Image Completion with Adaptive Multi-Temperature Mask-Guided Attention

Xiang Zhou[1]
zhoux2020@mail.sustech.edu.cn
Yuan Zeng[†2]
zengy3@sustech.edu.cn
Yi Gong[†3]
gongy@sustech.edu.cn

[1] Southern University of Science and Technology(SUSTech), China
[2] Academy for Advanced Interdisciplinary Studies, SUSTech, China
[3] University Key Laboratory of Guangdong Province, SUSTech, China
[†] Corresponding authors

Abstract

Leveraging distant contextual information and self-similarity of natural images in deep learning-based models is important for high-quality image completion with large missing regions. Most of the deep generative adversarial network (GAN)-based image completion methods attempt this via increasing receptive field size of convolutions and integrating an attention module. However, existing attention mechanisms treat the softness of the attention for different types of features with the same scale, which may be inferior since the same softness of the attention may lead attention made on limited spatial locations in feature space. To address this limitation, we design a new two-stage image completion model and propose an attention mechanism called Adaptive multi-Temperature Mask-guided Attention (ATMA). The ATMA performs non-local processing and adaptively adjusts the softness of attention by means of multiple learnable temperatures. The proposed model infers a coarse inpainting result via a gated convolution neural network in the first stage and refines appearance consistency between generated regions and known regions via ATMA in the second stage. Experiments demonstrate superior performance compared to state-of-the-art methods on benchmark datasets including CelebA-HQ, Paris StreetView and Places2.

1 Introduction

Image completion targets at filling reasonable contents into missing regions of such that the completion is visually realistic. Research on image completion has often been motivated by various applications such as image editing, image restoration and object removal. Since it is impossible to exactly restore the missing regions, image completion is an ill-posed problem. To address this problem, early image completion approaches like PatchMatch [3] assume that a source image contains appropriate information of the missing regions, such as similar structures or patches, and synthesize the missing regions by searching similar patches within the image and copy-pasting them into the missing regions. These approaches work well for synthesizing texture-consistent outputs but fail to generate semantically meaningful contents, especially when missing regions contain complicated scenes like objects and faces.

The main challenge of image completion is to synthesize both local textured patterns and global semantics that are coherent with known regions. To address this challenge, recent studies focus on directly learning to infer semantic contents and meaningful hidden representations using deep neural networks. These approaches can be categorized into two classes: one-stage approaches and two-stage approaches. One-stage approaches treat image inpainting as a conditional generation problem and use only one deep generative network to synthesize new contents, while two-stage approaches consist of a content inference network for coarse image completion and a refinement network for generating high-quality results. In order to draw upon information from a sufficiently broad context in deep learning-based models, most network architectures have been explored to increase the receptive field size [13, 21, 23]. However, these networks trade off context size for localization accuracy and may not be able to include the receptive field with the features of interest.

Inspired by traditional image completion algorithms, where non-local patch matching is performed to increase the receptive field, recent works attempt to refine the visual appearance by optimizing similarities between generated patches and the matched patches in known regions. Incorporating patch matching into deep neural networks for image completion has been considered very recently [14, 22]. To improve non-local processing for better image completion, these methods employed a patch matching process via replacing the filtering of matched patches with a convolution network, which was designed as an attention mechanism. However, these methods weight the similarity with a constantly scaled softmax to get attention score for each pixel, making attention on limited spatial locations in feature space.

To tackle the problem mentioned above, we propose an attention mechanism called Adaptive Multi-temperature Mask-guided Attention (ATMA), which is integrated into a deep generative adversarial network (GAN)-based model for image completion. Our model consists of two networks: the first one is a gated convolution based coarse image completion network and the second one is an ATMA-based appearance refinement network. In addition, we introduce multiple temperature parameters in ATMA, which can be learned by the model. With these temperature parameters, our model is able to automatically generate multiple sets of attention scores via tuning the degree of softness, which eventually enriches the feature representation. Our main contributions are summarized as follows:

- We introduce multiple self-adaptive temperature parameters to control the scale of the softness of the attention in image completion, learning different attention scores to extract the related features from different locations in feature space.
- We develop a novel attention layer, called Adaptive Multi-temperature Mask-guided Attention (ATMA), with multiple learnable temperatures. It attends on matched feature patches at multiple distant spatial locations and enables end-to-end trainable non-local patch matching based on the principle of self-similarity.
- Experimental results on three challenge inpainting datasets including CelebA-HQ, Paris StreetView and Places2 demonstrate that our proposed model achieves higher-quality image completion than existing state-of-the-arts.

