DETAIL GENERATION AND FUSION NETWORKS FOR IMAGE INPAINTING

Wu Yang and Wuzhen Shi

College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China

ABSTRACT

Recent image inpainting methods based on end-to-end models have achieved great success with the help of generative adversarial training and structure generation. However, the generated conditional structure priors cannot support the model to reconstruct finer texture. This is mainly because these models lack good texture prior knowledge, which leads to their limited ability to generate finer texture. In this paper, we propose a novel detail generation and fusion network (DGFNet) to strengthen the generation of texture details for image inpainting, which includes a dual-stream texture generation network and a multi-scale difference perception fusion network. The dual-stream texture generation network can explicitly model the missing texture information and generate a texture map to compensate for the coarse result produced by the parallel network. Furthermore, to merge two different kinds of information effectively, a fusion network based on the differential perception fusion module (DPFM) is introduced for multi-scale perception fusion in feature level. Extensive qualitative and quantitative experiments on the benchmark dataset show that the proposed DGFNet achieves state-of-the-art performance.

***Index Terms*—** image inpainting, deep learning, detail generation, dual-stream network

1. INTRODUCTION

Image inpainting refers to fill the target regions with alternative contents deduced from known information such as context or external datasets. It has a widespread application in imaging and graphics, e.g. face editing, object removal [1]. And it can also be extended to other tasks such as image/video stitching, compositing, compression and image-based rendering [2].

The traditional image inpainting approaches [3, 4] can not generate new content for lack of high-level semantic understanding. The studies have shown that deep learning-based inpainting methods [5, 6, 7, 8] can more effectively generate new and realistic image content. Recently, most state-of-the-art methods take two-stage optimization where the network in first stage is developed to model the hallucinate structures of missing parts for emphasizing the consistency of the global structure [9, 10, 11, 12]. They attempt to use the generated structure to guide the repair of details. However, when repairing complex scenes or large holes, these methods are prone to produce content with ambiguous semantics and vague textures. High-frequency details are difficult to recover, and these models lack good texture prior knowledge, which leads to their limited ability to generate finer texture. What's more, only cascading two or more generators is sub-optimal for parameter optimization in these cases.

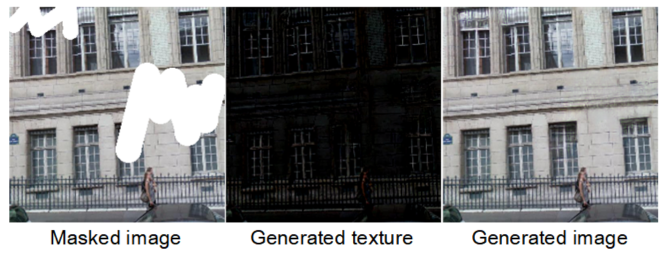

Fig. 1. A visual example of the generated texture and the final generated result by using the proposed DGFNet.

In this paper, we design a detail generation and fusion network to strengthen the detail repair which consists a dual-stream network and a fusion network. The dual-stream network is composed of a coarse branch for repairing coarse image and a texture branch for repairing details. The fusion network is embedded with multiple difference perception fusion modules (DPFM) for fusing different input information in multi-scale. Due to the dual-stream generation network as well as the specifically designed DPFM, our approach is able to achieve more visually convincing textures, as shown in Fig.1. Specifically, the dual-stream network take the masked image and the masked texture extracted from the masked image as input and output the coarse repaired image and complementary texture map. Then coarse result and the texture map are firstly encoded as high dimensional feature map via several layers of convolution respectively for performing fusion in feature level. After that, the DPFM adaptively generate gating weights to select the reasonable feature for fusing based on different input information. Besides, our method performs multi-scale fusion to better exploit different information. The main contributions of this work are as follows:

- We propose a detail generation and fusion network that emphasizes fine texture generation for image inpaint-

Corresponding author: Wuzhen Shi (Email:wzhshi@szu.edu.cn). This work was supported in part by the National Science Foundation of China under Grant 62101346, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011702 and in part by the Stable Support Plan for Shenzhen Higher Education Institutions under Grant 20200812104316001.

