Learning Prior Feature and Attention Enhanced Image Inpainting

Chenjie Cao*, Qiaole Dong*, and Yanwei Fu†

School of Data Science, Fudan University
{20110980001,qldong18,yanweifu}@fudan.edu.cn

Abstract. Many recent inpainting works have achieved impressive results by leveraging Deep Neural Networks (DNNs) to model various prior information for image restoration. Unfortunately, the performance of these methods is largely limited by the representation ability of vanilla Convolutional Neural Networks (CNNs) backbones. On the other hand, Vision Transformers (ViT) with self-supervised pre-training have shown great potential for many visual recognition and object detection tasks. A natural question is whether the inpainting task can be greatly benefited from the ViT backbone? However, it is nontrivial to directly replace the new backbones in inpainting networks, as the inpainting is an inverse problem fundamentally different from the recognition tasks. To this end, this paper incorporates the pre-training based Masked AutoEncoder (MAE) into the inpainting model, which enjoys richer informative priors to enhance the inpainting process. Moreover, we propose to use attention priors from MAE to make the inpainting model learn more long-distance dependencies between masked and unmasked regions. Sufficient ablations have been discussed about the inpainting and the self-supervised pre-training models in this paper. Besides, experiments on both Places2 and FFHQ demonstrate the effectiveness of our proposed model. Codes and pre-trained models are released in https://github.com/ewrfcas/MAE-FAR.

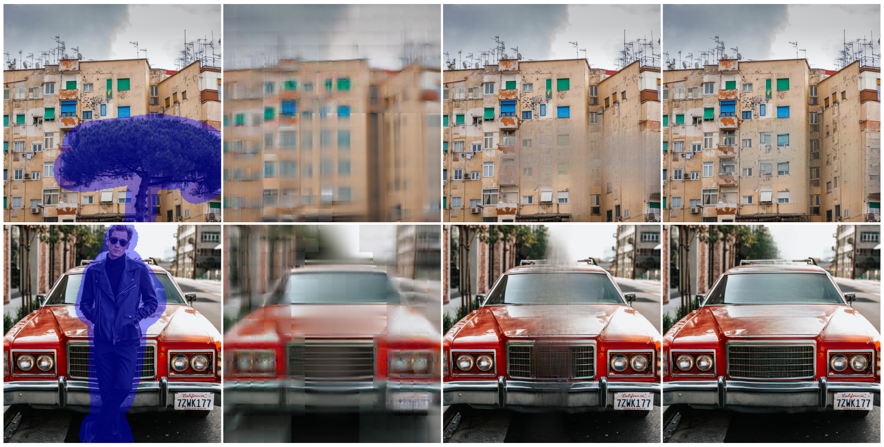
Keywords: Image Inpainting, Attention, Vision Transformer

1 Introduction

Image inpainting aims to fill missing regions of images with semantically consistent and harmoniously textured contents. It has a wide range of practical applications, including object removal [12], image editing [22], and so on. Conventional inpainting algorithms [3,27,34,17,7] rely on visual low-level assumptions and structures to search similar patches for the reconstruction. But they fail to tackle complicated inpainting situations with the limited feature representation.

Deep Neural Networks (DNNs) have recently been exploited by researchers to achieve prominent improvements in image inpainting [30,28,5,36,38,51], which thanks to the great capabilities of Convolutional Neural Networks (CNNs) [25],

∗ Equal contributions. † Corresponding authors.



(a) Masked image (b) MAE (c) LaMa (d) Ours

Fig. 1. The comparison of 1024×1024 high-resolution image inpainting. From left to right are masked images, results of pre-trained Masked AutoEncoder (MAE) [18], results of LaMa [36], and results from our method.

Generative Adversarial Networks (GAN) [13], and attention-based Transformers [37]. However, repairing images corrupted by arbitrary masks with reasonable results is still challenging, especially in high-resolution cases. Because the inpainting model needs to understand the semantic information from masked images, which demands data-driven priors and sufficient model capacities. Furthermore, the dilemmas shown below should be solved.

(i) *Limited capacities for modeling good priors.* Many pioneering works have tried to introduce prior information to inpainting models. Some works [5,30,35,46,38,47] propose multi-stage models, which repair various auxiliary information and corrupted images sequentially to enhance the image inpainting. These methods learn priors in specific fields, such as structures [30,5] or semantics [35] with good visual interpretability rather than features with more informative priors. Other methods [28,42] leverage auxiliary losses to introduce additional prior information without extending sufficient model capacities. Complex loss functions cause sophisticated hyper-parameter tuning and a more difficult inpainting training. Besides, some transformer based inpainting methods [38,47] heavily depend on low-resolution (LR) images generated by the time-consuming iterative sampling, and then upsample them with CNNs. Lahiri *et al.* [26] learn global latent priors with GAN, which can only solve simple scenes with single-object. Our method incorporates effective prior features from the transformer based representation learning [18] to enhance the inpainting, which make our method achieve superior results without overfitting the transformer results.

(ii) *Informative priors for high-resolution cases.* High-resolution (HR) image inpainting enjoys more practical implications with advanced electronic products and high-quality images in real-world. Some researches devote to facilitating

HR image inpainting with larger receptive fields [36,48], attention transfer for the high-frequency residual [43], and two-stage upsampling [49]. Unfortunately, these methods still tend to copy meaningless existing textures rather than really *understand* semantics of HR masked images without directly training with costly HR data. Our method leverages the continuous positional encoding to upsample the prior features for superior inpainting performance in HR images.
(iii) *Missing discussions about representation learning for inpainting.* Recently, self-supervised pre-training language models [32,33,4,10] have achieved great success in Natural Language Processing (NLP) fields. Such *masking and predicting* idea and transformer based architectures have been also well explored in vision tasks [18,41,2,53]. But these vision transformers only consider representation learning for classification tasks. To the best of our knowledge, no one has explored the application of self-supervised pre-training vision models to generative tasks, let alone image inpainting. We present comprehensive discussions about the pre-training based representation learning for image inpainting in this paper.