2 Related work

2.1 Image Completion

Early traditional diffusion-based or patch-based image completion approaches propagate pixels from known regions into the missing regions by variational algorithms [2] or patch matching [3, 5]. Those approaches produce convincing continuations of the background but cannot fill in missing regions with novel structure and semantics. Recent breakthroughs in deep learning enabled significant improvements in image completion. Phatak et al. [17] introduced GANs [6] to image completion, where an encoder-decoder network was trained using reconstruction and adversarial losses for better recovering semantics. Iizuka et al. [8] introduced global and local discriminators in GAN-based model to improve local texture and overall image layout. In [27], a variational auto-encoders [10] (VAE)-based model with two parallel paths was presented for pluralistic image completion. Partial convolution [12] was proposed to better handle irregular holes, where the convolution was masked and re-normalized to utilize valid pixels only.

Our work is more closely related to two-stage approaches which first infer a coarse image and afterwards fill in visually realistic appearance in a second stage. Yang et al. [20] presented a high resolution image completion approach using multi-scale neural patch synthesis in the second stage. Nazeri et al. [16] proposed a two-stage model EdgeConnet first predicting salient edges and then generating completion result guided by edges. Zeng et al. [25] proposed to do coarse image completion using a U-Net based deep generative model in the first stage and compose novel high-quality outputs by copying high-frequency missing information from different training exemplars in the second stage. Yu et al. [23] proposed a generative image completion network with contextual attention, where a GAN was first used for coarse image completion and a contextual attention module is then integrated into a refinement network for borrowing features from background in the second stage. Zheng et al. [28] proposed a transformer-based image completion network to fill reasonable content into the missing regions in a first phase and designed an attention-aware layer in a second phase to better exploit long-range context relations.

2.2 Attention Learning

Attention learning has benefited from recent advances in deep learning. Many studies on deep learning-based image generation have demonstrated that learning attention enables the generation of more realistic compared to classic GANs. For example, Zhang et al. [26] proposed self-attention GANs to consider non-local relationships in the feature space using a self-attention mechanism, which produces globally realistic images. Attention learning has also been explored in GAN-based image completion approaches, where an attention layer is integrated into an encoder-decoder network to capture distant contextual information and generate more visually realistic outputs. Yu et al. [22] proposed a contextual attention to explicitly attend on related features at distant spatial locations. Xie et al. [19] presented a learnable bidirectional attention map module for image completion. Zheng et al. introduced an attention-aware layer to better exploit distantly related features in [28]. Our proposed attention mechanism is inspired by self-attention mechanisms in [22] and [18], but we introduce multiple self-adaptive temperatures to control the softness of the attention, learning different attention scores for different locations in feature space.

3 Approach

Given a masked image $\mathbf{I}_{in}$, which is a degraded version of the target image $\mathbf{I}_{gt}$, image completion aims to estimate a map $\mathcal{F} : \mathbf{I}_{in} \rightarrow \mathbf{I}_{gt}$ from the masked image $\mathbf{I}_{in}$ to the target image $\mathbf{I}_{gt}$. We design a two-stage framework for image completion. Specifically, we modify the

