcal{L}_{pl}, \tag{16}$$

where L_{pl} represents the high receptive field perceptual loss proposed in (Suvorov et al. 2022). We apply the same loss terms to both the coarse and refinement networks except for L_{pl}, which is only applied to the refinement network.

Experiments

Datasets

Currently, there are no datasets specifically designed for human image completion tasks. Previous works utilize the LIP dataset (Gong et al. 2017) and some fashion portrait parsing datasets (Liang et al. 2015; Lassner, Pons-Moll, and Gehler 2017) for training and evaluation. However, only the rectangular area of the human body is preserved in the images of the LIP dataset. This restriction can make it hard for image completion methods to learn and complete specific human structures based on the image content since the aspect ratio of the image becomes closely tied to the approximate posture and position of the person in the image. Moreover, if the image is uniformly scaled to a consistent size, it may disrupt the original scale and structure of the image content, further complicating the image completion process. As for the fashion portrait analysis datasets, the human posture is too simple and singular, which is also not suitable for our task. As a result, we opted to use the AHP dataset (Zhou et al. 2021), which consists of 56,599 images collected from several large-scale instance segmentation and detection datasets, including COCO (Lin et al. 2014), VOC (Everingham et al. 2010), LIP (Gong et al. 2017), Objects365 (Shao et al. 2019) and OpenImages (Kuznetsova et al. 2020).

To prepare the dataset, we first cropped the square area where the human was located in the image, uniformly scaled to a size of 256 $\times$ 256. For testing, we used the original dataset's validation set, which amounted to 3400 pieces, while for training, we utilized the original dataset's training set of 53199 images.

To generate masks for our training and testing data, we utilized the same methods as TFill (Zheng et al. 2022a) for generating central square masks, random regular masks, and random irregular masks. In addition, we also adopted the mask generation method used in Deepfill v2 (Yu et al. 2019) to generate a set of irregular masks. We randomly generated these four types of masks in real time during training. We evaluated the performance of methods on the central square masks, rectangular masks and three sets of object masks provided by (Zeng et al. 2020).

Implementation Details

Our network training procedure consists of three distinct stages. First, we pre-train the coarse network refer to (He et al. 2022). To encourage long-distance pixel associations, we employ a random mask with a 90% masking ratio, as the adjacent visible pixels under a 75% masking ratio random mask are still relatively close to each other. We then retain the entire pre-trained coarse network and subject it to normal training. Finally, we train the refinement network.

Thanks to the reduced computational requirements at lower resolutions, we were able to design a deeper coarse network without sacrificing performance. Specifically, our coarse network includes two encoders, each with 32 layers and a width of 128. The decoder has 32 layers and a width of 256, which matches the combined width of the two encoder tokens. Similar to LaMa (Suvorov et al. 2022), our refinement network employs 4x downsampling and upsampling for image features, and utilizes 9 FFC blocks to complete image features at lower scales.

We obtain the segmentation map of human body through U-Net (Ronneberger, Fischer, and Brox 2015).

In our training process, we utilized Adam as the optimizer for all components. We trained the coarse network with a batch size of 128 and set the learning rate to 1e-4. For the refinement network, we used a smaller batch size of 16 and set the learning rate to 1e-3. The learning rates of all discriminators are set to 1e-6. We use the Pytorch framework for our implementation and train on an Nvidia A100 GPU.