To address these dilemmas, we propose to guide the image inpainting with an efficient Masked AutoEncoder (MAE) pre-training model [18], which is called as prior Feature and Attention enhanced Restoration (FAR). Specifically, an MAE is firstly pre-trained with the masked visual prediction task. We replace some random masks with large and contiguous masks to make the MAE more suitable for the downstream task. Then, features from the MAE decoder are added to the inpainting CNN for the prior guided image inpainting. Moreover, we find that attention relations of MAE among masked and unmasked regions are compatible with CNN inpainting learning. So group convolutions are used to aggregate CNN features with attention scores from MAE, which can improve the inpainting performance a lot. Furthermore, our model can be effectively extended to HR inpainting with a little finetuning of bilinear resized MAE features and the Cartesian spatial grid. Besides, we discuss some pre-training and finetuning tricks to better utilize MAE for superior inpainting performance.

We highlight our contributions as follows. (1) We propose to learn image priors from pre-trained MAE features, which contain informative high-level knowledge and strengthen the inpainting model. (2) We propose to aggregate the inpainting CNN feature with attention scores from MAE to improve the performance. (3) Several pre-training and finetuning tricks are exploited to make our FAR learn better prior features and attentions from MAE. (4) Our method can be simply extended to HR cases and achieve state-of-the-art results. Extensive experiments on Places2 [52] and FFHQ [23] reveal that our proposed model performs better than other competitors.

2 Related work

Image Inpainting with Auxiliaries. Inpainting with auxiliaries has demonstrated success in many previous works. Various priors have been leveraged to enhance the inpainting, such as edges [30,15], lines [5], gradients [42], segmentations [35,28], low-resolution images [38,47], and even latent priors [26]. Specifically, these methods can be categorized into two types. One is to employ the approach of first correcting auxiliary information and then guiding image inpainting with multi-stage models [30,26,35,46,5,38,47]. Since these methods enjoy superior performance and good interpretability, priors leveraged by them are not comprehensive enough to handle the image inpainting properly. Other methods [28,42] supervise the inpainting model with auxiliary information directly for introducing more positive priors. But capacities of these models are still stuck to tackle tough corrupted cases. Although work [15] combines both advantages mentioned above with the dual structure and texture learning, such low-level features are still insufficient to achieve results with rich semantics.

High-resolution Image Inpainting. HR image inpainting with large mask areas is still challenging. In [49], Yu *et al.* propose an iterative inpainting method with a feedback mechanism to progressively fill holes. They further learn an upsampling network to handle HR inpainting results based on LR ones. Yi *et al.* [43] design a contextual residual aggregation mechanism for the restoration of high-frequency residuals, which are added to LR predictions. Besides, various dilated convolutions are used in [48] to enlarge receptive fields for HR inpainting. Furthermore, Suvorov *et al.* [36] leverage Fast Fourier Convolution (FFC) to learn a global receptive field in the frequency and achieve amazing HR inpainting results with periodic textures. However, these methods still suffer from copying meaningless textures without really 'understanding' semantics in HR images. In contrast to previous methods, we transfer prior features from masked autoencoders to HR cases with a continuous positional encoding, which greatly improves the quality of HR inpainting results with meaningful semantic priors.

Masked Visual Prediction. Masked Visual Prediction (MVP) is a self-supervised task for representation learning by masking and predicting image patches. This work is originated in the masked language model [9] of NLP. The Vision Transformer (ViT) [11] has studied self-supervised pre-training by masking patches. BEiT [2] and iBOT [53] learn MVP on high-level discrete tokens and self-distillation respectively. Moreover, MAE [18] proposes an efficient transformer-based masked autoencoder for visual representation learning. MaskFeat [41] further studies MVP, and proposes to use HOG features [8] to get excellent results efficiently. These MVP pre-training models can be finetuned to achieve excellent classification results. However, few discussions about MVP are explored for image generation. To the best of our knowledge, our work firstly studies MVP-based pre-training for image inpainting.

3 Method

Overview. The overall pipeline of our FAR is shown in Fig. 2. For the given masked image $\mathbf{I}_m$, we resize it into 256×256 and further enlarge the mask to patch-wise of 16×16. Thus we can get the masked image $\mathbf{I}'_m \in \mathbb{R}^{256\times256}$. Then the MAE is applied to encode the prior features as $\mathbf{F}_p = \mathrm{MAE}(\mathbf{I}_p^{(i)}), i \in \{1, 2, ..., N\}$, where $\mathbf{I}_p^{(i)}$ indicate total N unmasked patches from $\mathbf{I}'_m$ (Sec. 3.1). Prior features are resized to 1/8 of the original image size and concatenated with the Cartesian