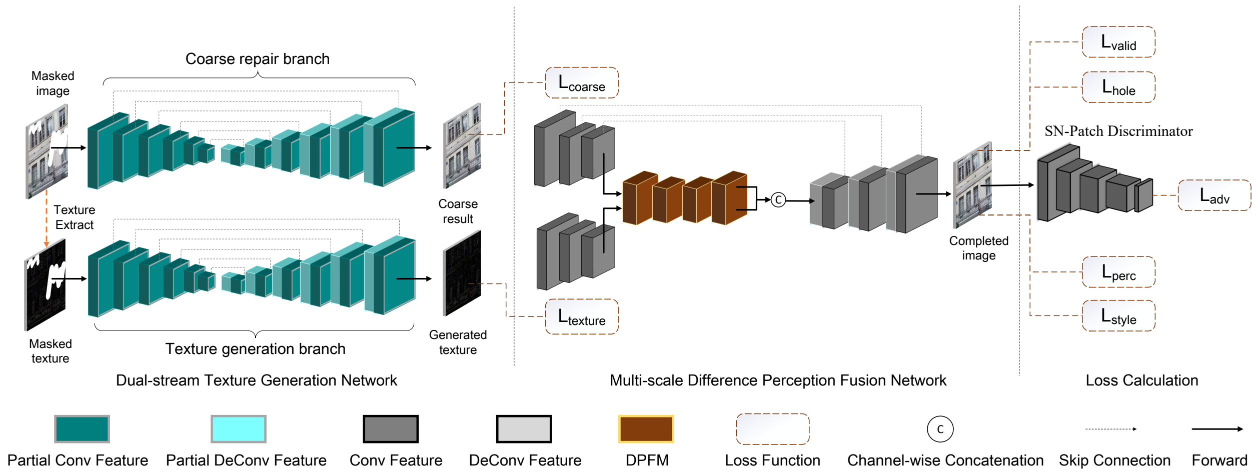

Fig. 2. The network structure of DGFNet.

ing. Extensive experiments on public datasets show that the proposed method achieves state-of-the-art performance both quantitatively and qualitatively.

- To the best of my knowledge, we are the first to propose using an independent texture generation branch to repair the image details, which explicitly model the hallucinate texture of missing regions. The repaired texture complementary to the coarse result produced by the parallel network can enhance the reconstruction of details
- We design the difference perception fusion module to merging the different information which can extract higher dimensional features from the input for fusing in multi-scale feature level.

2. RELATED WORK

The widespread use of deep learning has greatly improved the performance of image inpainting. Pathak et al. [13] introduce GANs to inpainting and Iizuka et al. [14] introduce local and global discriminators to improve local consistency and global consistency. Yu et al. [15] propose the SN-PatchGAN which distinguishes the authenticity of different image patches and use spectral normalization [16] to stabilize the training process. Furthermore, Liu et al. [6] introduce the Feature Patch Discriminator combining the advantages of the feature discriminator and patch discriminator. Liu et al. [17] propose the partial convolutional layer whose output is conditioned on only known parts. Later, Yu et al. [15] proposed a gated convolutional layer which providing a learnable dynamic feature selection mechanism.

Since some deep learning-based methods suffer from structural damage and texture blur, a number of structure prior guided approach are proposed to solve this issue. [11, 9, 10] first predict intermediate edge-preserved smooth images, edge maps and foreground contours respectively, and then provide conditional prior to guide the final result. [12, 18] also use the edge map to strengthen the structural characteristics of generated image. However, these methods cannot produce visually realistic details when the holes become large or the missing area involve multiple semantic objects for lack of prior with high frequency information. [19, 20] adopt a method of repairing the texture and structure separately and then fusing. But they all utilize the ground truth image to guide the generation of textures which is not accurate. Therefore, we extract the texture map from the image and model the texture generation separately.

3. PROPOSED METHOD

3.1. Overview of DGFNet

As shown in Fig.2, the whole model consists of three parts: the dual-stream texture generation network, the multi-scale difference perception fusion network and the Patch discriminator. The dual-stream network takes the masked image and the masked texture extracted from masked image as input, and simultaneously generates a coarse repaired image and a texture detail map. The multi-scale difference perception fusion network integrates the coarse repaired result and the generated texture in different feature levels to output the final repaired image. In the training process, the Patch discriminator with spectral normalization as in [15] is used for adversarial training which judges different locations and different semantics of input image.