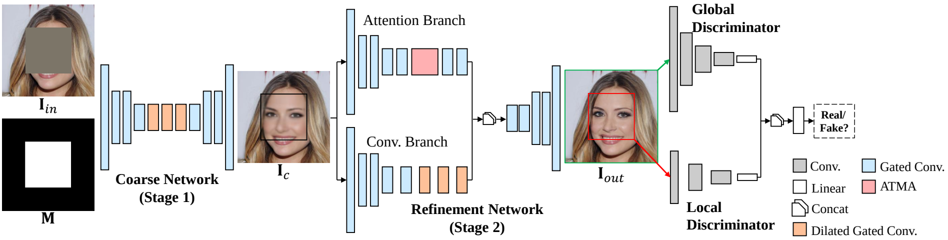

Figure 1: An overview of our two-stage image completion framework

image completion architecture in [23] by using the proposed ATMA and WGAN-based adversarial losses in the second stage. An overview of our framework is shown in Figure 1. The framework consists of a coarse network and a refinement network. In the first stage, an encoder-decoder network with gated convolution at both ends and dilated gated convolution in the middle is designed to generate a rough estimation of missing regions. The inputs of the network is a masked image and a binary mask indicating the hole regions, and the output $\mathbf{I}_c$ is an intermediate image. $\mathbf{I}_c$ is then used as the input of the refinement network. Since the intermediate inpainting result is semantically plausible and coherent with known regions, it allows the encoder of the refinement network to learn better feature representation for better appearance consistency. The refinement network is an encoder-decoder network with two encoders, where the first encoder tries to do non-local feature matching with gated convolution layers and a ATMA layer, and the second one tends to aggregate local information using gated and dilated gated convolution layers. Later, the features extracted from the two encoders are concatenated and fed into a single decoder to generate the final output. We present the architecture details of our model in supplementary material.

3.1 Revisiting Gated Convolution

Gated convolution [23] performs dynamic element-level feature selection in each channel and each spatial location. It consists of two transformations with convolutions or sets of convolutions, and aims to transform an input data $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ into an output $\mathbf{O} \in \mathbb{R}^{H' \times W' \times C'}$. The two transformations are $\mathcal{F}_c : \mathbf{X} \rightarrow \mathbf{V}$ and $\mathcal{F}_d : \mathbf{X} \rightarrow \mathbf{G}$, where $\mathbf{V}, \mathbf{G} \in \mathbb{R}^{H' \times W' \times C'}$ are the output of transformation $\mathcal{F}_c$ and $\mathcal{F}_d$, respectively. The output $\mathbf{O} \in \mathbb{R}^{H' \times W' \times C'}$ of a gated convolution is given by

$$\mathbf{O} = \sigma(\mathcal{F}_d(\mathbf{X})) \odot \phi(\mathcal{F}_c(\mathbf{X})), \tag{1}$$

where σ is sigmoid function, and $\odot$ denotes element-wise product. ϕ can be any activation function, *e.g.* ReLU and LeakyReLU. Gated convolution performs dynamic feature selection between the mask regions and existing regions by applying a soft mask to the feature map. It gradually fills in missing features and generates high-quality outputs.

3.2 Adaptive Multi-Temperature Mask-Guided Attention

The proposed Adaptive Multi-Temperature Mask-Guided Attention (ATMA) module is illustrated in Figure 2. It aims to transform an input feature maps $\mathbf{F}_{in} \in \mathbb{R}^{H \times W \times C}$ to a refined feature maps $\mathbf{F}_{out} \in \mathbb{R}^{H \times W \times C}$. The module first performs two blocks: similarity comparison and embedding network, in parallel, and then their outputs are fed into the Temperature

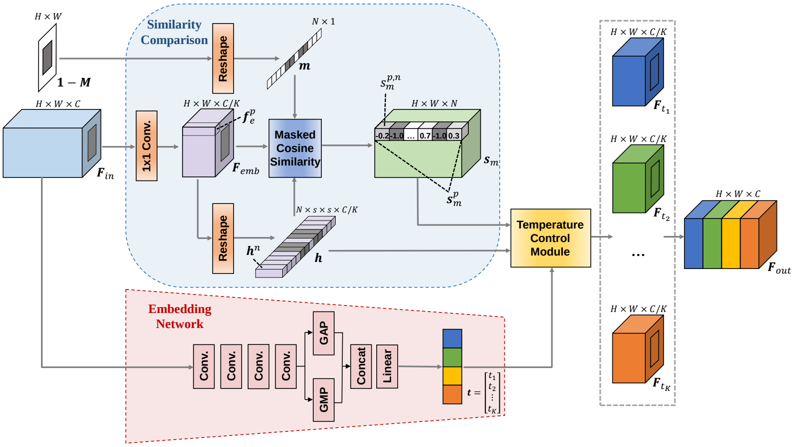

Figure 2: Illustration of the Adaptive Multi-Temperature Mask-Guided Attention Module

Control Module (TCM). The similarity comparison block aims to match patches in generated regions to patches in known regions using masked cosine similarity $\mathbf{s}_m \in \mathbb{R}^{H \times W \times C}$. The embedding network block is performed to learn K temperature parameters $\mathbf{t} = [t_1, t_2, \cdots, t_K]$. The TCM is then used to generate an attention refined feature maps $\mathbf{F}_{out}$ via attending on the extracted feature cube $\mathbf{h}$ with attention scores $\mathbf{s}_m$, which are controlled by the learned K temperatures.

