Image Inpainting with Cascaded Modulation GAN and Object-Aware Training

Haitian Zheng[1,2], Zhe Lin[2], Jingwan Lu[2], Scott Cohen[2], Eli Shechtman[2], Connelly Barnes[2], Jianming Zhang[2], Ning Xu[2], Sohrab Amirghodsi[2], and Jiebo Luo[1]

[1] University of Rochester
[2] Adobe Research

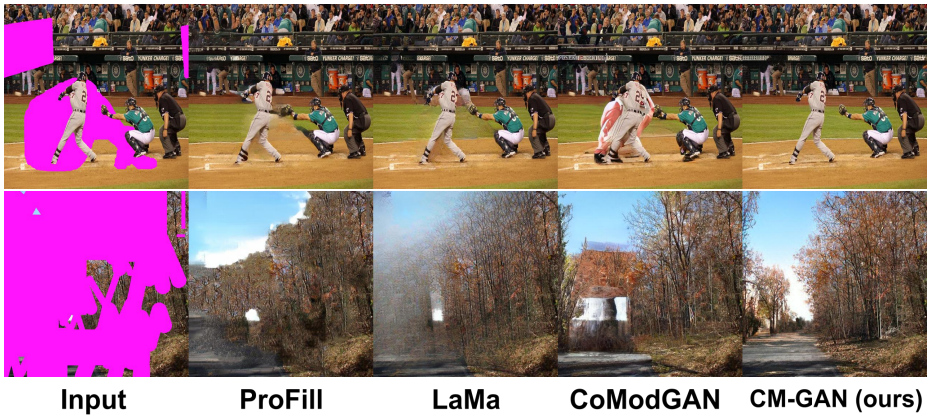

Fig. 1: Results of CM-GAN in comparison to state of the art methods: ProFill [57], LaMa [44] and CoModGAN [59]. CM-GAN generates more plausible and realistic results for the distractor removal scenario (1st row) and large holes (2nd row).

Abstract. Recent image inpainting methods have made great progress but often struggle to generate plausible image structures when dealing with large holes in complex images. This is partially due to the lack of effective network structures that can capture both the long-range dependency and high-level semantics of an image. We propose cascaded modulation GAN (CM-GAN), a new network design consisting of an encoder with Fourier convolution blocks that extract multi-scale feature representations from the input image with holes and a dual-stream decoder with a novel cascaded global-spatial modulation block at each scale level. In each decoder block, global modulation is first applied to perform coarse and semantic-aware structure synthesis, followed by spatial modulation to further adjust the feature map in a spatially adaptive fashion. In addition, we design an object-aware training scheme to prevent the network from hallucinating new objects inside holes, fulfilling the needs of object removal tasks in real-world scenarios. Extensive experiments are conducted to show that our method significantly outperforms existing methods in both quantitative and qualitative evaluation. Please refer to the project page: https://github.com/htzheng/CM-GAN-Inpainting.

Keywords: Image Inpainting, Generative Adversarial Networks

1 Introduction

Image inpainting refers to the task of completing missing regions of an image as shown in Fig. 1. It is one of the fundamental tasks in computer vision and has many practical applications, such as object removal [57,55] and manipulation [33,34], image retargeting [41,46,4], image compositing [7], and 3D photo effects [32,24].

Early inpainting methods leverage patch-based synthesis [4,25,9] or color diffusion [3,6,42,10] to fill the holes by propagating repeating textures and patterns from the visible regions. To facilitate the completion of more complex image structures, recent research efforts have shifted to adopting a data-driven scheme where deep generative networks are learned to predict visual content and appearance [54,55,57,44,59]. By training on a large corpus of images and with assist of both reconstruction and adversarial losses, generative inpainting models have shown to produce more visually appealing results on various types of input data including natural images and faces.

While existing works have shown promising results on completing simple image structures, generating complex holistic structures and image contents with high-fidelity details remains a huge challenge, espeically when the holes are large. Essentially, how to 1) *accurately propagate global context into the incomplete region* while 2) *synthesizing realistic local details that are coherent to the global clue* is the key question for image inpainting. To tackle global context propagation, existing networks leverage encoder-decoder structure [37,18,36], dilated convolution [53,55], contextual attention [54,55,52], or Fourier convolution [44] to incorporate long range feature dependency for expanding the effective receptive field [28]. Furthermore, two-stage approaches [50,54,55,31,49] and iterative hole filling [57] predict a coarse result such as a smoothed image, edge/semantic maps or partial completion to enhance the global structure. However, those models lack a mechanism to capture high-level semantics in the unmasked region and effectively propagate them into the hole to synthesize a holistic global structure. With the typical shallow bottleneck designs, the designed feature propagation layer is less aware of global semantics and more prone to generating incoherent local details potentially leading to visual artifacts.

More recently, feature modulation-based methods [20,21,36] have shown very promising results on controlling image generation with a global style code. Benefiting from a global code that captures the context of the entire image, CoModGAN [59] attempts to inject global context into the generator for filling in very large holes [59]. However, due to the lack of spatial adaptation, global modulation is sensitive to the corrupted encoding feature inside the inpainting region (shown in Fig. 3). Their results show that passing global information in this way is insufficient for synthesizing high quality global structures which may lead to severe artifacts such as the large unseen color blobs [59] or inconsistent visual appearances such as distorted structures, cf. Fig. 5.

To seek a better way to inject global context into the missing region in inpainting, we investigate a new modulation scheme by cascading global and spatial modulations. We propose **C**ascaded **M**odulation **GAN** (**CM-GAN**), a new generative network that can synthesize better holistic structure and local details, cf. Fig. 1 and Fig. 5. Different from [59] that attempts to globally modulates the partially invalid encoder feature as shown Fig. 3, our spatially-adaptive modulation scheme cascaded after a global modulation is much more effective in processing invalid features inside the hole. Although several spatially-adaptive modulation schemes [35,22] have been proposed in the past to tackle inpainting, our cascaded modulation approach is significantly different from those works in that: i) our spatial code comes from the *decoding stage* rather than the encoding stage to avoid modulating the decoder with invalid encoding feature

(cf. Fig. 3), 2) we incorporate the global code into spatial modulation for enforcing the global-local consistency, 3) we introduce a dual-branch design that decouples global and local features for structure-details separation, and 4) we design spatially-aware demodulation instead of instance or batch normalization to avoid potential 'droplet artifact" [21]. Furthermore, on the encoder side, we inject the Fast Fourier convolution [8] at each stage of the encoder network to expand the receptive field of the encoder at early stages, allowing the network to capture long-range correlations across the image.