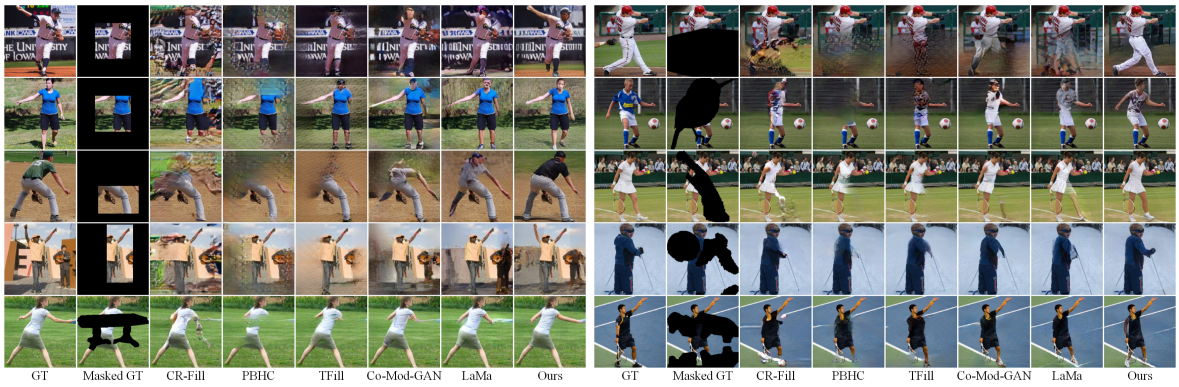

Figure 7: Visual comparison of completion images of our method and other methods. Zoom-in to see the details.

Mask	Metric	Whole Image							Human Area						
		Mask Ratio	CR-Fill	PBHC	TFill	Co-Mod	LaMa	Ours	Mask Ratio	CR-Fill	PBHC	TFill	Co-Mod	LaMa	Ours
Center	PSNR	75.00%	12.920	14.930	14.818	14.543	14.517	**14.931**	42.88%	21.721	22.729	22.973	22.946	22.877	**23.722**
	SSIM		0.4568	0.5204	0.5311	0.5359	0.5249	**0.5486**		0.9176	0.9210	0.9225	0.9255	0.9248	**0.9281**
	LPIPS		0.4256	0.3975	0.3779	0.3579	0.3351	**0.3165**		0.0804	0.0796	0.0715	0.0668	0.0646	**0.0610**
	FID		93.894	74.791	55.430	56.192	40.649	**37.468**		25.280	42.565	17.103	15.368	13.192	**11.957**
Rectangle	PSNR	75.26%	12.772	14.270	14.093	14.180	14.400	**14.626**	57.13%	20.532	21.828	21.710	21.865	21.881	**22.382**
	SSIM		0.4452	0.5058	0.4997	0.5213	0.5139	**0.5257**		0.8870	0.8937	0.8906	0.8972	0.8965	**0.8994**
	LPIPS		0.4477	0.4488	0.4230	0.4049	0.3686	**0.3515**		0.1134	0.1124	0.1084	0.0995	0.0950	**0.0923**
	FID		105.173	89.363	57.808	61.248	45.974	**40.246**		38.341	51.193	26.697	24.921	19.849	**19.615**
VOC	PSNR	42.27%	18.392	19.605	19.299	19.603	19.684	**20.458**	57.51%	22.550	23.153	23.489	23.736	23.932	**24.016**
	SSIM		0.7077	0.7256	0.7088	0.7398	0.7278	**0.7501**		0.8922	0.8954	0.8921	0.9019	0.8991	**0.9053**
	LPIPS		0.2384	0.2309	0.2234	0.1867	0.1896	**0.1669**		0.1070	0.1046	0.1004	0.0919	0.0901	**0.0857**
	FID		40.236	42.597	31.967	19.882	21.697	**18.606**		32.771	47.093	23.560	19.173	18.537	**18.484**
XPIE	PSNR	34.94%	19.125	20.309	20.264	20.388	20.476	**21.309**	54.10%	22.937	23.479	24.074	24.159	24.376	**24.543**
	SSIM		0.7588	0.7729	0.7610	0.7854	0.7747	**0.7952**		0.9007	0.9032	0.9014	0.9097	0.9069	**0.9131**
	LPIPS		0.2001	0.1928	0.1792	0.1519	0.1560	**0.1357**		0.0991	0.0961	0.0893	0.0827	0.0821	**0.0776**
	FID		31.221	33.806	24.488	15.083	16.849	**14.571**		29.036	42.545	20.120	15.512	15.731	**15.217**
HKU IS	PSNR	32.72%	20.303	21.693	21.486	21.727	21.660	**22.665**	40.30%	26.048	26.661	27.041	**27.358**	27.346	26.994
	SSIM		0.7858	0.7975	0.7869	0.8102	0.7997	**0.8193**		0.9282	0.9292	0.9282	0.9353	0.9327	**0.9369**
	LPIPS		0.1779	0.1608	0.1577	0.1289	0.1331	**0.1152**		0.0739	0.0699	0.0660	0.0595	0.0596	**0.0563**
	FID		25.676	26.053	20.517	12.534	13.565	**12.146**		17.034	24.236	12.746	9.770	9.680	**9.502**