Masked Cosine Similarity Let us consider a problem of matching a query pixel p in generated regions to pixels in known regions. We first use 1×1 convolution to convert the input feature cube $\mathbf{F}_{in}$ with dimension of $H \times W \times C$ into embedded feature cube $\mathbf{F}_{emb}$ with dimension of $H \times W \times C/K$. Then, we extract patches $(s \times s)$ from $\mathbf{F}_{emb}$ and reshape them to $\mathbf{h} \in \mathbb{R}^{N \times s \times s \times C/K}$, where N is the number of extracted patches. Next, we extract patches centered on pixel p from $\mathbf{F}_{emb}$ and build a feature cube $\mathbf{f}_e^p$ with size of $s \times s \times C/K$. After that, we calculate cosine similarity between pixel p and all extracted patches $\mathbf{h}$. Specifically, the cosine similarity between feature representation of pixel p and the n-th feature cube $\mathbf{h}^n$ is given as

$$s^{p,n} = \frac{\mathbf{f}_e^p \cdot \mathbf{h}^n}{\left\|\mathbf{f}_e^p\right\| \left\|\mathbf{h}^n\right\|}. \tag{2}$$

In parallel, we convert the input mask matrix $\mathbf{1} - \mathbf{M}$ into a vector $\mathbf{m} \in \mathbb{R}^{N \times 1}$ with values 0 for missing regions and 1 for elsewhere. After that, the similarity is recalculated based on the mask vector, that is

$$\mathbf{s}_m^p = (\mathbf{s}^p + \lambda) \odot \mathbf{m} - \lambda, \tag{3}$$

where $\mathbf{s}_m^p \in \mathbb{R}^{N \times 1}$ denotes the masked cosine similarity of query pixel p and λ $(\lambda \geq 1)$ is a hyper-parameter. The masked cosine similarity tends to reduce the similarity score $s^{p,n}$ to $-\lambda$ when n-th feature cube $\mathbf{h}^n$ is extracted from missing regions.

Embedding Network Similar to the term temperature in softmax function, we use temperature to change the weight distribution. For our ATMA module, a higher temperature indicates that the module is more confident to use the most similar neural patch, and may produce repetitive patches. A lower temperature makes the module utilize a greater number of neural patches for patch generation. We design an embedding network $\mathcal{F}_e$ to learn temperatures and adaptively adjusting the softness of the weight distribution. As shown in Figure 1, the network first uses convolutions for feature embedding. Then, global average pooling (GAP) and global max pooling (GMP) are used to aggregate global spatial information. After that, the outputs of GAP and GMP are concatenated and transformed by a linear layer and LeakyReLU, generating the final K temperatures, that is $\mathbf{t} = \mathcal{F}_e(\mathbf{F}_{in})$.

Temperature Control Module For each temperature t_k, we convert $\mathbf{s}_m$ into a weight matrix $\mathbf{w}_{t_k} \in \mathbb{R}^{H \times W \times N}$ using channel-wise softmax with temperature parameter t_k. Each element in $\mathbf{w}_{t_k}$ is given by

$$w_{t_k}^{p,n} = \frac{\exp(t_k s_m^{p,n})}{\sum\limits_n \exp(t_k s_m^{p,n})}, \tag{4}$$

with $\sum_{n=1}^{N} w_{t_k}^{p,n} = 1$. The refined feature map $\mathbf{F}_{t_k}$ can then be calculated by doing transposed convolution with the weight matrix $\mathbf{w}_{t_k}$ and kernels $\mathbf{h}$. Finally, all K refined feature maps are concatenated to get the final attention refined feature cube as $\mathbf{F}_{out} = \text{concat}\left([\mathbf{F}_{t_1}, \mathbf{F}_{t_2}, \cdots, \mathbf{F}_{t_K}]\right)$.