Another design aspect to consider is how to generate synthetic masks used during inpainting training. We need to design the masks tailored for real-world inpainting use cases, such as object removal and partial object completion. Previous methods generate training data by randomly locating rectangular [37,18] or irregularly-shaped [55,59] masks. However, the masks users draw in common use cases likely have a shape of an existing object, or part of an object or other relatively simple shapes such as a scribble or a blob. Moreover, users usually expect the removed objects to be filled by background textures and structures, thus new objects are not expected to appear inside the holes. However, models trained with randomly located masks tend to have the color-bleeding effect across object boundaries and generate object-like artifact blobs inside the hole [59]. In a better attempt, a recent work [57] leverages saliency annotation to simulate the holes left by distracting objects occluded by the foreground salient objects. However, saliency annotations only capture large dominant foreground objects, thus resulting in the algorithm constructing large holes that include most of the other less salient objects. This is different from the real use cases where only a few distracting objects need to be removed.

We found that the tendency to generate spurious objects, blobs and color-bleeding effects across object boundaries can be addressed by a proper and more carefully designed object-aware training scheme. In terms of mask sampling for training, different from random mask [55,59,44,27,39] or saliency-based mask [57], we leverage an instance-level panoptic segmentation model [26] to generate object-aware masks that better simulate real distractors or clutter removal use cases. To avoid generating distorted objects or color blobs inside the hole, we filter out cases where the entire object or a large part of the object is covered by the mask. Furthermore, the panoptic segmentation provides precise object boundaries and thus prevents the trained model from leaking colors at object boundaries.

Finally, for the training losses, we propose a masked R_1 regularization specifically designed for inpainting and augment the adversarial loss with a perceptual loss extracted by segmentation model for improving robustness. The new regularization avoids penalizing the model outside the mask and thus imposes a better separation of input condition from the generated contents. Consequently, the new regularization eliminates the potential harmful impact of computing regularization on the background. Our contributions are four-fold:

- Cascaded Modulation GAN, a new inpainting network architecture formed by a masked image encoder with Fourier convolution blocks and a cascaded global-spatial modulation-based decoder.
- An object-aware mask generation scheme preventing the model from generating new objects inside holes and mimicking realistic inpainting use cases.
- A masked R_1 regularization loss to stabilize the adversarial training for the inpainting tasks.
- State-of-the-art results on the Places2 dataset for various types of masks.

2 Related works

2.1 Image Inpainting

Image inpainting has a long standing history. Traditionally, the patch-based methods [11,25,4,9] search and copy-paste patches from known region to progressively fill in the target hole. Meanwhile, the diffusion-based methods [3,6,42,10] describe and solve the color propagation inside the hole via partial differential equation. The above methods can produce high-quality stationary textures while completing simple shapes, but they lack the mechanisms to model the high-level semantics for completing new semantic structure inside the hole. Recently, inpainting methods have shifted to a data-driven scheme where deep generative models are learned to directly predict the filled in content inside the hole in an end-to-end fashion. By training deep generative models via adversarial training [12], the learned model can capture higher-level representation of images and hence can generate more visually plausible results. Specifically, Pathak *et al.* [37] first leverage an encoder-decoder network with a bottleneck layer to predict the missing structures for hole filling. Iizuka *et al.* [18] propose a two discriminator design to encourage the global and local consistency separately. To make the generator better at capturing the global context, several attempts have been made. Motivated by the structure-texture decomposition principle [2,5], two-stage networks predict an intermediate representation of image with smoothed image [54,15,52,47,57], edge [31,49], gradient [51] or segmentation map [43] for enhancing the final output. Yu *et al.* [54] design contextual attention to explicitly let the network borrow patch features at a global scale. Aiming at expanding the receptive field of the network, Iizuka *et al.* [18] Yu *et al.* [54] incorporate dilated convolutions to the generator. Likewise, Suvorov *et al.* [44] leverage Fourier convolution [8] to acquire a global receptive field. Furthermore, feature gating such as partial convolution [27] and gated convolution [55] is proposed to handle invalid features inside the hole. To enhance global prediction capacity, Zhao *et al.* [59] propose an encoder-decoder network that leverages style code modulation for global-level structure inpainting. To augment the adversarial loss and to suppress artifacts, existing works [54,55,44,57,31] often train the generator with additional reconstruction objectives such as ℓ_1, perceptual [19] or contextual [37] loss. Recently, Suvorov *et al.* [44] propose to use segmentation networks to compute perceptual loss which achieve better performance.

2.2 Feature Modulation for Image Synthesis

Originating from style transfer [16], feature modulation [17,20,21] has been widely adopted to incorporate input conditions for controlled generation [48,36,1,59,35,60,45]. Existing modulation methods usually leverage batch normalization or instance normalization to normalize the input feature. The modulation is then achieved by scaling and shifting the normalized activation according to affine parameters predicted from input conditions. Recently, Karras *et al.* [21] find that normalization would cause the "droplet artifact" as the network can create a strong activation spike to sneak signal through normalization layer. Consequently, StyleGAN2 [21] replaces the feature normalization with a proposed demodulation step [21] for better image synthesis. To modulate the input feature according to a spatially-varying feature map, spatial modulation [35,22,1,48] are proposed. Essentially, those methods leverage convolutional layers to predict 2d affine parameters for spatially-controlled modulation. However, feature normalization makes the existing approaches [1,48] less consistent with the design principle of StyleGAN2.