Table 1: Quantitative comparison of complete image and human area in image. Bold for best results.

Performance of Our Approach

To evaluate the effectiveness of our proposed method, we compared it with several state-of-the-art image completion methods, including LaMa (Suvorov et al. 2022), Co-Mod-GAN (Zhao et al. 2021a), TFill (Zheng et al. 2022a), and CR-Fill (Zeng et al. 2021), as well as a human body completion method PBHC (Zhao et al. 2021b). We quantitatively measured the performance of each method using metrics such as PSNR, SSIM, FID, and LPIPS.

Quantitative evaluation. For the human visual experience, the quality of the human area in the image has a more significant impact on the authenticity of the image's appearance. Therefore, we conducted additional evaluations on the human area in the image based on the segmentation map of the original image. As illustrated in Table 1, LOHC achieves state-of-the-art performance on the AHP dataset. In particular, LOHC exhibits significantly superior performance over other methods when completing images with large areas of occluded human body parts.

Visual quality. We provide a visual comparison between our method and existing approaches. As shown Fig. 7 , our method is able to more completely and reasonably complete the human body, while maintaining a clear boundary between the character and the environment.

User study. We randomly selected 50 images from the AHP validation set and occluded them with 10 masks. We invited 10 human scorers for evaluation, the scorers were presented with the completed images by different methods in random order and asked to select the best result. The number of user preferences for each method is shown in Table 2. Our method was most frequently preferred by the human scorers.

Ablation Study

In this part, we study the specific effect of each part of our method. The following experiments all use unified model parameters and experimental settings. To speed up the experiment, we no longer conduct pre-training on the coarse network and only test the images on VOC masks. The ex-

Mask Mask Ratio	Center 75.00%	Rectangle 75.26%	VOC 42.27%	XPIE 34.94%	HKU IS 32.72%
Ours	**82**	**56**	**44**	**72**	37
LaMa	11	26	20	8	**40**
Co-Mod-GAN	3	12	26	14	14
TFill	4	6	6	5	5
PBHC	0	0	1	1	2
CR-Fill	0	0	3	0	2

Table 2: Subjective comparison. Bold for best results.

Category	Option	PSNR	SSIM	LPIPS	FID
Baseline		19.8768	0.7375	0.2009	32.4197
Prior	I	18.9394	0.7149	0.2181	39.0026
	II	19.1940	0.7152	0.2115	38.4703
Coarse network	III	17.9160	0.7096	0.2364	39.7605
	IV	18.1886	0.7122	0.2315	37.8431
	V	19.0473	0.7324	0.2142	33.0043
	VI	19.5710	0.7257	0.2085	32.9269
DFM	VII	17.2023	0.7188	0.2757	35.9658
	VIII	19.3467	0.7210	0.2133	33.2503
	IX	19.7791	0.7273	0.1990	33.3295
Discriminator	X	19.7978	0.7216	0.2229	38.3030

Table 3: Ablation study on different configurations of LOHC.

perimental results were shown in Table 3.

