SLTFILL: SPATIAL AND LIGHT TRANSFORMER FOR MULTI-REFERENCE IMAGE INPAINTING

Yuliang Fan, Yue Zhou, Zonghao Yang, Zhenyu Tong*

Department of Automation, School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Image inpainting aims to restore the missing regions of corrupted images plausibly. Recent techniques of image inpainting handle it as a pure generation problem and utilize large-scale training data to learn semantic and texture information to fill the missing area. However, generative networks frequently distort local structures or introduce visible artifacts while filling large missing regions. To address this issue, we propose SLTFill, an image inpainting method that uses a set of reference images which share scene contents with the corrupted image to guide the synthesis of missing regions. A key challenge addressed by our approach is aligning the transformations of the reference images with the context of the corrupted image and making sure they are consistent in spatial structure and light space. Qualitative and quantitative experiments demonstrate the state-of-the-art performance of our approach on the image inpainting task.

Index Terms— Image Inpainting, Image Harmonization, Deep Learning, Transformer, Attention

1. INTRODUCTION

Image inpainting aims to restore specific parts of the incomplete images. The regions to be filled are defined as hole regions, and they may contain undesired foreground objects, distracting components, or picture corruption. Existing inpainting approaches can be divided into two categories. Traditional image inpainting methods are mainly diffusion-based [1] or patch-based [2]. These methods seek to fill the holes by feature propagation from known image backgrounds to corrupted regions. These traditional methods work well in simple circumstances, such as filling the hole with a uniform textured background. However, because of a lack of high-level semantic comprehension, they are unable to address challenging scenarios involving complicated semantic layouts or nonrepetitive structures[3]. Deep methods [4]–[8] adopt deep convolutional neural networks (CNNs)[9], especially generative adversarial networks (GANs) [10] in image inpainting. These deep methods learn the contextual semantic information representation of known regions and hallucinate the missing patches. However, learning powerful representational spaces from incomplete images is tough for the encoder. As a result, these deep methods will provide visible color contrasts and artifacts, particularly in boundary areas.

These difficulties can be handled if there are reference images of the same scene providing desirable image content that can be transferred to the hole. This task is referred to as reference-guided image inpainting in the literature [11]. Generally speaking, there are two methods for obtaining reference pictures. One is that the photographer can either shoot from the same perspective when the object moves, or switch the camera to a different viewpoint to expose the scene behind the occluder. The other is collecting source photos from the Internet [12] that share the same scenario with the inpainted target image.

Despite the fact that the reference source picture makes inpainting easier, reference-guided inpainting is quite challenging for a variety of reasons. First, the exposure, white balance, or lighting environment of the reference images may vary greatly due to differences in lighting conditions, camera focus, point of view, and photography settings. Current existing purely warping-based solutions are unable to address the complicated color harmonization challenge. Besides, the multi-reference images represent the different transformations of the corrupted target image in 3D space, which induce large parallax in the hole-filling task.

To properly suppress parallax and color discrepancy in boundaries, we propose a unified three-stage framework named SLTFill, which explicitly disentangle reference-guided image inpainting into multi-reference alignment, spatial transformation, and illumination adjustment. Our key insight is that we address the issue of parallax by a spatial attention fusion module and tackle the challenge of illumination harmonization by a Light Harmonization Transformer. The core idea of the spatial attention fusion module is to calculate the similarity between the synthesized content and the known region, merging all the transformed reference image proposals by learning a set of attention fusion maps. Observing the effectiveness of Transformer [13] in modeling long-term correlations between distant contextual information and the hole regions, we presented a novel light contextual attention layer in the illumination adjustment stage, which used the light features of known patches as a guide to process the generated hole patches. This work is the first attempt to adapt the study of vision Transformer [14] for reference-guided image inpainting, diving deeper into Transformer for image harmonization and image alignment as well.

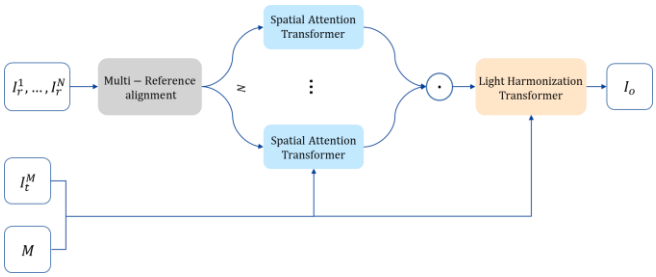


Fig. 1. The system pipeline of SLTFill.

