Artistic Face Image Inpainting with Transformer

Chengyang Li 30920221154274,[1] **Qifeng Dai 30920221154265,** [1] **Bin Luo 31520221154249,** [1] **Ziying Pan 30920221154280,** [1] **Chenyang Xie 30920221154285,** [1]

School of Informatics, Xiamen University
{ harshwinter, daiqifeng, 31520221154249, 30920221154280, 30920221154285 }@stu.xmu.edu.cn

Abstract

Recent research in the area of image inpainting has yielded promising results. However, there is still a lack of inpainting algorithms for artistic face images due to the lack of paired training data. To solve this problem, we propose a transformer-based artistic image inpainting algorithm. Our framework consists of two modules: 1) **Style Transfer.** We use the style transfer algorithm to generate artistic face images. 2) **Image Inpainting.** The generated images are then masked to train our image inpainting model, which allows us to obtain high-resolution images. Instead of using a traditional convolutional neural network (CNN), we use a transformer model because it focuses more on global features, which is important to ensure a consistent global style in the output of the image inpainting algorithm. Also, we redesign a new loss function by observing the completion results of the images, and our ablation study shows that our loss function has a better performance for the completion around the eyes. With a qualitative and quantitative evaluation of our method, we show that the proposed method works well for painting artistic face images.

1 Introduction

As an important branch in the field of computer vision, image inpainting algorithms have been studied by scholars in recent years and have become a hot issue. In our daily life, we often encounter the situation of partial loss of image data. To solve this problem, image repair comes into our view. In short, the goal of image inpainting is to infer a reasonable structure and color of the missing parts based on the existing parts of the image we need to repair. It has a wide range of applications, from object removal (Barnes et al. 2009) and image manipulation (Jo and Park 2019) to photo and artistic image restoration (Wan et al. 2020). In this work, we will focus only on the repair of artistic images.

In recent years, deep learning-based image repair methods have become mainstream in this field and are gradually replacing patch-based approaches, such as PatchMatch (Barnes et al. 2009). As a pioneer, ContextEncoder (Pathak et al. 2016) firstly uses a convolutional neural network to restore images. Immediately after, (Iizuka, Simo-Serra, and Ishikawa 2017) introduces global and local environment discriminators to train so that the generated image is indistinguishable from the real image. Since the convolution kernel is not sensitive to the information difference between the valid and invalid regions, Partial Convolution (Liu et al. 2018) is proposed to improve the operational efficiency by adding the mask to participate in the convolution operation. For the repair of contextual semantic information, (Yu et al. 2018a) incorporates the attention mechanism, and (Zheng et al. 2022) treats image completion as a directionless sequence-to-sequence prediction task, and deploy a transformer to directly capture long-range dependence in the encoder. AOT-GAN (Zeng et al. 2022) compliments high-resolution images for the problem of fine-grained texture synthesis of large missing regions. These previous models can achieve good results in terms of image realism, contextual information, image resolution, etc. However, all of them do not perform well in artistic face image inpainting. That is because, unlike ordinary photographs and landscapes, artistic images have more distinctive brushstrokes and textures that are not found inside the training set(FFHQ (Karras, Laine, and Aila 2019), CelebA (Karras et al. 2017), etc.) commonly used in image inpainting algorithms. In real life, there are no open-source large-scale artistic face images, which limits the application of image inpainting methods in artistic face image repair.

The existing style transfer algorithm can transfer special strokes and textures to real face images to obtain artistic face images. Style transfer aims to re-render features such as textures and colors of one image to another image and ensure the structure of the latter remains unchanged. The current style transfer algorithms are mainly divided into optimization-based approaches and feedforward-based approaches. As optimization-based methods tend to produce higher-quality images, we leverage an optimization-based style transfer (Kolkin, Salavon, and Shakhnarovich 2019) algorithm to build a high-quality artistic face image dataset base on FFHQ (Karras, Laine, and Aila 2019) which can be used to train our image completion model.

In addition to a high-quality dataset, an appropriate network model is also essential for repairing artistic face images. Recent work ICT (Wan et al. 2021) has achieved an excellent result in the image completion task using Transformer (Vaswani et al. 2017). In particular, the convolution operation of CNN has trouble modeling the global structure of the image and has a unique output result. Transformer, however, understands the global structure through the attention mechanism and supports pluralistic output. ICT in-

tegrates the advantages of CNN and Transformer, utilizing the global structure understanding capability and pluralistic support of Transformer and the local texture refinement of CNN to accomplish high-fidelity pluralistic image completion, showing a powerful performance. Thus, we utilize our established dataset of artistic face images to train ICT, thereby filling the gap in the artistic face image completion task. To get better results, we attempt to replace l1 loss with smooth l1 loss and add a TV loss to make the results continuous.