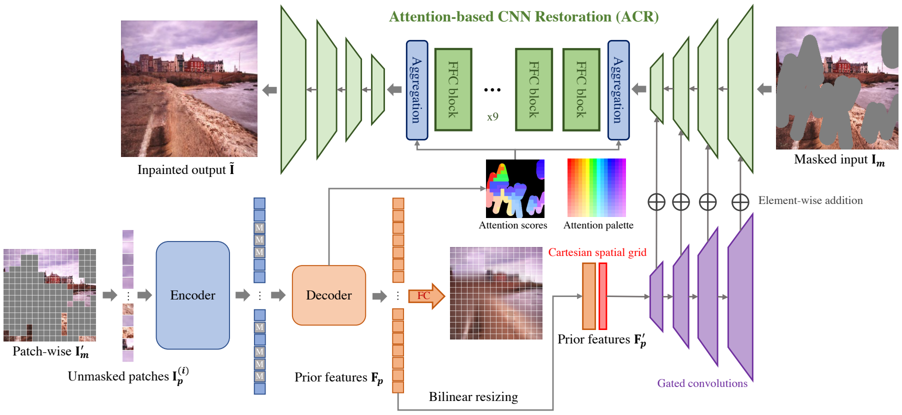


Fig. 2. The overview of our proposed FAR.

spatial grid as $\mathbf{F}'_p$. After that, $\mathbf{F}'_p$ is encoded by gated convolutions and added to the encoder of Attention-based CNN Restoration (ACR), which is used to restore the original masked image $\mathbf{I}_m$ (Sec. 3.2). Moreover, we leverage mean attention scores from the MAE decoder to aggregate unmasked features to masked regions in ACR and achieve the final inpainted image $\tilde{\mathbf{I}}$ (Sec. 3.2).

In this section, to better discuss the influence of pre-trained MAE on the inpainting model, we provide ablation studies on the subset of Places2 [52] with 5 scenes (about 25,000 training images, and 500 validation images, detailed in the Appendix). All methods are trained with 150k steps in 256×256. Although our MAE is pre-trained on the total Places2 training set, ablations among all MAE enhanced methods are still fair and meaningful. For the 512×512 ablations, we finetune the 256×256 model trained on the whole places2 for 150k steps with the dynamic resizing (Sec. 4) and test them on 1,000 validation images.

3.1 Masked Autoencoder for Inpainting

Training Settings. We use ViT-Base [11] as the backbone of MAE, which contains 12 encoder layers and 8 decoder layers. Although He *et al.* [18] have released pre-trained MAEs based on ImageNet-1K with random masks, there are still some domain gaps for the inpainting. Instead, we pre-train the MAE on the whole 1.8 million Places2 [52] and 68,000 FFHQ [23] for scene and face inpainting respectively. Validations of both datasets are excepted from the training set for a fair comparison. Moreover, the random mask used in the standard MAE is not amenable to inpainting. Although the masking ratio is high (75%), such noisy masks are easier to be restored by DNNs compared with continuous and large masks with even lower masking ratios [31]. Therefore, we blend continuous masks with 10% to 50% masking ratios and random masks, while the total masking rate remains at 75% as shown in Fig. 3. Specifically, both irregular and segmentation masks [5] are considered in continuous masks. Then we downsample continuous masks to 16×16 patch-wise forms with *max pooling* to ensure all masked patches are set to 1, while unmasked ones are 0, which enlarges masked regions. The mixed masking strategy can improve the learning of prior attention as shown in Tab. 4 of Sec. 3.2. Other training details are discussed in Sec. 4.

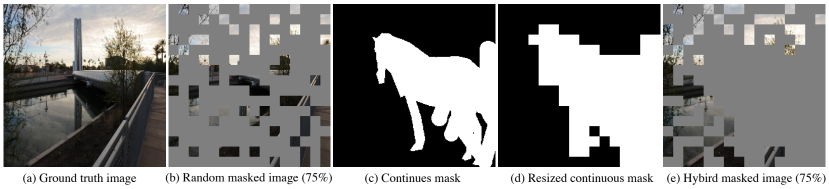

(a) Ground truth image (b) Random masked image (75%) (c) Continues mask (d) Resized continuous mask (e) Hybird masked image (75%)

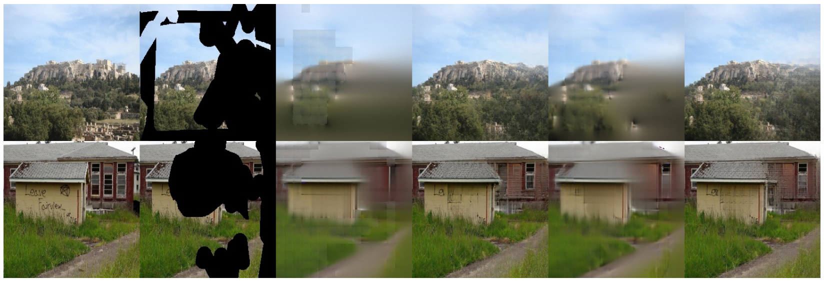
Fig. 3. The illustration for masking strategies. In (c), the continuous mask is combined with an irregular mask and a segmentation mask. We upsample the enlarged continues mask (16×16) in (d) for a better visualization.

Prior Features from MAE. In MAE [18], only 25% unmasked patches are applied into the encoder, while learnable masked tokens are shared in the decoder for the reconstruction. These masked tokens will be used to predict pixel values in masked regions. This trick makes the encoder enjoy much lower memory cost, and it can be pre-trained with more capacities for better performance in classification tasks. However, during the inpainting, features from the MAE encoder are insufficient. Because we should achieve masked features encoded by masked tokens with stacked attention modules. Therefore, features from decoder layers are more compatible with the inpainting task. To ensure good decoder learning, we chose to use balanced encoder-decoder ViT-Base architecture, which contains 12 encoder layers and 8 decoder layers rather than further enlarging the encoder capacity. In this work, we choose to use features from the last layer of the MAE decoder before the pixel linear projection as the prior features. Since the predicted images are blurry, which may contain limited information, especially for the HR inpainting. Thus an interesting future work would be exploring features from different transformer layers for inpainting.