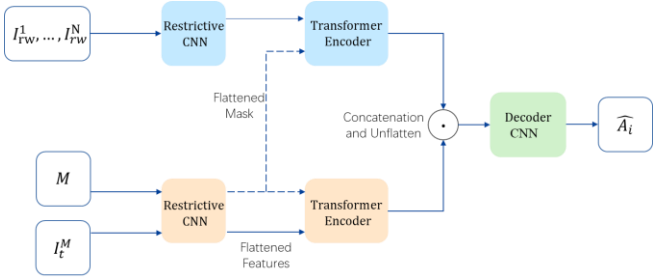


Fig. 2. The architecture of Spatial Attention Transformer.

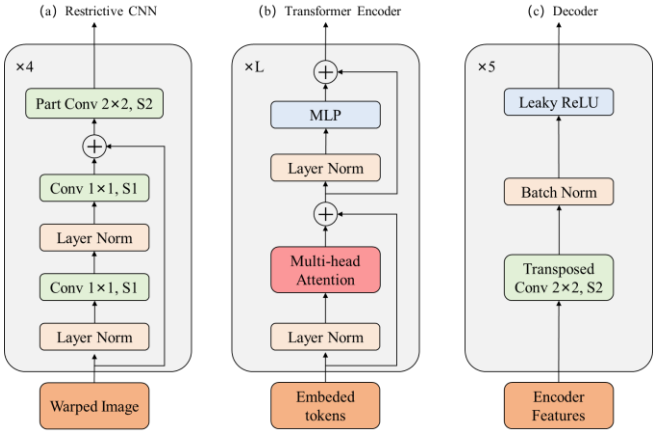


Fig. 3. The detailed design of essential blocks in Spatial Attention Transformer (SAT).

Qualitative and quantitative experiments on multiple datasets are utilized to evaluate our proposed method. The experimental results demonstrate that our proposed method outperforms state-of-the-art methods in generating high-quality inpainting results.

2. METHOD

Our image inpainting technology is embedded in a large system that follows three key steps. In Section 2.1, we present an overview of the multi-reference image inpainting pipeline. The second and third stage, which we address in Sections 2.3 and 2.4, is our key novelty.

2.1. System Overview

An overview of our reference-guided image inpainting framework is shown in Figure 1. To generate plausible images inpainting output, the pipeline employs three key modules sequentially. The first module transforms the *N* reference images by estimating global homography to obtain *N* warped references that are roughly aligned with the target image. The first stage is referred to as the Multi-Reference alignment. The second module aims to utilize the spatial attention mechanism to fuse multiple roughly aligned reference images transformed in the prior stage. We implement the second step through a spatial attention transformer(SAT). The third stage performs illumination adjustment on the finely aligned hole pixels to ensure the light consistency between the content of the hole region and the known region. This subtask is performed with a light harmonization transformer(LHT). We describe these three stages in detail below.

2.2. Multi-Reference Alignment

Our pipeline takes in an incomplete target image $\mathbf{I_t} \in \mathbb{R}^{\mathbf{W}\times\mathbf{H}\times\mathbf{3}}$, reference images $\mathbf{I_r}^{\mathbf{1,2,\ldots,N}} \in \mathbb{R}^{\mathbf{W_r}\times\mathbf{H_r}\times\mathbf{3}}$, and binary hole mask $\mathbf{M} \in \mathbb{R}^{\mathbf{W}\times\mathbf{H}\times\mathbf{1}}$. $\mathbf{M}$ represents the hole area with a value of 1, and a value of 0 elsewhere. Then the masked target image is symbolized by $\mathbf{I_t^M} = \mathbf{I_t}\odot(\mathbf{1}-\mathbf{M})$.