To summarize, the key contributions of this paper are: 1) A transformer-based artistic image inpainting algorithm; 2) A high-quality artistic image dataset that can be used for training; 3) Two new loss functions were tried: smooth l1 loss, TV loss.

2 Related Work

2.1 Style Transfer

Style transfer is the transformation of a content image into another artistic style. It has been developed for more than 20 years. Early style transfer did not use CNN. Image Analogies (Hertzmann et al. 2001) learned the mapping between source images and style images through supervised learning. Others combine image processing-related filters to render content images (Tomasi and Manduchi 1998), the algorithms are usually simple but limited in the number of styles. Since Neural Style Transfer (Gatys, Ecker, and Bethge 2016) was proposed in 2016, style transfer technology has undergone tremendous changes. More and more people use deep learning technology for style transfer. This model uses CNN to extract features from images and build content, and match features between content and style. In 2017, both Microsoft Research and Google solved the limitation that the model can only train one style for one image at a time, they implemented a network to train N styles (Dumoulin, Shlens, and Kudlur 2016; Chen et al. 2017).

2.2 Image Inpainting

Image inpainting aims to fill in missing areas of images, which has been a long-standing challenge in computer vision areas. As early as 20 years ago, Efros and Freeman (Efros and Freeman 2001) proposed an image quilting method, which synthesizes a new image by stitching together small patches of existing images. Then, PatchMatch (Barnes et al. 2009) finds approximate patches in the image and paste them into the missing area. These traditional image inpainting methods require the inclusion of structures or patches in the input image that are similar to the missing parts, which is impractical in many cases. In recent years, thanks to the excellent performance of CNN and GAN, solutions based on deep learning have begun to dominate the field of image inpainting. Pathak et al. (Pathak et al. 2016) propose a feature learning algorithm driven by context-based pixel prediction, which is capable of giving reasonable results for semantic hole-filling. Yu et al. (Yu et al. 2018b) design a new contextual attention module to capture the correlations at distant spatial locations. Li et al. Liu et al. (Yang, Qi, and Shi 2020) propose a mutual encoder-decoder network that can simultaneously learn features structure and texture corresponding to different layers. Wan et al. (Wan et al. 2021) propose to combine the advantages of transformers and CNN to improve the image fidelity and the diversity of results.

2.3 Transformer

Transformer (Vaswani et al. 2017) was first utilized for machine translation, and it has shown outstanding performance in natural language processing (NLP) through the attention mechanism. With Transformer reaching the mainstream model in NLP, more and more work (Carion et al. 2020; Dosovitskiy et al. 2020; Bao, Dong, and Wei 2021; Liu et al. 2021) is trying to apply Transformer to the field of computer vision. For example, DERT (Carion et al. 2020) introduces Transformer to do the target detection task. ViT (Dosovitskiy et al. 2020) employs the standard Transformer and processes images into a form similar to token sequences in NLP to solve the image recognition problem. BEiT (Bao, Dong, and Wei 2021) utilizes masked image patches to pre-train the visual transformer. Swin Transformer (Liu et al. 2021) employs a hierarchical structure to extract visual features at various levels, making it more appropriate for tasks such as segmentation and detection. In the area of image completion, ICT (Wan et al. 2021) achieves diverse and high-fidelity results by Transformer to recover the structure and coarse texture of missing parts, which solves the problem of the weak performance of CNN in understanding global structure or supporting diverse completion. In our work, we similarly apply Transformer to accomplish the completion of art-style face images, which was not addressed in previous work.

3 Method

In order to make up for the shortcomings of the current image inpainting models that cannot repair artistic images, we propose a transformer-based artistic image inpainting algorithm. As shown in Figure 1, our method consists of two stages. First, we exploit the existing style transfer method (Kolkin, Salavon, and Shakhnarovich 2019) to stylize the face images (Karras, Laine, and Aila 2019) into artistic images and use the stylized results as the dataset for our image repair phase. Then, we utilize this dataset to retrain an image repair model which is capable of repairing artistic images. The details of these two stages are described in section 3.1 and section 3.2, respectively.