3.2. Dual-stream Texture Generation Network

Traditional image processing method performs low-pass filtering on the image and then subtract the low frequency part from the original image to get the image with the high frequency information. Then the enhanced high frequency parts are added back to the original image to sharpen texture details. Inspired by this, we use RTV [21] to get the edge-preserved smooth image. And the texture parts can be extracted by subtracting the smooth image from the original image. As a result, the inputs of our dual-stream texture generation network are the masked image and the masked

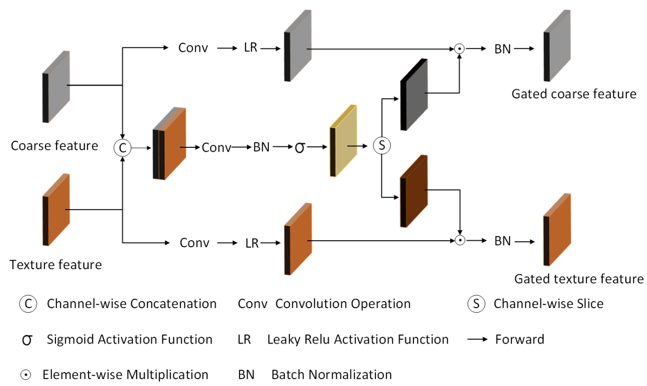

Fig. 3. The network structure of the difference perception fusion module.

texture obtained by RTV. The RFR network [22] is used as our coarse repair branch. The texture generation branch is a simple encoder-decoder structure.

3.3. Multi-scale Difference Perception Fusion Network

The main purpose of DPFM is to merge the coarse result and generated texture more effectively in high-dimensional feature level. It generates soft gated weights adaptively based on two complementary information and selects right features for fusing. To be specific, the coarse result and the generated texture are firstly coded as features F_c and F_t respectively by several convolution layers as shown in Fig.3. Then the DPFM produces soft gated weights W_c and W_t respectively, which are formulated as:

$$W_c, W_t = S(\sigma(BN(Conv(Concat(F_c, F_t))))) \tag{1}$$

where $Concat(\cdot)$ is channel-wise concatenation, $Conv(\cdot)$ is a convolution filtering operation, $BN(\cdot)$ is the batch normalization, $\sigma(\cdot)$ is the Sigmoid activation and $S(\cdot)$ is channel-wise slice. Then we use W_c and W_t to control two types of input features, which can be formulated as:

$$F_c^{'} = BN(W_c \circ LR(Conv(F_c))) \tag{2}$$

$$F_t^{'} = BN(W_t \circ LR(Conv(F_t))) \tag{3}$$

where LR is the LeakyRelu activation, $\circ$ denotes element-wise multiplication. There are four cascaded DPFMs in the fusion network. To perform multi-scale fusion, the strides of all convolution layers are set to 2. Finally, we fuse the feature map F_c and the feature map F_c to get the integrated feature F_{out} by channel-wise concatenation:

$$F_{out}^{'} = Concat(F_c^{'}, F_c^{'}) \tag{4}$$

Furthermore, we fuse the integrated feature and the high-dimensional feature of the coarse repair results with the skip connection in different scale.

3.4. Loss Functions

The whole model is trained with a mixed loss. Each sub-net branch has an independent loss function. We use the perceptual loss capturing high-level semantics and style loss ensuring overall style from a pre-trained and fixed VGG-16. Both of them are formalized as follows. ϕ_{pool_i} denotes feature maps from the i_{th} pooling layer in the pretrained VGG-16. The perceptual loss can be written as following:

$$\mathcal{L}_{perc} = \mathbb{E}\left[\sum_i \left|\phi_{pool_i}^{gt} - \phi_{pool_i}^{pred}\right|_1\right] \tag{5}$$

And the style loss is written as following:

$$\phi_{pool_i}^{style} = \phi_{pool_i}\phi_{pool_i}^T \tag{6}$$

$$\mathcal{L}_{style} = \mathbb{E}\left[\sum_i \left|\phi_{pool_i}^{style_{gt}} - \phi_{pool_i}^{style_{pred}}\right|_1\right] \tag{7}$$

The SN-Patch discriminator is used to enhance the detail repair ability of the generator. We use the hinge loss for adversarial training.The loss for discriminator D is:

$$\mathcal{L}_{adv_D} = \mathbb{E}\left[\mathrm{ReLU}(1 - D(I_{gt}))\right] + \mathbb{E}\left[\mathrm{ReLU}(1 + D(I_{com}))\right] \tag{8}$$

where I_{com}, I_{gt} denotes completed image and the ground truth respectively. The loss for generator is:

$$\mathcal{L}_{adv_G} = \mathbb{E}\left[\mathrm{ReLU}(1 - D(I_{com}))\right] \tag{9}$$