Finetuning for Partially Masked Patches. Since the input of MAE is patch-wise tokens in 16×16, we have to enlarge some partially masked patches as shown in Fig. 3(c)(d), which may lose information. Intuitively, we further finetune MAE for those partially masked patches with 50 epochs in Places2. Specifically, masked embeddings of these partially masked patches are re-encoded by a new initialized linear layer of decoder. Inputs of these partially masked patches are composed of the concatenation of RGB pixel values (masked pixels are all 0 in 3 channels) and 0-1 masking maps. From the ablations in Tab. 1, the model trained with finetuned MAE performs slightly worse than the original one in FID. As shown in Fig. 4, we think that the finetuned MAE learns more explicit results, which makes the CNN restoration overfit MAE features. So these enlarged masked regions increase training difficulty and can be seen as noise regularization.

Table 1. Ablations of our full model enhanced by MAE with/without finetuning for partially masked patches (Partial F.T.).

Partial F.T.	PSNR↑	SSIM↑	FID↓	LPIPS↓
×	24.51	**0.864**	**25.49**	0.113
✓	**24.67**	**0.864**	26.00	**0.111**



(a) Origin input (b) Masked input (c) MAE w/o ft (d) MAE w/o ft+ACR (e) MAE with ft (d) MAE with ft+ACR

Fig. 4. Qualitative results of our full model enhanced by MAE with and without finetuning for partially masked patches.

3.2 Attention-based CNN Restoration (ACR)

The design of CNN modules in ACR is referred to LaMa [36], which is an encoder-decoder model including 4 downsampling convolutions, 9 Fast Fourier Convolution (FFC) blocks, and 4 upsampling convolutions. FFC has been demonstrated that it can tackle some HR inpainting cases with strong periodic textures. We further enhance ACR with prior features and attentions from MAE as follows.

Prior Features Upsampling. To overcome inpainting tasks with arbitrary resolutions, the local feature ensemble [6] is leveraged to facilitate the feature warping from MAE to ACR in various resolutions. Given 16×16 patch-wise prior features $\mathbf{F}_p \in \mathbb{R}^{16\times 16\times d}$ with dimension d from the MAE decoder, we resize them into 1/8 of the original image size with the bilinear interpolation. To indicate the continuous position in HR, the Cartesian spatial grid [21] *i.e.*, normalized 2d coordinates is concated to resized features as

$$\mathbf{F}'_p = \mathrm{Concat}(\mathrm{BilinearResize}(\mathbf{F}_p), \mathbf{C}) \in \mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times(d+2)}, \tag{1}$$

where h, w represent the original image height and width respectively; $\mathbf{C} \in \mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times 2}$ means normalized 2d coordinates grid valued from -1 to 1. As shown in the ablations of Tab. 2, such positional information can improve the HR inpainting results, and is complementary to learn smooth feature representations.

Prior Features Combination. Four gated deconvolutions [45] are leveraged to upsample $\mathbf{F}'_p$. Then these upsampled features are applied to the encoder of ACR. The gated mechanism works for making ACR filter corrupted features adaptively. To integrate prior features to ACR, the general solution is to element-wise add upsampled prior features to the downsampled ones of ACR. However,

Table 2. Ablations of models finetuned in 512×512 Places2 and tested in 1,000 images.

2D-coordinate	norm-pixel	PSNR↑	SSIM↑	FID↓	LPIPS↓
×	✓	24.35	0.879	25.47	0.135
✓	✓	24.31	**0.880**	25.33	0.135
✓	×	**24.37**	**0.880**	**25.30**	**0.132**

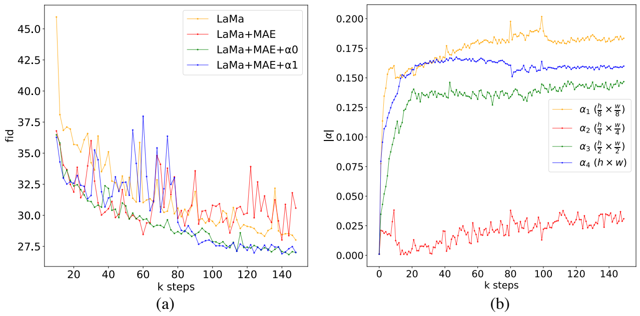


Fig. 5. (a) Ablations of various prior features combination methods. LaMa [36] is the baseline. 'LaMa+MAE' means directly adding MAE prior features to the ACR encoder. '+α0' and '+α1' indicate that prior features are multiplied with 0 and 1 initialized learnable-parameter α before the addition. (b) The line chart of $\alpha_j, j \in \{1,2,3,4\}$ shows absolute values for ACR features with different resolutions.

we observed an unstable training process with the vanilla element-wise addition as shown in the ablations of Fig. 5(a). We try to multiply trainable parameters $\alpha_j, j \in \{1,2,3,4\}$ to the prior features for the element-wise addition to ACR encoder features in $(\frac{h}{8} \times \frac{w}{8})$, $(\frac{h}{4} \times \frac{w}{4})$, $(\frac{h}{2} \times \frac{w}{2})$, and $(h \times w)$ respectively. Moreover, ablation studies of α_j initialized with 0 and 1 are shown in Fig. 5(a) and Tab. 3. The zero initialization enjoys a much more stable convergence. Therefore, the 0-initialized addition is adopted in our model, which is omitted in the follows for simplicity. We further analyze tendencies of different α_j trained with MAE layer 8 (*i.e.*, row 3 of Tab. 3) in Fig. 5(b). From Fig. 5(b), both high-level $(\frac{h}{8} \times \frac{w}{8})$ and low-level $(h \times w, \frac{h}{2} \times \frac{w}{2})$ features are important for the prior learning, since absolute values of $\alpha_1, \alpha_3, \alpha_4$ are large. And features in $\frac{h}{4} \times \frac{w}{4}$ seem have less effect during the inpainting.