We initially transform the reference images globally to align them with the masked incomplete image coarsely in this stage. Firstly, we extract SIFT [15] features from $\mathbf{I_t}$ and $\mathbf{I_r}^{\mathbf{1,2,\ldots,N}}$, then feed all the extracted feature points and their descriptors into the approximate nearest neighbor matching algorithm [16]to constructs the correspondences between $\mathbf{I_t^M}$ and $\mathbf{I_r}^{\mathbf{i}}$. The inliers constitute a point set $\mathbf{P_t^i}$ in $\mathbf{I_t^M}$ and its matching point set $\mathbf{P_g^i}$ in $\mathbf{I_g}^{\mathbf{i}}$. Further, the module adopts PROSAC algorithm [17]to exclude mistake matching points and estimate global homography symbolized by $\mathbf{H^i}, \mathbf{i} \in [\mathbf{1}, \mathbf{N}]$. Finally, we use the estimated homography matrix H^i to warp the reference images, yielding a set of warped reference photos $\{\mathbf{I_{rw}}^{\mathbf{i}}\}$, where $\mathbf{I_{rw}}^{\mathbf{i}} \in \mathbb{R}^{\mathbf{W}\times\mathbf{H}\times\mathbf{3}}$, $\mathbf{i} \in [\mathbf{1}, \mathbf{N}]$. We set $\mathbf{N} = 4$ in our study.

2.3. Spatial Attention Transformer

The backbone of the Spatial Attention Transformer is two-branch, with features generated separately by warped reference pictures and separately by target images, then concatenated and sent to the decoder. Each branch consists of a tensor-based token generator block and a transformer-encoder block, where tokens are fed. Figure 2 depicts the overall network architecture. The traditional convolutional network is employed to obtain embeddings used as tokens. However, traditional CNN encoders have a broad receptive field. That is, each token carries information about a large portion of the image, and all tokens have some correlated information, implying that the tokens are interdependent. As a result, the transformer may be unable to extract deep

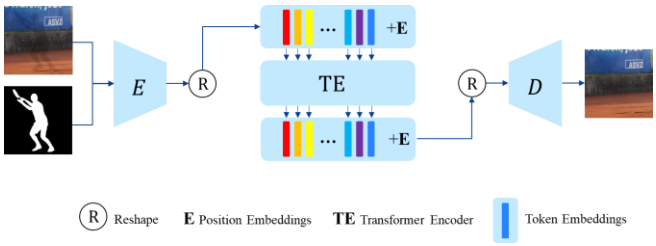


Fig. 4. Illustration of Light Harmonization Transformer (LHT). The architecture contains a convolutional encoder-decoder involving a Transformer encoder inside.

dependencies for inpainting. In this paper, we took the tokenizer implementation from [18], which they named Restrictive CNN. The diagram of Restrictive CNN is shown in Figure 3a.

Given the warped reference image set $\mathbf{I_{rw}}^{\mathbf{i}}$, corrupted target image $\mathbf{I_t}$ and its corresponding mask M, the output of Restrictive CNN can be denoted as:

$$\mathrm{T_t}, \mathrm{M}' = R(\mathrm{I_t}, \mathrm{M}) \quad (1)$$

$$\mathrm{T_{rw}}^{\mathrm{i}} = R\left(\mathrm{I_{rw}}^{\mathrm{i}}\right) \quad (2)$$

where T_t and $T_{rw}{}^{i}$ are generated tokens from corrupted target image and warped reference images set respectively, and M' is the modified mask.

Along with the flattened tokens, the final tokenizer mask is sent into the Transformer Encoder based on vanilla Transformer implementation in [14]. It comprises layer normalization (LN), multi-head self-attention (MSA), and Multilayer Perceptron (MLP), as shown in Figure 3b.