3.3 Loss Functions

We train our model with a weighted sum of per-pixel reconstruction losses and WGAN adversarial losses, where the per-pixel reconstruction losses regress the missing regions to the ground truth and WGAN adversarial losses learn to match potentially real images and generate visually realistic outputs.

Pixel Reconstruction Losses Given input image with hole $\mathbf{I}_{in}$, binary mask $\mathbf{M}$, the intermediate output $\mathbf{I}_c$ of the coarse network $\mathcal{F}_g$ and the final output $\mathbf{I}_{out}$ of the refinement network $\mathcal{F}_r$. We define our per-pixel reconstruction losses as $\mathcal{L}_c = \left\|\mathbf{M} \odot (\mathcal{F}_g(\mathbf{I}_{in}, \mathbf{M}) - \mathbf{I}_{gt})\right\|_1$ and $\mathcal{L}_r = \left\|\mathbf{M} \odot (\mathcal{F}_r(\mathcal{F}_g(\mathbf{I}_{in}, \mathbf{M}), \mathbf{M}) - \mathbf{I}_{gt})\right\|_1$, where $\mathcal{L}_c$ and $\mathcal{L}_r$ denote the l_1 losses on the intermediate network output $\mathbf{I}_c$ and the final network output $\mathbf{I}_{out}$. The mask $\mathbf{M}$ makes the losses to be computed only on the missing regions.

WGAN Adversarial Objective Adversarial loss [6] has been widely used in image generation to improve the visual fidelity of generated outputs. Our adversarial loss is based on Wasserstein Generative Adversarial Network (WGAN-GP) [1, 7], which has better training stability than GAN and works well for image generation tasks. To further enhance the global and local consistency of the generated outputs, we design two discriminators for global and local perception similar to [22]. Our adversarial objective is formulated as

$$\begin{aligned} \mathcal{L}_{adv} = \min_G \max_D \; & \mathbb{E}_{\mathbf{I}_{gt} \sim p_{\text{data}}(\mathbf{I}_{gt})} \left[D(\mathbf{I}_{gt})\right] - \mathbb{E}_{\mathbf{I}_{out} \sim p_{\text{data}}(\mathbf{I}_{out})} \left[D(\mathbf{I}_{out})\right] \\ & + \beta \; \mathbb{E}_{\hat{\mathbf{I}} \sim p_{\text{data}}(\hat{\mathbf{I}})} \left[(\left\|\nabla_{\hat{\mathbf{I}}} D(\hat{\mathbf{I}})\right\|_2 - 1)^2\right], \end{aligned} \tag{5}$$

where D denotes the global and local discriminators, and $G = \mathcal{F}_r(\mathcal{F}_g(\cdot))$ denotes the image completion model. $\hat{\mathbf{I}}$ is randomly sampled from $\mathbf{I}_{gt}$ and $\mathbf{I}_{out}$ with $\hat{\mathbf{I}} = \mathbf{I}_{out} + \alpha(\mathbf{I}_{gt} - \mathbf{I}_{out})$, where $\alpha \sim U(0,1)$ is a random value. We set $\beta = 10$ in our experiments.

Model objective The objective function of our model is defined as $\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_{adv} \mathcal{L}_{adv}$, where λ_c, λ_r and λ_{adv} are scales. We set $\lambda_c = 1.2$, $\lambda_r = 1$ and $\lambda_{adv} = 0.001$ in our implementations. Similar to the model in [23], our model can also be used for free-form with different inputs and loss functions. For image inpainting of irregular holes at arbitrary locations, we use $\mathcal{L}_c = \left\|\mathcal{F}_g(\mathbf{I}_{in}, \mathbf{M}) - \mathbf{I}_{gt}\right\|_1$ in our loss functions.