Prior Attentions. Many inpainting researches show that the attention mechanism is useful for the image inpainting [44,45,54,43]. The classical contextual attention [44] used in inpainting is to aggregate masked features $\mathbf{F}_m$ with unmasked ones $\mathbf{F}_u$. The aggregation is based on the attention score $\mathbf{R}_{u,m}$ as

$$\begin{aligned} \cos_{u,m} &= \left\langle \frac{\mathbf{F}_u}{||\mathbf{F}_u||}, \frac{\mathbf{F}_m}{||\mathbf{F}_m||} \right\rangle \\ \mathbf{R}_{u,m} &= \mathrm{softmax}_u(\cos_{u,m}), \end{aligned} \tag{2}$$

Table 3. Ablations of models enhanced with different initialized parameters α. Column 'MAE' means that whether to use prior features from MAE. Column 'init-α' indicates initialized values of learnable parameter α for prior feature combination.

MAE	init-α	PSNR↑	SSIM↑	FID↓	LPIPS↓
×	no	24.12	0.859	28.01	0.124
✓	no	24.14	0.860	28.01	0.127
✓	0	**24.34**	0.860	**26.83**	0.117
✓	1	24.27	**0.861**	26.90	**0.115**

where $cos_{u,m}$ indicates normalized cosine similarities of masked and unmasked features; softmax$_u$ means the softmax normalization among all unmasked features. Then we can get the aggregated masked features $\mathbf{F}'_m = \sum_u \mathbf{R}_{u,m}\mathbf{F}_u$.

Unfortunately, the improvement of the attention module is not orthogonal to other effective inpainting strategies. We add two contextual attention (CA) modules [44] to ACR in the same positions as the prior feature aggregation shown in Fig. 2. But no improvement is achieved by the trainable contextual attention as shown in Tab. 4. We think that the restricted improvement is caused by the limited capacity of CNN models. Therefore, we try to use attention relations from the decoder of MAE to overcome the limited capacity. For the decoder layer l of MAE, we can get attention scores $\mathbf{R}^{(l)}_{u,m}$ as

$$\mathbf{R}^{(l)}_{u,m} = \text{softmax}(\frac{\mathbf{Q}^{(l)}\mathbf{K}^{(l)^T}}{\sqrt{d}} - inf \cdot \mathbf{M}), \tag{3}$$

where $\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}$ mean query and key of the attention; d is the channels; and $\mathbf{M}$ indicates a 0-1 mask map, where 1 means masked regions. We make the scores of masked regions to 0 in Eq. (3) that means all masked regions should only pay attention to unmasked ones. Then we average values of total 8 decoder attention scores, and get the prior attention scores $\mathbf{R}_p$ as

$$\mathbf{R}_p = \frac{\sum_{l=1}^{L} \mathbf{R}^{(l)}_{u,m}}{L}, L = 8. \tag{4}$$

The aggregations for getting masked features $\mathbf{F}'_m$ in ACR are executed in the start and the end of FFC blocks as shown in Fig. 2. Then $\mathbf{F}'_m$ is added to the original features as the residual. Note that we also multiply a zero-initialized learnable parameter to the $\mathbf{F}'_m$ before the addition instead of using LayerNorm as discussed in [1]. Besides, from Tab. 4, random masking used in the vanilla MAE fails to learn proper attention relations compared with mixed one discussed in Sec. 3.1. The ablation study about different attention layers from MAE are discussed in the Appendix.

3.3 Loss Functions

Loss Functions of MAE. Our MAE loss is the mean squared error (MSE) between MAE predictions and ground truth pixel values for masked patches as

Table 4. Ablations of models with different attention and MAE masking strategies; 'mixed' mask type means MAE pre-trained with both random, enlarged irregular and segmentation masks; 'prior attention' means using prior attention aggregation from MAE; 'trainable CA' indicates contextual attention [44].

MAE mask type	attention type	PSNR↑	SSIM↑	FID↓	LPIPS↓
mixed	no attention	24.34	0.860	26.84	0.117
mixed	trainable CA	24.13	0.859	26.99	0.123
random	prior attention	24.39	0.861	26.25	0.117
mixed	prior attention	**24.51**	**0.864**	**25.49**	**0.113**

in [18]. Besides, He *et al.* study to use normalized pixel values of each masked patch as the self-supervised target, which normalizes each masked patch with the mean and standard of this patch. Such a trick can improve the classification quality in their experiments. However, for the inpainting, the global relation among different patches is also important. The patch-wise normalization makes MAE learn more bias for each patch rather than the global information. Thus we study the MAE pre-training ablations with and without the patch-wise normalization as shown in Tab. 2. We find that non-normalized targets can achieve slightly better quality in HR inpainting.

Loss Functions of ACR. ACR is trained as a regular adversarial inpainting model. We adopt the same loss functions as LaMa [36], which include L1 loss, adversarial loss, feature match loss, and high receptive field (HRF) perceptual loss. Specifically, L1 loss is only used to constrain unmasked regions. The discriminator loss is based on PatchGAN [20]. And WGAN-GP [14] is used as the generator loss. The feature match loss [39] based on L1 loss between true and fake discriminator features is also used to stable the GAN training. Furthermore, we also leverage the segmentation pre-trained ResNet50 for the HRF loss as proposed in [36] to improve the inpainting quality. More details about the loss functions of ACR are in the Appendix.