3.1 Stylized Datasets

In the style migration phase, we need two input images, one is the content image I_C and the other is the style image I_S. We follow (Kolkin, Salavon, and Shakhnarovich 2019) and use the gradient descent variant RMSprop to minimize the loss function of the style transfer part, which is defined as:

$$L_{stage1}(X, I_C, I_S) = \frac{\alpha l_C + l_m + l_r + \frac{1}{\alpha} l_p}{2 + \alpha + \frac{1}{\alpha}} \quad (1)$$

where X is the stylized image, l_C is the content loss and $l_m + l_r + \frac{1}{\alpha} l_p$ is the style term. In particular, the hyperparameter α represents the weight of content loss.

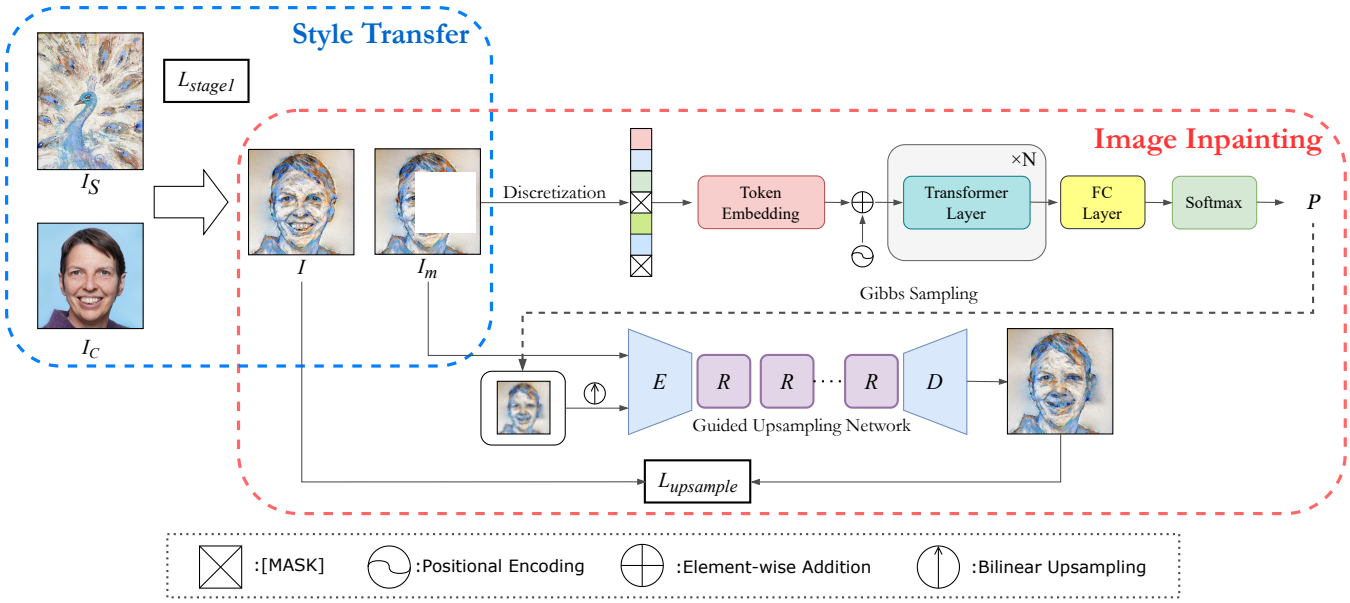

Figure 1: Pipeline Overview. Our framework consists of two modules: 1) **Style Transfer.** We use the style transfer algorithm to generate artistic face images. 2) **Image Inpainting.** This module consists of two networks, the top one is a bi-directional transformer, which produces the probability distribution of missing regions, and the bottom one is CNN, which upsamples the appearance prior and then generates the high-resolution images. **E**: Encoder, **D**: Decoder, **R**: Residual block, **P**: the probability distribution of missing regions.

Specifically, inspired by self-similarity, l_C is defined as:

$$L_C(X, I_C) = \frac{1}{n} \sum_{i,j} \left| \frac{D^X_{i,j}}{\sum_i D^X_{i,j}} - \frac{D^C_{i,j}}{\sum_j D^{I_C}_{i,j}} \right| \quad (2)$$

where $D^X_{i,j}$ is the pairwise cosine distance matrix of all hypercolumn feature vectors extracted from X and the definition of $D^C_{i,j}$ is the same.