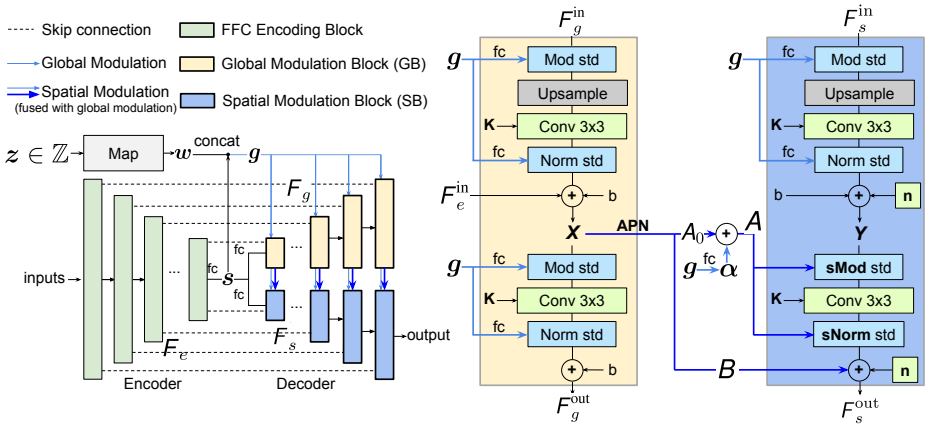

Fig. 2: **Left**: The CM-GAN architecture, which consists of an encoder with FFC blocks and a two-stream decoder with a cascade of global modulation block (GB) and subsequent spatial modulation block (SB). This cascaded modulation scheme extracts spatial style codes from the globally modulated feature map (instead of from the encoder feature map used in the previous work) to make spatial modulation more effective for inpainting. **Right**: Cascaded modulation at each scale. GB and SB take F_g^{in} and F_s^{in}, respectively, as inputs and produce the upsampled feature F_g^{out} and F_s^{out}. Specifically, we apply joint global-spatial modulation to ensure the generation consistency both at the global and local scales.

2.3 Regularization for Adversarial Training

Adversarial training is known to be challenging [29] as it is hard for the adversarial networks to reach global Nash equilibrium [14]. Consequently, various regularization are proposed to stabilize the GAN training. In particular, weight normalization [40] and spectrum normalization [30] are proposed to constrain the the Lipschitz continuity of the discriminator. Likewise, Gulrajani *et al.* [13] propose a gradient penalty to impose a K-Lipschitz constraint to the discriminator. Mescheder *et al.* [29] propose R_1 regularization to penalize the discriminator gradient on real data, which is later used by [21,59,44]. Karras *et al.* [21] propose perceptual path length regularization on the generator to ensure smoothness mappings and lazy regularization to optimize the training efficiency.

3 Method

3.1 Cascaded Modulation GAN

To better model the global context for image completion [54,55,57,56,52,59,44], we propose a novel mechanism that *cascades global code modulation with spatial code modulation* to facilitate the processing of the partially invalid feature while better injecting the global context into the spatial region. It leads to a new architecture named Cascaded Modulation GAN (CM-GAN), which can synthesize holistic structures and local details surprisingly well as shown in Fig. 1.

Network Overview. As shown in Fig. 2 (left), CM-GAN is based on an encoder branch and two parallel cascaded decoder branches to generate visual output. Specifically, it starts with an encoder that takes the partial image and the mask as inputs to produce multi-scale feature maps $F_e^{(1)}, \cdots, F_e^{(L)}$ at each scale $1 \leq i \leq L$ (L is the highest level with the smallest spatial size). Unlike most encoder-decoder methods [54,44] and to facilitate the completion of holistic structure, we extract a global style code $\boldsymbol{s}$ from the highest level feature $\boldsymbol{F}_e^{(L)}$ with a fully connected layer followed by a ℓ_2 normalization. Furthermore, an MLP-based mapping network [21] is used to generate a style code $\boldsymbol{w}$ from noise, simulating the stochasticity of image generation. The code $\boldsymbol{w}$ is joined with $\boldsymbol{s}$ to produce a global code $\boldsymbol{g} = [\boldsymbol{s}; \boldsymbol{w}]$ for the consequent decoding steps.

Global-Spatial Cascaded Modulation. To better bridge the global context at the decoding stage, we propose *global-spatial* **C***ascaded* **M***odulation* (CM). As shown in Fig. 2 (right), the decoding stage is based on two branches of Global Modulation Block (GB) and Spatial Modulation Block (SB) to respectively upsample global feature F_g and local features F_s in parallel. Different from existing approaches [55,57,47,44,59], the CM design introduces a new way to inject the global context into the hole region. At a conceptual level, it consists of a cascade of global and spatial modulations between features at each scale and naturally integrates three compensating mechanisms for global context modeling: 1) *feature upsampling* allows both GB and SB to utilize the global context from the low-resolution features generated by both of the previous blocks; 2) the *global modulation* (cyan arrows of Fig. 2) allows both GB and SB to leverage the global code $\boldsymbol{g}$ for generating better global structure; and 3) *spatial modulation* (blue arrows of Fig. 2) leverages spatial code (intermediate feature output of GB) to further inject fine-grained visual details to SB.

More specifically, as shown in Fig. 2 (right), CM at each level of the decoder consists of the paralleled GB block (yellow) and SB block (blue) bridged by spatial modulation. Such parallel blocks takes F_g^{in} and F_s^{in} as input and output F_g^{out} and F_s^{out}. In particular, GB leverages an initial upsampling layer following a convolution layer to generate the intermediate feature X and global output F_g^{out}, respectively. Both layers are modulated by the global code $\boldsymbol{g}$ [21] to capture the global context.