4 Experiments

Datasets. Our model is trained on Places2 [52] and FFHQ [23]. For Places2, both the pre-training of MAE and training of our full model are based on the whole 1.8 million training images, and tested with 36,500 testing images. We prepare additional 1,000 testing images for 512×512 experiment. Detailed settings about main experiments and ablations are in the Appendix. We pre-train another MAE on 68,000 training set of FFHQ, which is also used to train the face inpainting model. And other 2,000 images work for testing. Our pre-trained MAE on Places2 and FFHQ can be well generalized to most real-world cases.

Implementation Detials. Our FAR is implemented with Pytorch based on two 48 GB NVIDIA RTX A6000 gpus. MAE is pre-trained on Places2 and FFHQ for 200 and 450 epochs respectively with batch size 512, while other

Table 5. Left: Results on 256×256 FFHQ and Places2 with mixed masks compared among Co-Mod (Co.) [51], LaMa (La.) [36], EC [30], CTSDG (CT.) [16] and ours. **Right:** Results on 512×512 Places2 testset with mixed masks compared among HiFill (Hi.) [43], Co-Mod (Co.) [51], LaMa (La.) [36] and ours. Metrics are PSNR (P.), SSIM (S.), FID (F.) and LPIPS (L.).

	FFHQ			Places2				
	Co.	La.	Ours	EC	Co.	La.	CT.	Ours
P.↑	25.2	26.6	**26.8**	23.3	22.5	24.3	23.4	**24.5**
S.↑	.889	.903	**.906**	.839	.843	.869	.835	**.871**
F.↓	**5.85**	6.38	6.15	6.21	1.49	1.63	11.2	**1.31**
L.↓	.085	.078	**.074**	.149	.246	.155	.185	**.101**

	P.↑	S.↑	F.↓	L.↓
Hi.	20.1	.764	65.4	.291
Co.	22.0	.843	30.0	.166
La.	24.1	.877	27.8	.149
Ours	**24.3**	**.880**	**25.3**	**.119**

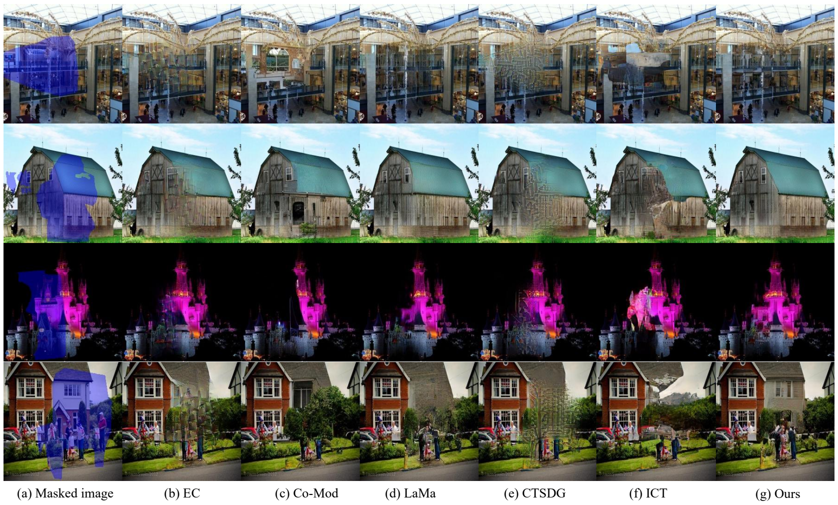


Fig. 6. Qualitative results of places2 256×256 images. From left to right are masked image, EC [30], Co-Mod [51], LaMa [36], CTSDG [16], ICT [38], and our results.

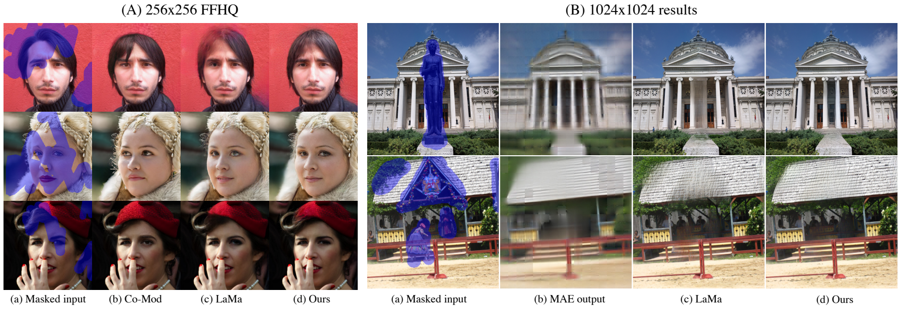


Fig. 7. Qualitative (A) 256×256 FFHQ, and (B) 1024×1024 results from network.

settings follow the released codes* except the masking strategy and the partially masked finetuning. For ACR, we employ the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the learning rates are 1e-3 and 1e-4 for the generator and the discriminator. We train our models with 850k steps on Places2 and 150k steps on FFHQ. For every 200k steps on Places2 and 100k steps on FFHQ, the learning rate is reduced by half. To save the computation, we finetune our model on images with higher resolutions, which are dynamically resized from 256 to 512 for 150k steps on Places2.

Masks Settings. To solve real-world application problems, we adopt the masking strategy in [5]. The masks consist of irregular brush masks and COCO [29] segmentation masks ranged from 10% to 50%. During the training, we combine these two types masks in 20%.

Comparison Methods. We compared our model with other state-of-the-art models including Edge Connect (EC) [30], Co-Modulation GAN (Co-Mod) [51], Large Mask inpainting (LaMa) [36], Conditional Texture and Structure Dual Generation (CTSDG) [16] and Image Completion with Transformers (ICT) [38] using the official pre-trained models. We also retrain LaMa on Places2 and FFHQ, and further finetune it with the same settings as ours in the HR inpainting for a fair comparison.