In the style term, l_r is the Relaxed Earth Mover's Distance between X and I_S. And because l_r only considers the spatial relationship between vectors and ignores the impression of vector length, it is important to introduce l_m to reduce the artifacts caused by l_r. l_p, on the other hand, ensures that the stylized image and the stylized image are consistent in color.

What's more, unlike the general style transfer algorithm, (Kolkin, Salavon, and Shakhnarovich 2019) uses hypercolumns to extract and characterize the features of the image. We use the set of feature maps extracted from the 10 sub-layers of the VGG16 network, which has a total of 2179 feature maps. For an input image, we randomly select 1024 sampling points and use a bilinear interpolation algorithm to approximate the positions corresponding to these 1024 sampling points on the 2179 feature maps. Then, we arrange the points at each position into a vector of length 2179 as the hypercolumn of the corresponding sampled points. In the initialization stage, to speed up the convergence, we initialize the output image with the Laplace Pyramid of I_C instead of directly using I_C. Using the extracted features, we optimize the output image at increasing resolutions to finally obtain X.

We utilized several stylized images to stylize real face images, but most of them did not give good results. To ensure the quality of the training set, we only selected the results with the best results, as shown in Figure 2. Finally, we obtained a total of 900 artistic face images and divided the training set, validation set, and test set according to the ratio of 8:1:1.

Figure 2: The first column is the reference style image. And in the second to fourth columns, the first row is the content image and the second row is the stylized artistic face image.

3.2 Image Inpainting

In the image inpainting phase, we follow ICT (Wan et al. 2021) and use the combination of Transformer and convolution to accomplish image inpainting in order to overcome the problems that convolution does not understand global information well and does not support diverse outputs. Therefore,

our image inpainting is in two stages, appearance prior reconstruction and low-resolution upsampling. Given an input image $I_m \in \mathbb{R}^{H\times W\times 3}$ with missing regions, we first utilize Transformer to sample a low-resolution inpainting result, called the appearance prior X. Subsequently, the upsampling CNN model is leveraged to obtain high-fidelity inpainting results $I_{pred} \in \mathbb{R}^{H\times W\times 3}$ guided by the appearance prior and the input image. The details of these two stages will be described in sections 3.2.1 and 3.2.2, respectively. The overall pipeline can be found in Figure 1.

3.2.1 Appearance Prior Reconstruction

Recently, Transformer has demonstrated powerful capability in modeling long-term relationships and generating diverse results, but its computational complexity is quadratic in the input length. So we employ Transformer only to recover complex coherent structures and some coarse textures, called appearance prior reconstruction, which is represented by a low-resolution image. Similarly, the RGB space of pixels is also large and we followed ICT for a discretized representation, using an extra visual vocabulary. To this end, the missing regions are expressed in terms of a special learnable token [MASK] similar to the Masking Language Model (MLM) in BERT (Devlin et al. 2018).

For discrete sequences, we project them to a d-dimensional feature vector, adding the learnable positional embedding, with the final $E \in \mathbb{R}^{L\times d}$ as the input to the Transformer. For the network architecture, we only use the Transformer's decoder, which consists mainly of 30 self-attention layers. To capture the full available information, we utilize bi-directional attention to make each token attend to the full positions, thus achieving consistency between the generated content and the regions that are unmasked. The output of the final layer of the Transformer is further projected as a pixel-by-pixel distribution with a visual vocabulary using a fully connected layer and softmax. The MLM combined with bidirectional attention ensures that the Transformer model is capable of capturing the entire contextual information to predict the probability distribution of the missing regions.

Therefore, we take advantage of the powerful representation capabilities of the Transformer to reconstruct the appearance prior which contains global structure and coarse texture. Moreover, due to the small amount of data in our dataset, we initialize the network with the same weights as ICT pre-trained on the FFHQ dataset in order to better utilize the semantic information of faces, and later fine-tune it on our data. This way, we can use our small amount of data to get better results more quickly. After obtaining the distribution generated by the Transformer, we use Gibbs sampling to iteratively sample tokens at different locations.