The outputs of Transformer Encoders are concatenated, unflattened, and transferred to the CNN Decoder, as shown in Figure 3c. It's a progressive upsampling decoder, with several stacked up-sampling layers consisting of transposed convolution with kernel 2x2, batch normalization, and leaky ReLU for each layer. The spatial attention maps $\mathbf{A^i} \in \mathbb{R}^{\mathbf{W\times H\times 3}}$ is given by:

$$\mathrm{A^i} = D\left(Unflatten\left(TE(\mathrm{T_t}, \mathrm{M}'), TE\left(\mathrm{T_{rw}}^{\mathrm{i}}, \mathrm{M}'\right)\right)\right) \quad (3)$$

where D is the Decoder, TE is the Transformer Encoder. The pixel-wise weights of spatial attention map $\mathbf{A^i}$ are learned through the Decoder with softmax. The output of Spatial Attention Transformer $\widehat{\mathrm{I_r}}$ is obtained by merging different parts of the warped multi-reference guide images as shown below,

$$\widehat{\mathrm{I_r}} = \sum \mathrm{A^i} \odot \mathrm{I_{rw}}^{\mathrm{i}} \quad (4)$$

2.4. Light Harmonization Transformer

We employ Transformer in a basic convolutional encoder-decoder architecture to harmonize the hole region with the known region in the corrupted image, as shown in Figure 4.

The CNN encoder E and decoder D are responsible for compressing the input image to light token embeddings as Transformer input and reconstructing the Transformer output back to the harmonized image, respectively. Formally, given a spatial structure consistent image $\widehat{\mathbf{I_r}}$ and a hole region mask $\mathbf{M}$ which indicates the inharmonious region as input, the goal of Light Harmonization Transformer is to produce a light harmonized image $\mathbf{I_o}$ as output. Specifically, CNN encoder E(·) generates a lower-resolution feature map $\mathbf{F} \in \mathbb{R}^{\mathbf{w\times h\times c}}$, where $\mathrm{w} = \frac{W}{4}$, $\mathrm{h} = \frac{H}{4}$, and $\mathrm{c} = 256$. Then it is reshaped to sequence $\mathbf{F'} \in \mathbb{R}^{\mathbf{hw\times c}}$, with pixels (corresponding to image patches) as Transformer Encoder TE input tokens and channel aggregation of each pixel as light latent token embedding, added with position embeddings $\mathbf{E}$ using the sinusoidal version of vanilla Transformer. Moreover, the output sequence of TRE is inversely shaped into a feature map with the same size as F, and fed to CNN decoder D(·). The harmonized result I_o can be symbolized as follows:

$$I_o = D\left(Unflatten\left(TE\left[Flatten\left(E\left(\widehat{\mathrm{I_r}}, M\right)\right) + \mathbf{E}\right]\right)\right) \quad (5)$$

3. EXPERIMENT

3.1. Datasets

To demonstrate the effectiveness of the Spatial Attention Transformer (SAT), we evaluate our method on DAVIS [19] and YouTube-VOS [20] datasets. DAVIS dataset contains 150 high-quality video image sequences annotated with the pixel-wise foreground object masks. There are 4,453 videos in YouTube-VOS [20], divided into 3,471 for training, 474 for validation, and 508 for testing. In these datasets, 5 frames are uniformly sampled from each video sequence to form a group of image set. One of the 5 images is used as the target image, while the other 4 images are used as the references.

For training the Light Harmonization Transformer (LHT), although DAVIS contains sufficient samples with real multi-view images and different exposures across multi-reference image sets, it lacks image sets with large light inconsistency. Therefore, we synthesized composite light-different images from the Adobe5K[21] dataset for training. Adobe5K contains 5000 images, and for each image it provides five additional expert-retouched images to form 5000 sets in total. Thus We can merge the ordinary image with expert-retouched images, and synthesized 25000 composite color-different image sets, which includes synthesized composite images, binary masks of composite images, and corresponding real images.

3.2. Implementations

We pre-trained the Light Harmonization Transformer module using the mixed dataset in advance for 500 epochs, and then froze it. After freezing the network weights, the whole pipeline was trained end-to-end for 1000 more epochs. We



(a) Target Image (b) Reference Images (c) OPN (d) DeepFill V2 (d) SLTFill(Ours)

Fig. 5. Comparison of qualitative results over DAVIS [19] with exiting models.

resize target corrupted images as 512×512 for training and arbitrary size for inference. We employed the Adam optimizer [22] with betas (0.9, 0.999). It is worth noting that a single $\mathcal{L}_1$ loss gives the best reconstruction quality, due to the use of multi-reference images:

$$\mathcal{L}_1 = \mathbb{E}_{(I_o, I_{gt})} \left[\left\| I_o - I_{gt} \right\|_1 \right] \tag{6}$$

3.3. Ablation Study

This ablation study was conducted on the validation set of DAVIS datasets. For each alternative sequence, we re-trained the model for fairness. The sequence between Spatial Attention Transformer and Light Harmonization Transformer makes significant effect, as seen in Table 1. We found that adjusting the spatial structure first helped the training converge faster, since the spatial attention fusion module did not use light latent data. Table 1 also showed that both Spatial Attention Transformer and Light Harmonization Transformer were essential modules.