4 Experiments

We evaluate the proposed framework on three datasets including CelebA-HQ [11], Paris StreetView [4] and Places2 [29]. This section conducts three experiments to analyze the inpainting performance of our approach. First, we study the effect of a different number of learnable temperature parameters on inpainting outputs. Later, we compare our model with a baseline model. After that, we compare our approach with previous state-of-the-art methods: *i.e.*, PatchMatch [3], DeepFillv1 [22], ParConv [12], PENNET [24], MEDFE [15] and DeepFillv2 [23]. All the masks and images for training and testing are with the size of 256×256. For a fair evaluation on model generalization abilities, we conduct experiments on filling center holes and irregular holes on the input images. The center hole is a fixed center mask with the size of 128×128. We generate irregular masks using the method in [23]. We use Content-Aware Fill function from Photoshop for producing results of PatchMatch, and use the provided models to generate results of DeepFillv1, PENNET, MEDFE and DeepFillv2. We train ParConv on the same training set and report its results on the same validation set.

The proposed model is implemented using PyTorch v1.8. We train our model on a single NVIDIA Geforce RTX 3090 GPU (24GB) with a batch size of 16, and use the Adam [9] with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ for optimization. The learning rate is 0.0001. The hyperparameter λ is 2. The training process is terminated when the validation loss converges. It takes around 5 days for training the CelebA-HQ model, 3 days for training the Paris StreetView model and 5 days for training the Places2 model.

4.1 Ablation study

Single-temperature vs Multi-temperature Figure 3 shows the inpainting results of the proposed model with different numbers of learnable temperature parameters, which are $K = 1$, $K = 2$ and $K = 4$. Compared to the results of the model with 1 temperature, the model with 2 temperatures provides outputs with fewer artifacts. In addition, the model with 4 temperatures generates more blurring results than the models with $K = 1$ and $K = 2$. This can be explained that we consider the trade-off between similarity comparison and information loss in our attention module. Similarity comparison in high dimension is inevitably affected by the curse of dimensionality. Our attention module uses a 1×1 convolution to reduce channel dimension to $1/K$ times. This can alleviate the dimensional problem in similarity comparison, but will cause information loss. The larger amount of temperatures, the fewer output channels, and the more information loss, which eventually causes the degeneration of the inpainting results. In addition, the channel dimension reduction can significantly reduce memory and computation cost in experiments.

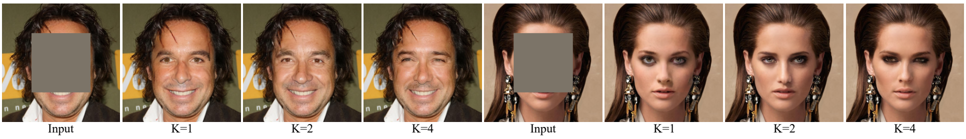

Figure 3: Image completion using different numbers of learnable temperatures. The outputs of the model with 1 temperature are sharper than the models with 2 and 4 temperatures, but are not well coherent. The outputs of the model with 4 temperatures are smoother than the models with 1 and 2 temperatures but details are still blurry (*e.g.*, eyes).

4.2 Comparison to baseline model

To investigate the effectiveness of the proposed ATMA, we compare our model with a baseline model. The baseline model is a modified version of our model, where the ATMA module is replaced by the contextual attention module proposed in [22]. The comparison results are shown in Figure 4. We observe that the generated outputs of the proposed model are more natural than those of the baseline model. This can be explained that ATMA attention using multiple temperatures can control the softness of the attention and effectively capture long-range context information, which helps to produce high-quality image completion results.

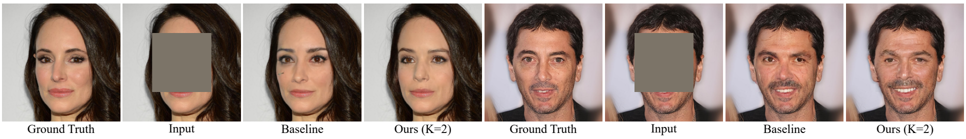

Figure 4: Visual comparison between our model and the baseline model on CelebA-HQ.