Due to the limited expressive power of the global code $\boldsymbol{g}$ to represent a 2-d scene, and the noisy invalid features inside the inpainting hole [55,27], the global modulation alone generates distorted features inconsistent with the context as shown in Fig. 3 and leads to visual artifacts such as large color blobs and incorrect structure as demonstrated in Fig. 7. To address this critical issue, we cascade GB with an SB to correct invalid features while further injecting spatial details. SB also takes the global code $\boldsymbol{g}$ to synthesize local details while respecting global context. Specifically, taking the spatial feature F_s^{in} as input, SB first produces an initial upsampled feature Y with an upsampling layer modulated by global code $\boldsymbol{g}$. Next, Y is jointly modulated by X and $\boldsymbol{g}$ in a spatially adaptive fashion following the *modulation-convolution-demodulation* principle [21]:

- *Global-spatial feature modulation.* A spatial tensor $A_0 = \text{APN}(X)$ is produced from feature X by a 2-layer convolutional affine parameter network (APN). Meanwhile, a global vector $\boldsymbol{\alpha} = \text{fc}(\boldsymbol{g})$ is produced from the global code $\boldsymbol{g}$ with a fully connected layer (fc) to incorporate the global context. Finally, a fused spatial tensor $A = A_0 + \boldsymbol{\alpha}$ leverages both global and spatial information extracted from $\boldsymbol{g}$ and X, respectively, to scale the intermediate feature Y with element-wise product $\odot$:

$$\bar{Y} = Y \odot A. \tag{1}$$

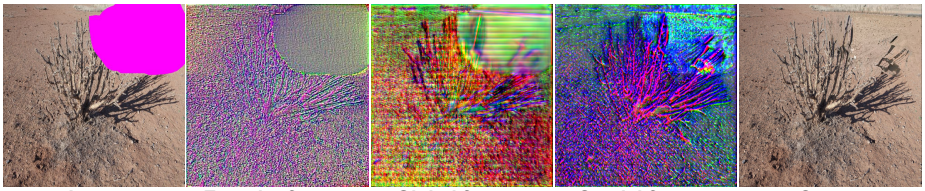

Fig. 3: Visualization of the intermediate features for inpainting. From left to right are the incomplete image, encoded feature, our globally modulated feature and spatially modulated features at the 256×256 layer and the output image.

– *Convolution.* The modulated tensor $\bar{Y}$ is then convolved with a 3×3 learnable kernel K, resulting in $\hat{Y}$

$$\hat{Y} = \bar{Y} * K. \tag{2}$$

– *Spatially-aware demodulation.* Different from existing spatial modulation methods [35,22,1], we discard instance or batch normalization to avoid the known "water droplet" artifact [21] and propose a spatially-aware demodulation step to produce normalized output $\widetilde{Y}$. Specifically, we assume that the input features Y are independent random variables with unit variance and after the modulation, the expected variance of the output does not change, i.e. $\mathbb{E}_{y \in \widetilde{Y}}[\mathrm{Var}(y)] = 1$. This assumption gives the demodulation computation:

$$\widetilde{Y} = \hat{Y} \odot D, \tag{3}$$

where $D = 1/\sqrt{K^2 \odot \mathbb{E}_{a \in A}[a^2]}$ is the demodulation coefficient. Eq. (3) is implemented with standard tensor operations as elaborated in the supplementary material.

– *Adding spatial bias and broadcast noise.* To introduce further spatial variation from feature X, the normalized feature $\widetilde{Y}$ is added to a shifting tensor $B = \mathrm{APN}(X)$ produced by another affine parameter network from feature X along with the broadcast noise $\boldsymbol{n}$ to generate the new local feature F_s^{out}:

$$F_s^{\mathrm{out}} = \widetilde{Y} + B + \boldsymbol{n}. \tag{4}$$

As shown in the 4th column of Fig. 3, the cascaded SB block helps generate fine-grained visual details and improves the consistency of feature values inside and outside the hole.

Expanding the Receptive Field at Early Stages. The fully convolutional models suffer from slow growth of the effective receptive field [28], especially at early stages of the network. For this reason, an encoder based on strided convolution usually generates invalid features inside the hole region, making the feature correction at the decoding stage more challenging. A recent work [44] shows that fast Fourier convolution (FFC) [8] can help early layers achieve large receptive fields that cover the entire image. However, the work [44] stacks FFC at the bottleneck layer and is computationally demanding. Moreover, like many other works [55], due to the shallow bottleneck layers, [44] cannot capture global semantics effectively, limiting its ability to handle large holes. We propose to replace every convolutional blocks of the CNN encoder with FFC. By adopting FFC at all scale levels, we enable the encoder to propagate features at early stages and thus address the issue of generating invalid features inside the holes, helping improve the results as shown in the ablation study in Tab. 2.

Fig. 4: Examples of our object-aware masks generated for training (right) in comparison to the real inpainting requests (left) and the masks generated by CoModGAN [59] (middle). Note that our masks are more consistent with real user requests.

3.2 Object-aware Training

The algorithm to generate masks for training is crucial. In essence, the sampled masks should be similar to the masks drawn in realistic use-cases. Moreover, the masks should avoid covering an entire object or most of any new object to discourage model from generating object-like patterns. Previous works generate mask with square-shaped masks [37,18] or use random strokes [27] or a mixture of both [55,59] for training. The oversimplified mask schemes may cause casual artifacts such as suspicious objects or color blobs.

To better support realistic object removal use cases while preventing the model from trying to synthesize new objects inside the holes, we propose an object-aware training scheme that generate more realistic masks during training as shown in Fig. 4. Specifically, we first pass the training images to PanopticFCN [26] to generate highly accurate instance-level segmentation annotations. Next, we sample a mixture of free-form holes [59] and object holes as the initial mask. Finally, we compute the overlapping ratio between a hole and each instance from an image. If the overlapping ratio is larger than a threshold, we exclude the foreground instance from the hole. Otherwise, the hole is unchanged to mimic object completion. We set the threshold to 0.5. We dilate and translate the object masks randomly to avoid overfitting. We also dilate the hole on the instance segmentation boundary to avoid leaking background pixels near the hole into the inpainting region.