Table 1. Ablation study on SLTFill. S: Spatial L: Light.

Order	PSNR ↑	SSIM ↑	FID ↓
S→L	**29.55**	**0.9047**	**5.97**
L→S	29.39	0.8849	6.10
Only S	29.42	0.8732	6.22
Only L	28.54	0.8587	7.25

3.4. Comparison and Discussion

We compare our proposed method with state-of-the-art video inpainting and single image inpainting methods as follows,

OPN[11]: The Onion-Peel Network address video inpainting with an asymmetric attention block. It fills the holes with muti-references.

DeepFill V2[8]: DeepFill V2 is a state-of-the-art image inpainting method with gated convolution and SN-Patch GAN. We used the official pre-trained model from the authors.

For evaluation metrics, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Fréchet Inception Distance (FID) are utilized. These data show that SLTFill can achieve better qualitative (Figure 5) and quantitative performance(Table 2) in all aspects of the indicators. It can be seen that DeepFill tend to synthesise blurred final results and OPN performed worse than SLTFill in light harmonization. It demonstrates that spatial attention fusion module and light harmonization module play an important role in reference-guided image inpainting.

Table 2. Quantitative comparison on validation images of DAVIS and YouTube-VOS.

Dataset	Metric	OPN	DeepFill V2	SLTFill(Ours)
DAVIS	PSNR ↑	28.34	27.57	**29.55**
	SSIM ↑	0.8892	0.8730	**0.9047**
	FID ↓	6.28	7.73	**5.97**
YouTube-VOS	PSNR ↑	27.45	27.12	**28.63**
	SSIM ↑	0.8814	0.8163	**0.9014**
	FID ↓	7.01	7.24	**6.37**

4. CONCLUSION

In this paper, we propose a novel way of multi-reference guided image inpainting with Transformer, aiming to eliminate the structure and illumination inharmony by leveraging Transformer's modeling ability of long-range context dependencies. Spatial Attention Transformer is designed to build the connection between multi-reference images and target images. Light Harmonization Transformer framework is proposed to reason about illumination. Experiments on DAVIS and YouTube-VOS datasets show the superiority of our proposed pipeline in generating more plausible and coherent inpainting results, especially when the hole contains complicated scene.

5. REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, pp. 417–424, 2000.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 24–33,2009.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018.

[4] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[5] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 637–647, 2017.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018.

[7] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 331–340, 2018.

[8] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019.

[9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[11] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. "Onion-peel networks for deep video completion," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4403–4412, 2019.

[12] Oliver Whyte, Josef Sivic, and Andrew Zisserman. "Get out of my picture! internet-based inpainting," in *Proceedings of the 20th British Machine Vision Conference*, London, 2009.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need," in *Advances in Neural Information Processing System*, pp. 5998– 6008, 2017

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929, 2020.*

[15] David G Lowe. "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, 60(2):91–110, 2004.

[16] A. Andoni and P. Indyk . "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions," in *Proceedings of the IEEE Symposium on Foundations of Computer Science*, 2006.

[17] O. Chum and J. Matas, "Matching with PROSAC – Progressive Sample Consensus," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[18] C. Zheng, T.-J. Cham, and J. Cai. "Tfill: Image completion via a transformer-based architecture," *arXiv preprint arXiv:2104.00845*, 2021.

[19] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[20] Ning Xu, Linjie Yang, Dingcheng Yue, Jianchao Yang, Yuchen Fan, Yuchen Liang, and Thomas Huang. "Youtube-vos: A large-scale video object segmentation benchmark," in *arXiv preprint arXiv:1809.03327*, 2018.

[21] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fr´ edo Durand. "Learning photographic global tonal adjustment with a database of input / output image pairs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[22] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.