Human segmentation prior. To examine the effect of human body segmentation prior to image completion quality, we first fully exclude the segmentation information from the baseline network (Option I). Here, only the occluded image is used as input into the network to generate the completed image. The input of the global discriminators and the human image encoder in the coarse network is replaced by the complete image, and the loss function related to segmentation information is removed. Furthermore, we analyzed on the effects of unprocessed occluded prior on image completion. We input both the occluded image and the occluded segmentation map into the model, without providing any supervision or completing the segmentation map (Option II).

The results indicate that the integration of prior information on human body segmentation has enhanced the network's image completion performance to some extent. Moreover, the proposed cooperative completion strategy can efficiently utilize segmentation information, resulting in a significant improvement in the network's performance.

Coarse network. We verified the effectiveness of the coarse network and dual encoder settings by conducting experiments. We first completely remove the coarse network in the baseline, and the channel attention part used in the DRM for fusion was also removed (Option III). Next, to illustrate the effect of separately encoding the human part of the image in the coarse network, we constructed three comparative schemes to replace the encoders of the coarse network in the baseline: encoding only the image with a single large encoder (Option IV); encoding both the image and the human part of the image simultaneously with the large encoder (Option V); and using the original two encoders in the baseline but making them both encode only the image (Option VI). All three methods have the same parameters as the baseline.

The results show that the coarse completed images generated by the coarse network can significantly improve the structural authenticity of the finally completed images. Although the human image part without a background has less information than the complete image, it can enable the decoder to better analyze the shape of the person and generate a more realistic human body, especially when the human part is separately encoded.

Dynamic fusion module. We conducted experiments to separately exclude the complete dynamic fusion module (Option VII) and its channel attention (Option VIII) and spatial attention (Option IX) parts in the refinement network to assess their impact on network performance.

The experimental results indicate that directly concatenating coarse images with occluded images substantially limits the performance of refinement networks. Both channel attention and spatial attention in the dynamic fusion module significantly improve the performance of the network, which is essential for achieving higher-quality output.

Dual discriminator. To evaluate the efficacy of the global discriminators, we removed them from both the coarse and refinement networks (Option X).

The experimental results show that the global discriminator plays a crucial role in guiding the overall structure of images in both coarse and refinement networks, and its removal significantly impacts the network's performance. Moreover, both the human part image and the complete image are indispensable for the effective operation of the global discriminator.

Conclusion

In this paper, we have investigated the completion of large occluded human images. we proposed a two-stage human image completion network based on an image-prior cooperative completion strategy. Our study highlights that integrating prior completion with the image completion process can be a more effective approach for utilizing prior information to generate more realistic images. We demonstrate the importance of providing additional supervision on human body parts during training for human body image completion tasks. Achieving adequate attention to both human structure and detailed texture using a single discriminator can be challenging, but our findings suggest that this issue can be effectively addressed by employing two discriminators - one for supervising global features and the other for supervising local features. Finally, extensive experimental results indicate our method performs better than state-of-the-art methods in the human image completion task. In addition, the application of our strategy to other models of backbone structure is still feasible in theory, and human image completion based on other backbones such as the diffusion model deserves further study in the future.

Acknowledgements

The paper is supported by the National Natural Science Foundation of China under grant No.62293540, 62293542,

62276045.

References

Balakrishnan, G.; Zhao, A.; Dalca, A. V.; Durand, F.; and Guttag, J. 2018. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8340–8348.

Ballester, C.; Bertalmio, M.; Caselles, V.; Sapiro, G.; and Verdera, J. 2001. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8): 1200–1211.

Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3): 24.

Chan, T. F.; and Shen, J. 2001. Nontexture inpainting by curvature-driven diffusions. *Journal of visual communication and image representation*, 12(4): 436–449.

Criminisi, A.; Pérez, P.; and Toyama, K. 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9): 1200–1212.