3.3 Training Objective and Masked-R_1 Regularization

Our model is trained with a combination of adversarial loss [59] and segmentation-based perceptual loss [44]. The experiments show that our method can also achieve good results when purely using the adversarial loss, but adding the perceptual loss can further improve the performance. In addition, we propose a masked-R_1 regularization tailored to stabilize the adversarial training for the inpainting task. Different from [59,44] that naively apply R_1 regularization [29], we leverage the mask m to avoid computing the gradient penalty outside the mask, specifically:

$$\bar{R}_1 = \frac{\gamma}{2}\mathbb{E}_{p_{\text{data}}}[\|m \odot \nabla D(x)\|^2], \tag{5}$$

where m is the mask indicating the hole region and γ is a balancing weight. The new loss eliminates the potential harmful impact of computing gradient on real pixels, and therefore stabilizes the training.

Table 1: Quantitative evaluation of inpainting on the Places evaluation set. We report FID [14], LPIPS [58], U-IDS [59] and P-IDS [59] scores.

Methods	FID↓	LPIPS↓	U-IDS↑	P-IDS↑	Methods	FID↓	LPIPS↓	U-IDS↑	P-IDS↑
CM-GAN	**1.628**	**0.189**	**37.42**	**20.96**	DS [38]	16.003	0.399	10.72	0.47
CoModGAN[59]	3.724	0.229	32.38	14.68	EC [31]	12.086	0.414	9.21	0.28
Lama [44]	3.864	0.195	29.57	10.08	ICT [47]	16.405	0.424	8.12	0.25
ProFill [57]	7.700	0.230	21.19	3.87	HiFill [52]	37.484	0.336	9.86	0.96
CRFill [56]	9.657	0.233	22.90	5.53	SF [39]	28.252	0.489	6.05	0.13
DeepFillv2 [55]	13.597	0.371	14.23	1.67	MEDFE [15]	35.454	0.445	7.10	0.35

4 Experiments

Implementation Details. We conduct the image inpainting experiment at resolution 512×512 on the Places2 dataset [61]. Our model is trained with Adam optimizer [23]. The learning rate and batch size are set to 0.001 and 32, respectively. Our network takes the resized image as input, so that the model can predict the global structure of an image. We apply flip augmentation to increase the training samples.

Evaluation Metrics. We report the numerical metrics on the validation set of Places2 which contains 36.5k images. For the numerical evaluation, we compute *Frchet Inception Distance* (FID) [14] and *Perceptual Image Patch Similarity Distance* (LPIPS) [58]. We also adopt the *Paired/Unpaired Inception Discriminative Score* (P-IDS/U-IDS) [59] for evaluation.

4.1 Comparisons to Existing Methods

We set channel numbers of our network to have a similar model capacity as CoModGAN and LaMa as shown in Tab. 4.

Quantitative Evaluation. Tab. 1 presents the comparison of our method against a number of recent methods using our masks. Results show that our method significantly outperforms all other methods in terms of FID, LPIPS, U-IDS and P-IDS. We notice that with the assist of perceptual loss, LaMa [44] and our CM-GAN achieve significantly better LPIPS score than CoModGAN and other methods, attributing to additional semantic guidance provided of the pre-trained perceptual model. Compared to LaMa, our CM-GAN reduces FID by over 50% from 3.864 to 1.628, which can be explained by the typically blurry results of LaMa versus ours which tend to be sharper.

We evaluate generalization of CM-GAN to other types of masks including the wide mask [44] and the mask of CoModGAN [59]. We also fine-tune CM-GAN with masks of [44] and [59] (denoted by CM-GAN†) and report the results. As shown in Tab. 3, our models with and without fine-tuning achieve clear performance gain and demonstrate its generalization ability. Notably, CM-GAN trained on our object-aware masks outperforms CoModGAN on the CoModGAN mask, confirming the better generation capacity of CM-GAN. The strong capacity of CM-GAN brings further performance gain after fine-tuning.

Qualitative Evaluation. Fig. 5, Fig. 6 and Fig. 8 presents visual comparisons of our method with state-of-the-art methods on our synthesized masks and other types of mask introduced by [44] and [59], respectively. ProFill [57] generates incoherent global structures such as the smoothed building and tends to bleed color to the background

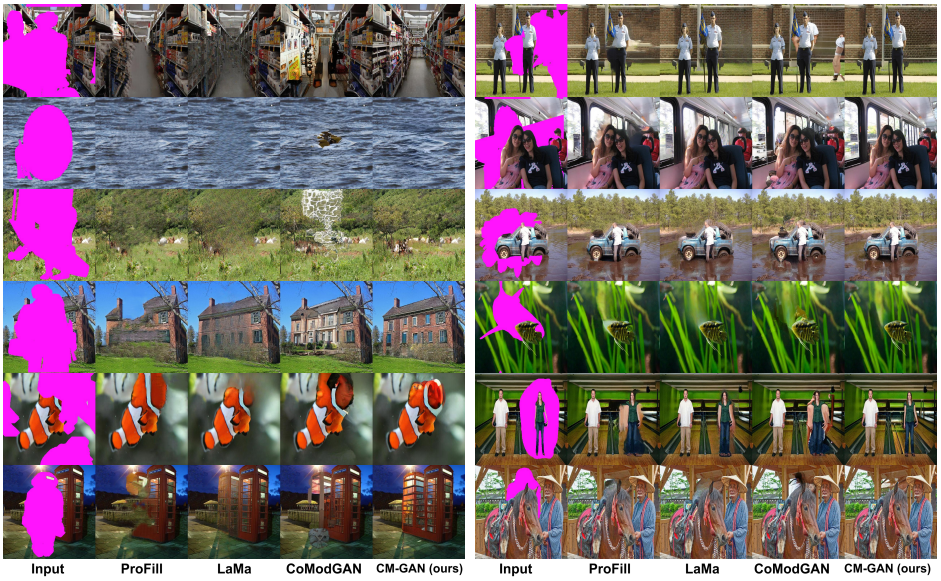

Fig. 5: Visual comparisons on Places2 with our synthesized masks including large random masks (left) and masks for the distractor removal scenario (right). We show the input images and the results of ProFill [57], Lama [44], CoModGAN [59] and CM-GAN (ours). Best viewed by zoom-in on screen.

for the object-removal case. CoModGAN [59] produces structural artifacts and color blobs. LaMa [44] is superior on repeating structures, but tends to generate blurry results on large holes, especially on nature scenes. In contrast, our method produces more coherent semantic structures and hallucinates cleaner textures for various types of scenarios.