3.2.2 Low-resolution Upsampling

After obtaining the low-dimensional appearance prior, we reshape X as $I_t \in \mathbb{R}^{\sqrt{L}\times\sqrt{L}\times 3}$ for the subsequent processing. Then, we need to convert the I_t to the original resolution $H \times W \times 3$ and ensure the boundary consistency of the inpainting image. After obtaining the low-dimensional appearance prior, we reshape X as It for the subsequent processing. Then, we need to convert the It to the original resolution HxWx3 and ensure the boundary consistency of the inpainting image. CNNs have demonstrated the ability to learn rich texture patterns. Therefore, we employ an extra upsampling network that is capable of rendering high-fidelity details using the appearance prior and the input masked image. In this way, we can utilize the CNN network to enhance the local texture details with a coarse prior to obtain the final inpainting image.

Unlike ICT, we optimize this guided upsampling network by minimizing the smooth L1 loss between the predicted image I_{pred} and the corresponding ground truth I. This makes the training more robust and this loss can be expressed as:

$$L_{smoothL1} = \begin{cases} |I_{pred} - I| - 0.5, |I_{pred} - I| > 1 \\ 0.5\,|I_{pred} - I|_2\,, |I_{pred} - I| < 1 \end{cases} \tag{3}$$

Furthermore, in order to ensure the consistency of the boundaries of the inpainting image, we additionally employ a Total Variation loss (TV loss) for minimizing the differences of adjacent pixels and ensuring the smoothness of the complete image, denoted as:

$$L_{TV} = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H} |I_{pred,i+1,j} - I_{pred,i,j}| \\ |I_{pred,i,j+1} - I_{pred,i,j}| \tag{4}$$

Similar to ICT, to generate more realistic details, we similarly employ adversarial loss in the training process, denoted as follows:

$$L_{adv} = \mathbb{E}[log(1 - D(I_{pred}))] + \mathbb{E}[logD(I)] \tag{5}$$

where D is the discriminator. We simultaneously train the upsampling network and the discriminator.

The final optimization was solved by the followings.

$$L_{upsample} = \alpha_1 L_{smoothL1} + \alpha_2 L_{adv} + \alpha_3 L_{TV} \tag{6}$$

The loss weights were set to $\alpha_1 = 1.0$, $\alpha_2 = 0.1$, and $\alpha_3 = 1.0$ in all experiments.

4 Experiments

4.1 Implementation Details

The image inpainting experiments were conducted on our self-made artistic face images dataset. The production process of the dataset and the split mode of training, testing, and validation are as mentioned in section 3.1. Randomly generated masks are used for training and evaluation, and the resolution of all the images and masks is 512 × 512 pixels.

The implementation of the model is based on PyTorch. The transformer is optimized with AdamW (Loshchilov and Hutter 2017) optimizer, where $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We first warm up the learning rate from 0 to 3e-4 for initial training, then decay it to 0 in the rest epochs. The guided upsampling network is trained with Adam (Kingma and Ba 2014) optimizer with the learning rate of 1e-4, $\beta_1 = 0.0$, and $\beta_2 = 0.9$. It takes about 24 hours to train our model on an RTX 3090 GPU.

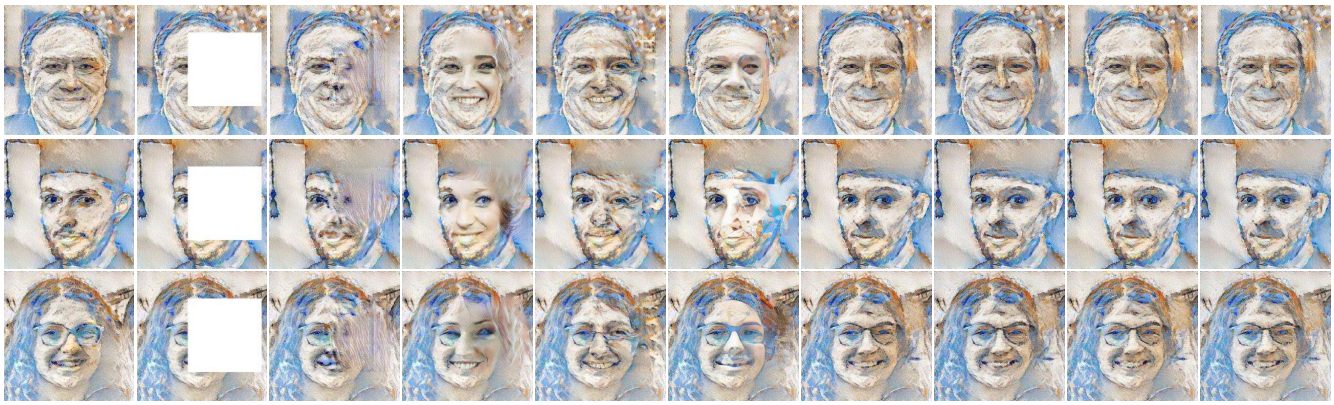

Figure 3: Qualitative Comparison with Deepfill v2, PICNet and ICT methods.