Efros, A. A.; and Freeman, W. T. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 341–346.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. 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*, 932–940.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Grigorev, A.; Sevastopolsky, A.; Vakhitov, A.; and Lempitsky, V. 2019. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12135–12144.

Han, X.; Wu, Z.; Huang, W.; Scott, M. R.; and Davis, L. S. 2019. Compatible and diverse fashion image inpainting. *arXiv preprint arXiv:1902.01096*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Iizuka, S.; Simo-Serra, E.; and Ishikawa, H. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4): 1–14.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2989–2998.

Ke, L.; Tai, Y.-W.; and Tang, C.-K. 2021. Occlusion-aware video object inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14468–14478.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Kwatra, V.; Essa, I.; Bobick, A.; and Kwatra, N. 2005. Texture optimization for example-based synthesis. In *ACM SIGGRAPH 2005 Papers*, 795–802.

Lassner, C.; Pons-Moll, G.; and Gehler, P. V. 2017. A generative model of people in clothing. In *Proceedings of the IEEE international conference on computer vision*, 853–862.

Liang, X.; Xu, C.; Shen, X.; Yang, J.; Liu, S.; Tang, J.; Lin, L.; and Yan, S. 2015. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, 1386–1394.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5084–5093.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F. Z.; and Ebrahimi, M. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A.; and Zhang, R. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33: 7198–7211.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023a. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, X.; Li, R.-L.; Zhang, F.-L.; Liu, J.-C.; Wang, J.; Shamir, A.; and Hu, S.-M. 2019. Deep portrait image completion and extrapolation. *IEEE Transactions on Image Processing*, 29: 2344–2355.

Xi, T.; Sun, Y.; Yu, D.; Li, B.; Peng, N.; Zhang, G.; Zhang, X.; Wang, Z.; Chen, J.; Wang, J.; et al. 2022. UFO: unified feature optimization. In *European Conference on Computer Vision*, 472–488. Springer.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; and Luo, J. 2019. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5840–5848.

Yang, Y.; and Guo, X. 2020. Generative landmark guided face inpainting. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part I 3*, 14–26. Springer.

Ye, H.; and Xu, D. 2022. Inverted pyramid multi-task transformer for dense scene understanding. In *European Conference on Computer Vision*, 514–530. Springer.

Yi, Z.; Tang, Q.; Azizi, S.; Jang, D.; and Xu, Z. 2020. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7508–7517.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5505–5514.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4471–4480.

Zeng, Y.; Lin, Z.; Lu, H.; and Patel, V. M. 2021. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14164–14173.

Zeng, Y.; Lin, Z.; and Patel, V. M. 2022. Shape-guided object inpainting. *arXiv preprint arXiv:2204.07845*.

Zeng, Y.; Lin, Z.; Yang, J.; Zhang, J.; Shechtman, E.; and Lu, H. 2020. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, 1–17. Springer.

Zhao, S.; Cui, J.; Sheng, Y.; Dong, Y.; Liang, X.; Chang, E. I.; and Xu, Y. 2021a. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*.

Zhao, Z.; Liu, W.; Xu, Y.; Chen, X.; Luo, W.; Jin, L.; Zhu, B.; Liu, T.; Zhao, B.; and Gao, S. 2021b. Prior based human completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7951–7961.

Zheng, C.; Cham, T.-J.; Cai, J.; and Phung, D. 2022a. Bridging global context interactions for high-fidelity image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11512–11522.

Zheng, H.; Lin, Z.; Lu, J.; Cohen, S.; Shechtman, E.; Barnes, C.; Zhang, J.; Xu, N.; Amirghodsi, S.; and Luo, J. 2022b. Image inpainting with cascaded modulation GAN and object-aware training. In *European Conference on Computer Vision*, 277–296. Springer.

Zhou, Q.; Wang, S.; Wang, Y.; Huang, Z.; and Wang, X. 2021. Human de-occlusion: Invisible perception and recovery for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3691–3701.