4.2 Ablation Study

We perform a set of ablation experiments to show the importance of each component of our model. All ablated models are trained and evaluated on the Places2 dataset. Results of the ablations are shown in Tab. 2. Below we describe the ablations from the following aspects:

Masked-R_1 regularization We start from a `baseline` with a simple encoder-decoder structure based on global-vector modulation [59] and skip connection. We compare the baseline trained with R_1 regularization with the model trained with masked R_1 regularization regularization (`baseline+m`R_1). From the result, the masked R_1 regularization improves the numerical metrics as the designed loss avoids computing gradient at the fixed input region.

Cascaded Modulation We next evaluate the cascaded modulation design on top of the baseline network and mR_1 loss. Specifically, we evaluate a baseline model with spatial modulation instead of the global code modulation, i.e. `baseline + s`. The performance improvement verify the effectiveness of the spatial adaptation introduced by spatial modulation. Next, we cascade global and spatial modulation on the baseline to get the

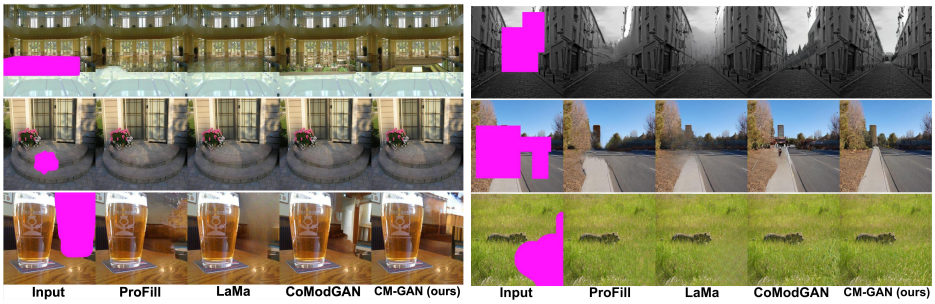

Fig. 6: Visual comparisons on Places2 with the masks proposed by Lama [44].

Fig. 7: Compared to a global modulation baseline, CM significantly improves the coherence of the synthesized color, textures, global structures and objects.

main CM model `baseline+CM (g-s) ours` which improves all numerical metrics. To better understand CM, we visually compare `baseline` with `baseline+CM (g-s) ours` in Fig. 7 and find that CM significantly improves the synthesized color, texture and global semantic and corrects the color blob artifact [59], which confirms the effectiveness of CM on correcting the incoherent feature in a global-semantic aware fashion.

Choices of Second-stage Modulation We evaluate other variant of the second-stage modulation choices: i) we replace our StyleGAN2-compatible spatial with skip connection, resulting in a model that cascades global modulation twice, i.e. `baseline+CM(g-g)`, ii) we test the CM baseline with the spatial modulation of [22], i.e. `baseline+CM(g-[22])` and iii) we drop the demodulation step (Eq. (3)) from the spatial modulation step, resulting a model with a plain spatial modulation operation, i.e. `Baseline + CM (g-s) plain`. From the results, our spatial modulation outperforms the global modulation version as we modulate feature using both global and spatial code. We found [22] does not improve CM as the instance normalization of [22] is designed for StyleGAN and is less compatible with our baseline. Furthermore, demodulation seems crucial to the model as it regularizes the intermediate feature activation. Finally, for the same reason as [22], we found SPADE [35] not compatible with our baseline due to the use of batch normalization.

Perceptual Loss The perceptual loss (perc.) provides additional semantic supervision to the network and can significantly improve the FID metrics. However, it slightly

decreases the discriminative U/P-IDS scores as perceptual loss may lead to certain visual patterns that are imperceptible to human.

Fast Fourier Convolutions Encoder The Fast Fourier Convolution (FFC) encoder further brings significant performance gains on top of the cascaded modulation and perceptual loss as shown in the table, which validates the importance of more effective encoder with wider receptive fields in early encoding stages.

Object-aware Training To study the effect of object-aware training (OT), we retrain LaMa [44] and CoModGAN [59] on our object-aware masks. The results show that object-aware training improves the performance of both of these models consistently. However, our full model still outperforms these retrained models significantly. Notably, our model reduces FID of the retrained CoModGAN (with OT) by 40% from 2.599 of to 1.628.

Table 2: Ablation study of our model design (architecture, loss, training scheme) including masked-R_1 loss (mR_1), cascaded modulation option (CM), Fourier convolution (FFC), perceptual loss (perc.) and object-aware training (OT). We report FID [14], LPIPS [58], U-IDS [59] and P-IDS [59] scores.

Ablations	Methods	FID↓	LPIPS↓	U-IDS↑	P-IDS↑
Masked-R_1 (mR_1)	Baseline	2.530	0.221	**36.59**	21.10
	Baseline + mR_1	**2.475**	**0.221**	36.58	**21.55**
Cascaded Modulation (CM)	Baseline + s	2.398	0.218	36.72	21.94
	Baseline + CM (g-g)	2.247	0.226	37.12	22.70
	Baseline + CM (g-[22])	2.915	**0.221**	35.72	20.44
	Baseline + CM (g-s) plain	2.392	0.224	37.23	22.67
	Baseline + CM (g-s)	**2.187**	0.225	**37.76**	**23.86**
FFC and perc.	Baseline + CM + perc.	1.730	0.195	36.12	19.73
	Baseline + CM + perc. + FFC	**1.628**	**0.189**	**37.42**	**20.96**
Object-aware Training (OT)	CoModGAN [59]	3.724	0.229	32.38	14.68
	CoModGAN [59] + OT	2.599	0.222	35.45	20.08
	Lama [44]	3.864	0.195	29.57	10.08
	Lama [44] + OT	2.884	0.192	31.32	15.10
	CM-GAN (full)	**1.628**	**0.189**	**37.42**	**20.96**

Table 3: Generalization evaluation on other types of mask including wide masks [44] and CoMoGAN masks [59]. CMGAN† are models fine-tuned on the two types of mask.