4.2 Method Comparison

4.2.1 Quantitative Comparison

We numerically compare our method with other baselines in Table 1. We use the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to quantify the similarity of the completed image to the ground truth. We found that PICNet pretrain achieved the best results on both SSIM and PSNR, but we can find in Figure 3 that the effect of PICNet pretrain on image perception is not good.

Method	SSIM	PSNR
Deepfill v2	0.7084	19.1971
PICNet pretrain	**0.7386**	**20.0410**
PICNet finetune	0.7295	20.0401
ICT pretrain	0.7267	19.4598
ICT finetune	0.7138	19.8218
ICT + Smooth L1 Loss	0.7087	19.5278
ICT + TV Loss	0.7129	19.9055
Ours	0.7129	19.8225

Table 1: **Quantitative results on 90 images.** Ours: ICT with Smooth L1 Loss and TV Loss

4.2.2 Qualitative Comparison

Although PICNet pretrain achieves the highest score on quantitative metrics, its real-world style image infilling does not meet the requirements of stylized completion. ICT pretrain also has similar problems, but the effect is relatively good. Ours works better on eyes than other methods.

In other methods, it is difficult to distinguish the sclera and pupil of the eye when the eye is completed, and Ours can better distinguish the sclera and pupil.

4.3 Ablation Study

We perform all ablation studies for the proposed image complementation algorithm to investigate the efficiency of the network designs. Our network utilizes smooth L1 loss instead of L1 loss in the upsampling network of ICT and also adds TV loss for optimization. The qualitative and quantitative results are presented in Figure 3 and Table 1 respectively.

First, it can be observed that ICT pretrain method has higher SSIM and lower PSNR than our method. And in the quantitative result, ICT pretrain produces the basic compositions of missing parts without the texture features of artistic images, which are more similar to realistic photographs. Then, replacing the L1 loss with smooth L1 loss in ICT shows decreases in both quantitative metrics. Next, ICT with TV loss shows a slight decrease of 1.26% in SSIM and a greater improvement of 4.22% in PSNR, which verifies the advantage of using TV loss. As shown in Figure 3, there is no significant difference between the methods trained in the artistic dataset, and each method inpaints the images well.

5 Conclusion and Discussion

In this paper, we combine a style transfer algorithm with an image inpainting algorithm, which yields a new framework capable of performing the artistic face image inpainting task. First, we infer the face images with style images by style migration algorithm to generate stylized datasets, and then use the datasets as the training set of image inpainting algorithm to train a model that can patch artistic face images. For the image inpainting algorithm, we redesign the network loss function based on ICT, replace L1 loss using smooth L1 loss on top of the image patching algorithm, and further add TV loss. Although the metrics of our method are not the highest in quantitative experiments, our results are visually better in terms of qualitative results. What's more, we conjecture that the lack of significant improvement in our results is due to the fact that the transformer model is too large and our training set is too small (720 images in total), which leads to a relatively serious overfitting of our model. In the future, we may be able to obtain better results by increasing the dataset.

References

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

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

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2017. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1897–1906.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dumoulin, V.; Shlens, J.; and Kudlur, M. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

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

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Hertzmann, A.; Jacobs, C. E.; Oliver, N.; Curless, B.; and Salesin, D. H. 2001. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 327–340.

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

Jo, Y.; and Park, J. 2019. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1745–1753.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolkin, N.; Salavon, J.; and Shakhnarovich, G. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10051–10060.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Tomasi, C.; and Manduchi, R. 1998. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, 839–846. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; and Wen, F. 2020. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2747–2757.

Wan, Z.; Zhang, J.; Chen, D.; and Liao, J. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4692–4701.

Yang, J.; Qi, Z.; and Shi, Y. 2020. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12605–12612.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018a. Generative Image Inpainting With Contextual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Zeng, Y.; Fu, J.; Chao, H.; and Guo, B. 2022. Aggregated Contextual Transformations for High-Resolution Image Inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 1–1.

Zheng, C.; Cham, T.-J.; Cai, J.; and Phung, D. 2022. Bridging Global Context Interactions for High-Fidelity Image Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11512–11522.