4.3 Comparison to existing work

Qualitative Comparison Figure 5 and Figure 6 shows the comparison results of central hole completion on CelebA-HQ and Paris StreetView, respectively. We observe that PatchMatch without learning from training images fails to recover complex structures and textures, and all deep learning-based models can generate plausible content. DeepFillv1 with vanilla convolution generates obvious visual artifacts. ParConv without GAN adversarial loss can effectively generate plausible semantics and structures but many areas (*e.g.*, eyes and mouths) are still blurry, which leads to the details in the generated regions are not as delicate as the background. PENNET produces obvious edge responses surrounding holes. MEDFE can generate more natural results, but it still generates outputs with visible color and structure inconsistency. DeepFillv2 with gated convolution and contextual attention produces better results but still exhibits observable unpleasant boundaries and artifacts. Our model with ATMA produces more visually pleasant results with fine details, since the proposed ATMA attention with multiple learnable temperatures allows the model to autonomously regulate

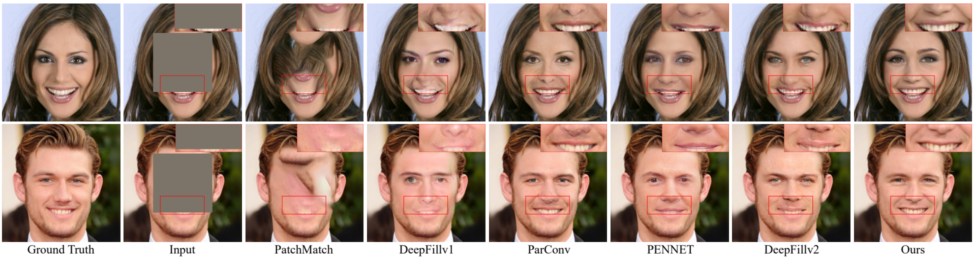

Figure 5: Qualitative comparison of central hole completion on CelebA-HQ. Zoom-in to see differences between methods (*e.g.*, details of eyes, noses and mouth).

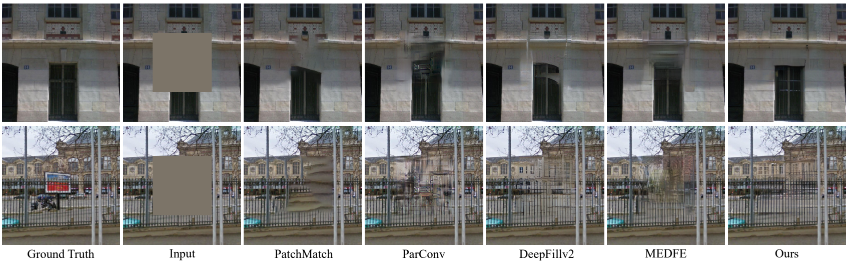

Figure 6: Comparison results of central hole completion on Paris StreetView.

the attention softness, conducting effective distant contextual correlations and resulting in high-quality inpainting results.

Figure 7 and Figure 8 shows the visual comparison results of irregular holes image inpainting on CelebA-HQ, Paris StreetView and Places2. We observe that our model can leverage distant high-quality contextual information and consequently generates results with higher-quality textures than the other inpainting approaches.

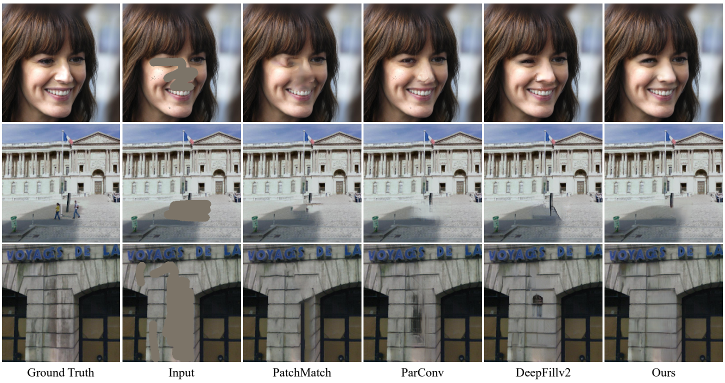

Figure 7: Comparison samples of irregular holes inpainting on CelebA-HQ and Paris StreetView.