Wide masks [44]					**CoModGAN masks [59]**				
Methods	FID↓	LPIPS↓	U-IDS ↑	P-IDS ↑	Methods	FID↓	LPIPS↓	U-IDS ↑	P-IDS ↑
CM-GAN	1.521	0.129	39.24	23.24	CM-GAN	6.811	0.313	26.13	10.84
CM-GAN†	**1.329**	0.126	**40.20**	**25.59**	**CM-GAN†**	**5.863**	**0.310**	**27.92**	**12.45**
LaMa [44]	1.838	**0.123**	35.00	15.12	CoModGAN[59]	7.790	0.344	24.87	10.47
CoModGAN[59]	1.964	0.140	37.69	21.42	LaMa [44]	12.442	0.316	18.71	4.36
ProFill [57]	3.333	0.142	29.12	8.39	ProFill [57]	20.314	0.352	11.21	1.23

Table 4: The inference complexities. Our model has a similar number of parameters and FLOPs as other recent models.

Models	#Params of $\mathcal{G}$	#Params of $\mathcal{D}$	FLOPs
CoModGAN [59]	79.79M	28.98M	345.54G
LaMa [44]	51.25M	9.258M	395.30G
CM-GAN (ours)	75.28M	28.98M	373.60G

4.3 User Study

We conduct a user study to better evaluate the visual quality of our method. Specifically, we generate samples for evaluation using the Places2 evalation set and three types of masks: our object-aware mask, wide mask [44] and mask from real user request. The former data class contains 30 samples while the latter contains 13 real inpainting requests online following [56]. Each input image with the region to be removed and the results of different methods are presented to online users who are asked to select the best result. Finally, we collect votes from all users. Results in Tab. 5 shows that our method receives the majority of votes on both the synthetic data and realistic object removal requests.

Table 5: The user study. For each mask type, we show the number of votes and percentages for different methods (ProFill, LaMa, CoModGAN, and ours)

Masks	ProFill [57]	Lama [44]	CoModGAN [59]	CM-GAN
Our mask	20 (5%)	68 (17%)	83 (20%)	**234 (58%)**
Wide mask	45 (10%)	107 (25%)	94 (22%)	**186 (43%)**
User mask	15 (5%)	101 (35%)	55 (19%)	**120 (41%)**

5 Conclusion

In this paper, we present a new approach tailored to real-world image inpainting. Our method is based on a new modulation block that cascades global modulation with spatial modulation for better pass global context into the hole region. We further propose a training scheme based on object-aware mask sampling to improve generalization to real use cases. Finally, we propose specifically designed masked R_1 regularization to stabilize the adversarial training of image inpainting networks. Our method achieves the new state-of-the-art performance on the Places2 dataset and better visual quality.

Currently, our model is still limited in synthesizing large objects like humans or animals. One possible solution is to train a specialist inpainting model for specific types of objects. Another direction is to leverage depth and semantic segmentation for more precise structure-aware inpainting.

Acknowledgement

We would like to thank Qing Liu for the efforts and help on the demo interface and other experiments.

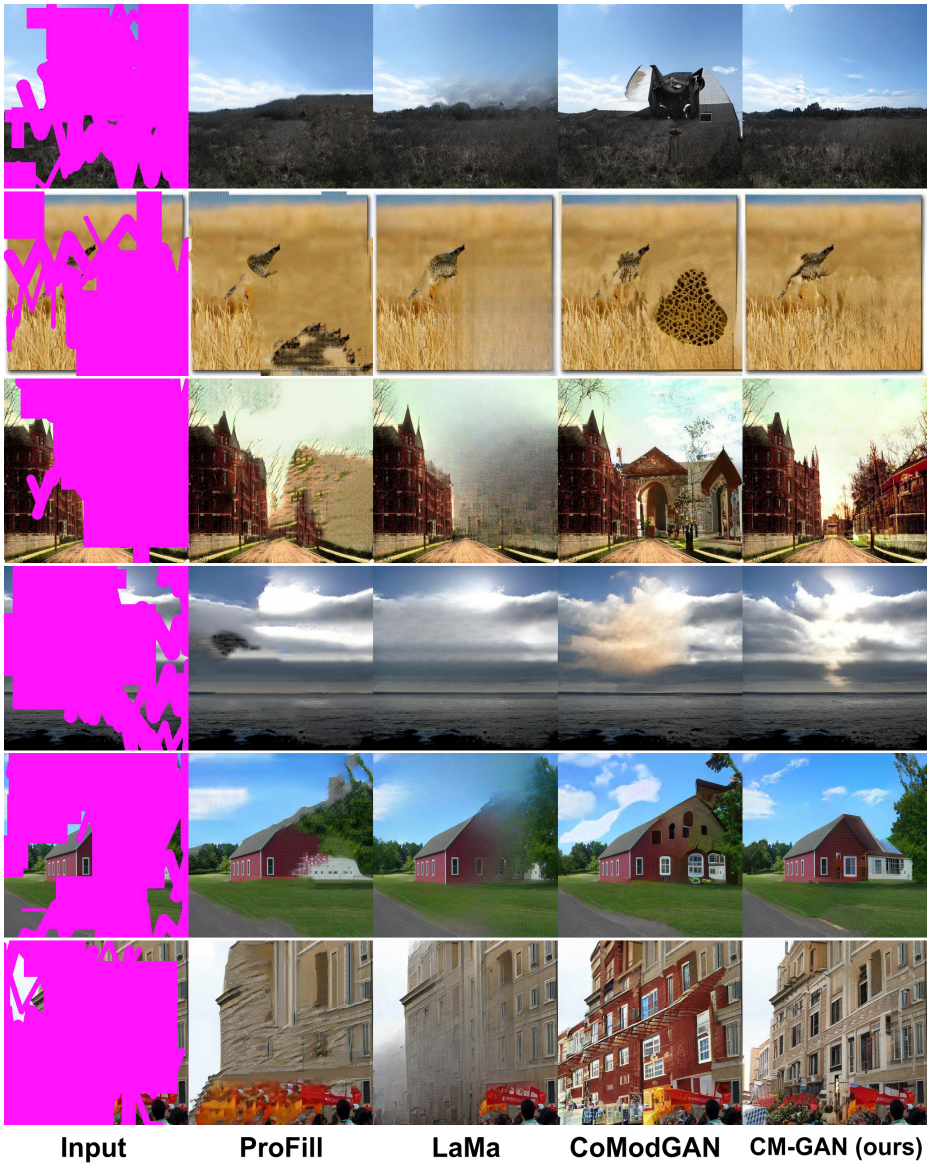

Fig. 8: **Visual comparison on the mask of CoModGAN [59]. Best viewed by zoom-in on screen.**