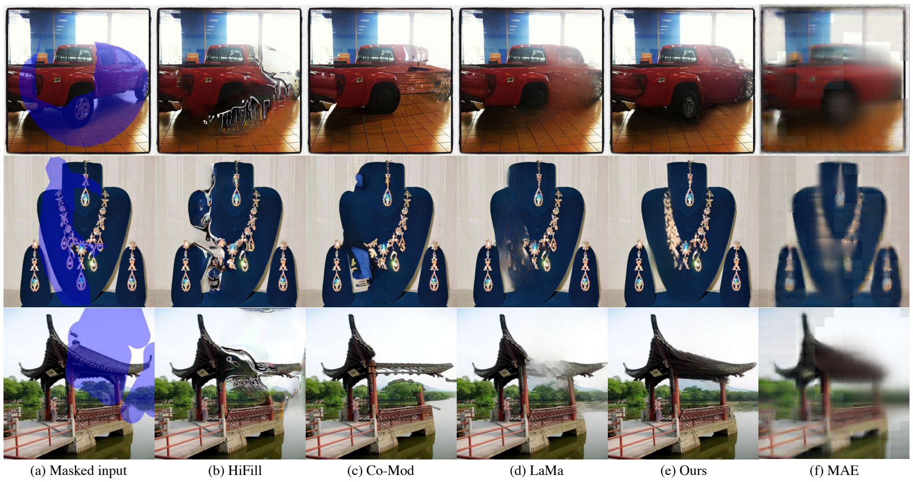

(a) Masked input (b) HiFill (c) Co-Mod (d) LaMa (e) Ours (f) MAE

Fig. 8. Qualitative results of places2 512×512 images. From left to right are masked image, HiFill [43], Co-Mod [51], LaMa [36], our results, and MAE predictions.

Quantitative Comparisons. We evaluate PSNR, SSIM [40], FID [19], and LPIPS [50] for both 256×256 and 512×512 in Tab. 5. For 256×256 results shown in the left of Tab. 5, our method achieves significant improvements based on the

* https://github.com/facebookresearch/mae

LaMa baseline. Moreover, Co-Mod results are also competitive. For the FFHQ results, since our MAE enhanced method can achieve much more stable and faithful results for face images, PSNR and SSIM of our method is better than Co-Mod. The powerful stylegan [24] architecture helps Co-Mod learn better textures with good FID, but our results have superior human perception (lower LPIPS). For the Places2, our results achieve best results in all metrics.

For the HR 512×512 results listed in the right of Tab. 5, our method can outperform all other competitors. Note that most methods fail to tackle HR inpainting tasks in complex scenes. Results of baseline method LaMa are competitive, but our methods still achieve certain advantages compared with LaMa.

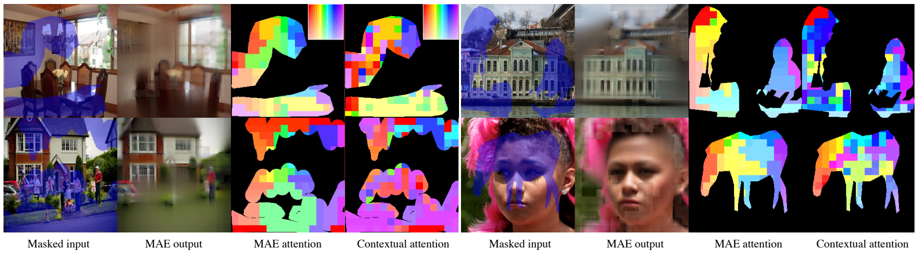


Fig. 9. Visualization of attentions from MAE and learned by contextual attention mechanism [44]. For the visualizations of attention maps, unmasked regions are ignored as black, and the patch attended with the highest attention score is shown in masked regions. Attention palettes are shown in the upper right corner of the first instance, which illustrate the exact positions attended mostly by masked patches.

Table 6. Efficiency of our model without pre-processing. Results are reported on 256×256 images with training and testing batch size 24 and 1 respectively.

MAE	Attention	Training(sec/batch)	Inference(sec/image)
×	×	1.0101	0.0425
✓	×	1.2510	0.0656
✓	✓	1.2604	0.0665

Qualitative Comparisons. We show 256×256 qualitative results of Places2 and FFHQ in Fig. 6 and Fig. 7(A) respectively. For the results in Fig. 6, results of EC and CTSDG are blurry and have obvious artifacts. The results of Co-Mod and LaMa have proper qualities. But these two methods generate some unreasonable building architectures, which also cause artifacts. For the results of ICT, they fail to get good quality due to the poor LR reconstruction. Moreover, the slow sampling strategy of ICT makes it hard to test large-scale image datasets for a quantitative comparison. Our method can achieve better inpainting results in both structures and textures. For face images, our method can achieve stable inpainted results with consistent eye gaze.

For qualitative results in 512×512, other compared algorithms fail to reconstruct reasonable semantics for certain objects in HR, such as car, necklace, and temple in Fig. 8, except our FAR. Such good results are benefited from informative prior features of MAE, which make the model understand the real categories of objects in HR cases. Besides, as shown in Fig. 8, our method avoids overfitting MAE outputs due to the enlarged masking strategy discussed in Sec. 3.1. We further provide some 1024 results in Fig. 7(B), which show that MAE can provide expressive priors for both structure and texture reconstructions.