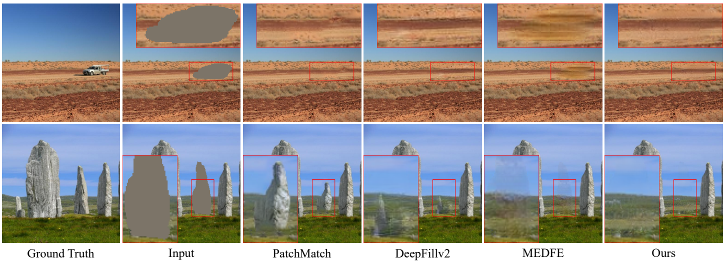

Figure 8: Qualitative comparisons of irregular holes image inpainting on Places2.

Quantitative Comparison We report our quantitative comparison results in Table 1 in terms of mean l_1 loss, mean l_2 loss, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) on the validation set on CelebA-HQ. Numerical results in Table 1 show that learning models perform much better than PatchMatch, and our approach outperforms all the other approaches.

Failure cases Although our model can generate high-quality inpainting results, it generally fails when a heavily structured objects is partially masked, see Figure 9.

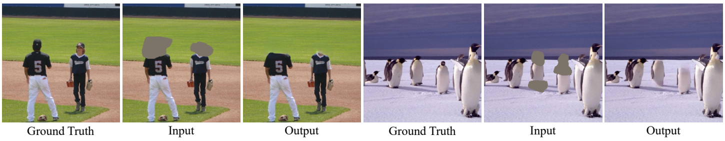

Figure 9: Failure cases of our method when a person or an animal is partially masked.

Table 1: Results of mean l_1, mean l_2, PSNR and SSIM on validation set on CelebA-HQ.

	mean l_1 ↓	mean l_2 ↓	PSNR (dB) ↑	SSIM ↑
PatchMatch	4.50 %	1.53%	18.8	82.8 %
DeepFillv1	2.29 %	0.43%	24.2	87.7 %
ParConv	1.86 %	0.30%	25.9	89.5 %
PENNET	1.93 %	0.31%	25.7	89.2 %
DeepFillv2	1.91 %	0.33%	25.7	89.0 %
ours ($K = 2$)	**1.68** %	**0.29** %	**26.1**	**89.9** %

5 Conclusion

We proposed a novel Adaptive Multi-temperature Mask-guided Attention (ATMA) mechanism to learn distant contextual correlations, which performs non-local matching based on the principle of self-similarity. In addition, we integrated it into a new two-stage model for high-quality image completion with large missing regions. Experiment results demonstrated that the outputs of our model are more natural than a baseline model with contextual attention. Furthermore, qualitative results and quantitative comparisons showed that the proposed image completion model with ATMA performs favorably against state-of-the-arts in generating more visually pleasant results.

6 Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This work is supported by National Key R&D Program of China under Grant 2019YFB1802800, National Natural Science Foundation in China under Grants 62071212 and 62106095, Guangdong Science and Technology Program under Grant 2019A1515110479, Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515130003, Shenzhen Science and Technology Program under Grant JCYJ202001091414409, and Educational Commission of Guangdong Province of China under Grants 2020ZDZX3057 and 2017KZDXM075.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8):1200–1211, August 2001. ISSN 1057-7149.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):1–11, July 2009.

[4] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4):1–9, 2012.

[5] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, 1999.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[10] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019.

[11] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[13] Guilin Liu, Kevin J. Shih, Ting-Chun Wang, Fitsum A. Reda, Karan Sapra, Zhiding Yu, Andrew Tao, and Bryan Catanzaro. Partial convolution based padding. In *arXiv preprint arXiv:1811.11718*, 2018.

[14] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[15] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 725–741. Springer, 2020.

[16] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[17] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2016.

[18] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural Information Processing Systems*, 31:1087–1098, 2018.

[19] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[20] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, July 2017.

[21] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, May 2016.

[22] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 5505–5514. IEEE Computer Society, 2018.

[23] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019.

[24] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1494, 2019.

[25] Yuan Zeng, Yi Gong, and Jin Zhang. Feature learning and patch matching for diverse image inpainting. *Pattern Recognition*, 119:108036, 2021. ISSN 0031-3203.

[26] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363. PMLR, 2019.

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

[28] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Tfill: Image completion via a transformer-based architecture. *CoRR*, abs/2104.00845, 2021.

[29] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.