References

1. AlBahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., Huang, J.B.: Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. ACM Transactions on Graphics (2021) 4, 7
2. Aujol, J.F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decompositionmodeling, algorithms, and parameter selection. International journal of computer vision **67**(1), 111–136 (2006) 4
3. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE transactions on image processing **10**(8), 1200–1211 (2001) 2, 4
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009) 2, 4
5. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. IEEE transactions on image processing **12**(8), 882–889 (2003) 4
6. Chan, T.F., Shen, J.: Nontexture inpainting by curvature-driven diffusions. Journal of visual communication and image representation **12**(4), 436–449 (2001) 2, 4
7. Chen, B.C., Kae, A.: Toward realistic image compositing with adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8415–8424 (2019) 2
8. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33** (2020) 3, 4, 7
9. Cho, T.S., Butman, M., Avidan, S., Freeman, W.T.: The patch transform and its applications to image editing. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008) 2, 4
10. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing **13**(9), 1200–1212 (2004) 2, 4
11. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 341–346. ACM (2001) 4
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014) 4
13. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017) 5
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017) 5, 9, 12
15. Hongyu Liu, Bin Jiang, Y.S.W.H., Chao, Y.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: Proceedings of the European Conference on Computer Vision (2020) 4, 9
16. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017) 4
17. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018) 4

18. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 1–14 (2017) 2, 3, 4, 8
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 4
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019) 2, 4
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. CVPR (2020) 2, 3, 4, 5, 6, 7
22. Kim, H., Choi, Y., Kim, J., Yoo, S., Uh, Y.: Exploiting spatial dimensions of latent in gan for real-time image editing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021) 2, 4, 7, 11, 12
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
24. Kopf, J., Matzen, K., Alsisan, S., Quigley, O., Ge, F., Chong, Y., Patterson, J., Frahm, J.M., Wu, S., Yu, M., Zhang, P., He, Z., Vajda, P., Saraf, A., Cohen, M.: One shot 3d photography **39**(4) (2020) 2
25. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. In: ACM SIGGRAPH 2005 Papers, pp. 795–802 (2005) 2, 4
26. Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., Jia, J.: Fully convolutional networks for panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 214–223 (2021) 3, 8
27. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018) 3, 4, 6, 8
28. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 4905–4913 (2016) 2, 7
29. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018) 5, 8
30. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018) 5
31. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019) 2, 4, 9
32. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Transactions on Graphics **38**(6), 184:1–184:15 (2019) 2
33. Ntavelis, E., Romero, A., Kastanis, I., Gool, L.V., Timofte, R.: Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In: European Conference on Computer Vision. pp. 394–411. Springer (2020) 2
34. Oh, B.M., Chen, M., Dorsey, J., Durand, F.: Image-based modeling and photo editing. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 433–442 (2001) 2
35. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 2, 4, 7, 11

36. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R.: Swapping autoencoder for deep image manipulation. arXiv preprint arXiv:2007.00653 (2020) 2, 4
37. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016) 2, 3, 4, 8
38. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10775–10784 (2021) 9
39. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: Structureflow: Image inpainting via structure-aware appearance flow. In: IEEE International Conference on Computer Vision (ICCV) (2019) 3, 9
40. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems **29**, 901–909 (2016) 5
41. Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B.: Automatic image retargeting. In: Proceedings of the 4th international conference on Mobile and ubiquitous multimedia. pp. 59–68 (2005) 2
42. Shen, J., Chan, T.F.: Mathematical models for local nontexture inpaintings. SIAM Journal on Applied Mathematics **62**(3), 1019–1043 (2002) 2, 4
43. Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.C.J.: Spg-net: Segmentation prediction and guidance network for image inpainting. arXiv preprint arXiv:1805.03356 (2018) 4
44. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
45. Tan, Z., Chen, D., Chu, Q., Chai, M., Liao, J., He, M., Yuan, L., Hua, G., Yu, N.: Semantic image synthesis via efficient class-adaptive normalization. arXiv preprint arXiv:2012.04644 (2020) 4
46. Vaquero, D., Turk, M., Pulli, K., Tico, M., Gelfand, N.: A survey of image retargeting techniques. In: Applications of Digital Image Processing XXXIII. vol. 7798, pp. 328–342. SPIE (2010) 2
47. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. arXiv preprint arXiv:2103.14031 (2021) 4, 6, 9
48. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 606–615 (2018) 4
49. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5840–5848 (2019) 2, 4
50. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4076–4084 (2017) 2
51. Yang, J., Qi, Z., Shi, Y.: Learning to incorporate structure knowledge for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12605–12612 (2020) 4
52. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition. pp. 7508–7517 (2020) 2, 4, 5, 9
53. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015) 2
54. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018) 2, 4, 5, 6
55. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480 (2019) 2, 3, 4, 5, 6, 7, 8, 9
56. Zeng, Y., Lin, Z., Lu, H., Patel, V.M.: Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision (2021) 5, 9, 13
57. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. arXiv preprint arXiv:2005.11742 (2020) 1, 2, 3, 4, 5, 6, 9, 10, 12, 13
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018) 9, 12
59. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021) 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14
60. Zheng, H., Liao, H., Chen, L., Xiong, W., Chen, T., Luo, J.: Example-guided image synthesis using masked spatial-channel attention and self-supervision. In: European Conference on Computer Vision. pp. 422–439. Springer (2020) 4
61. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 9