Visualization of Attention. We show the visualization about the attention scores in Fig. 9. Patches attended with the highest attention score are shown in MAE and contextual attention maps, which can be located by the palettes in the first instance of Fig. 9. Attention scores from MAE are more consistent and reasonable compared with learning-based contextual attention [44]. Specifically, some irrelevant patches are attended by masked regions in the contextual attention, which leads to confused attention maps. Besides, although some MAE results are blurry, their attention relations are still effective.

Computation and Complexity. Benefited by the efficient design from [18] with 25% unmasked input tokens to the encoder, we should claim that the MAE pre-training is not heavier than CNNs. As discussed in implemented details, our MAE has been trained for 200 epochs in Places2 with about 10 days on two NVIDIA RTX A6000 gpus, which costs almost the same time for training a LaMa in Places2 for just 800k steps (only 28.4 epochs). Besides, we list times for both training and inference stages tested on A6000 in Tab. 6. It shows that the computation of the prior attention is negligible and can be ignored. MAE and gated convolutions increase about 0.024 seconds for predicting each image, which affects the training a little compared with the time-consuming GAN training. And the inference with 0.0665 seconds for each image is efficient enough for the user interaction, *e.g.*, object removal.

5 Conclusions

This paper proposes an MAE enhanced image inpainting model called FAR. We utilize a masked visual prediction based vision transformer – MAE to provide features for the CNN based inpainting model, which contain rich informative priors with meaningful semantics. Moreover, we apply the prior attention scores from the pre-trained MAE to aggregate masked features, which is proved to work better than learning contextual attention from scratch. Besides, many constructive issues about the pre-trained MAE and image inpainting are discussed in this paper. Our experiments show that our method can achieve good improvements with the prior features and attentions from MAE. Social impacts of our model, especially working on the face dataset are discussed in the Appendix.

Acknowledgements. This work was supported by SMSTC (2018SHZDZX01).

References

1. Bachlechner, T., Majumder, B.P., Mao, H.H., Cottrell, G.W., McAuley, J.: Rezero is all you need: Fast convergence at large depth. arXiv preprint arXiv:2003.04887 (2020)
2. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Bertalmío, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. Proceedings of the 27th annual conference on Computer graphics and interactive techniques (2000)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. ArXiv **abs/2005.14165** (2020)
5. Cao, C., Fu, Y.: Learning a sketch tensor space for image inpainting of man-made scenes. arXiv preprint arXiv:2103.15087 (2021)
6. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8628–8638 (2021)
7. Criminisi, A., Pérez, P., Toyama, K.: Object removal by exemplar-based inpainting. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. **2**, II–II (2003)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 vol. 1 (2005). https://doi.org/10.1109/CVPR.2005.177
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv **abs/1810.04805** (2019)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), `https://arxiv.org/abs/2010.11929`
12. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Akbari, Y.: Image inpainting: A review. Neural Processing Letters **51**(2), 2007–2028 (2020)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
14. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
15. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14134–14143 (2021)
16. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14134–14143 (October 2021)

17. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (SIGGRAPH 2007) **26**(3) (2007)
18. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. ArXiv **abs/2111.06377** (2021)
19. 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 (2018)
20. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
21. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. CoRR **abs/1506.02025** (2015), http://arxiv.org/abs/1506.02025
22. Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1745–1753 (2019)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4396–4405 (2019). https://doi.org/10.1109/CVPR.2019.00453
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
26. Lahiri, A., Jain, A.K., Agrawal, S., Mitra, P., Biswas, P.K.: Prior guided gan based semantic inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13696–13705 (2020)
27. Levin, A., Zomet, A., Weiss, Y.: Learning how to inpaint from global image statistics. Proceedings Ninth IEEE International Conference on Computer Vision pp. 305–312 vol.1 (2003)
28. Liao, L., Xiao, J., Wang, Z., Lin, C.W., Satoh, S.: Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 683–700. Springer (2020)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
30. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
31. Ntavelis, E., Romero, A., Bigdeli, S., Timofte, R., Hui, Z., Wang, X., Gao, X., Shin, C., Kim, T., Son, H., et al.: Aim 2020 challenge on image extreme inpainting. In: European Conference on Computer Vision. pp. 716–741. Springer (2020)
32. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
33. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
34. Roth, S., Black, M.J.: Fields of experts: a framework for learning image priors. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) **2**, 860–867 vol. 2 (2005)

35. 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)
36. 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)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
38. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. arXiv preprint arXiv:2103.14031 (2021)
39. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
41. Wei, C., Fan, H., Xie, S., Wu, C., Yuille, A.L., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. ArXiv **abs/2112.09133** (2021)
42. Yang, J., Qi, Z., Shi, Y.: Learning to incorporate structure knowledge for image inpainting. CoRR **abs/2002.04170** (2020), https://arxiv.org/abs/2002.04170
43. 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 Conference on Computer Vision and Pattern Recognition. pp. 7508–7517 (2020)
44. 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)
45. 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/CVF International Conference on Computer Vision. pp. 4471–4480 (2019)
46. Yu, Y., Zhan, F., Lu, S., Pan, J., Ma, F., Xie, X., Miao, C.: Wavefill: A wavelet-based generation network for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14114–14123 (2021)
47. Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. arXiv preprint arXiv:2104.12335 (2021)
48. Zeng, Y., Fu, J., Chao, H., Guo, B.: Aggregated contextual transformations for high-resolution image inpainting. arXiv preprint arXiv:2104.01431 (2021)
49. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. In: European Conference on Computer Vision. pp. 1–17. Springer (2020)
50. 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)
51. 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)
52. 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)

53. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
54. Zhou, T., Ding, C., Lin, S., Wang, X., Tao, D.: Learning oracle attention for high-fidelity face completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7680–7689 (2020)