Besides, $\mathcal{L}_{hole}$ and $\mathcal{L}_{valid}$ are also used in our model which calculate L1 differences in the hole region and valid region respectively. In summary, the total loss is written as:

$$\begin{aligned}\mathcal{L}_{total} = {} & \lambda_{hole}\mathcal{L}_{hole} + \lambda_{valid}\mathcal{L}_{valid} + \lambda_{perc}\mathcal{L}_{perc} \\ & + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv_G}\mathcal{L}_{adv_G}\end{aligned} \tag{10}$$

4. RESULTS AND DISCUSSIONS

4.1. Experimental Settings

Our model was trained with the batch size of 6 on an NVIDIA A100 40G GPU. We used the Adam Optimizer to optimize our generator and discriminator. We first train our model with a learning rate of 5×10^{-4} And then we use 5×10^{-5} in the fine-tuning stage. For the hyper-parameters in loss functions, we set λ_{hole}, λ_{valid}, λ_{perc}, λ_{style} as 50, 50, 0.1, 180 respectively for texture branch. Similarly, for multi-scale difference perception fusion network, we set the same hyper-parameters and set the λ_{adv_G} as 0.1. For the coarse branch, we set λ_{hole}, λ_{valid}, λ_{perc}, λ_{style} as 6, 1, 0.1, 180. And we evaluated our model and other state-of-the-art methods on Paris Street View dataset [23], which contains 14900 images for training and 100 images for testing.

4.2. Comparisons with state-of-the-art methods

Quantitative Comparison. We compare our model with recent state-of-the-art methods on Paris Street View dataset [23]. The compared models include EdgeConnect [9], MEDFE [19], RFR [22] and CTSDG [20]. We use different types of masks with different mask ratios. We use the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), mean l_1

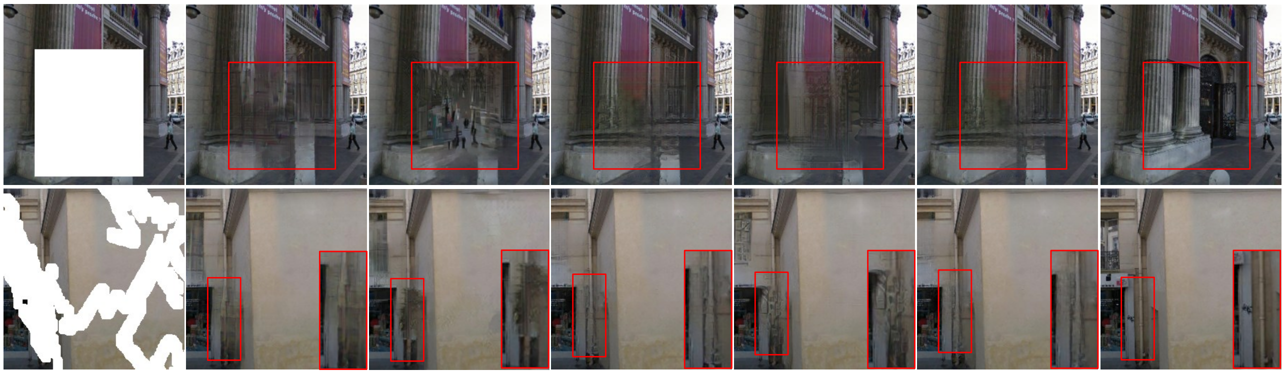

Fig. 4. Visual quality comparison of various image inpainting methods.

Table 1. Quantitative comparison of various image inpainting methods.

Mask Type		Irregular Mask			Square Mask		
Mask Ratio(%)		20-30	30-40	40-50	20-30	30-40	40-50
Mean L1↓	EC	0.0118	0.0216	0.0302	0.0175	0.0285	0.0402
	MEDFE	0.0125	0.0235	0.0334	0.0214	0.0348	0.0494
	RFR	0.0108	0.0200	0.0280	0.0176	0.0281	0.0403
	CTSDG	0.0110	0.0203	0.0283	0.0185	0.0295	0.0408
	Ours	**0.0107**	**0.0198**	**0.0277**	**0.0174**	**0.0279**	**0.0401**
SSIM↑	EC	0.9072	0.8454	0.7887	0.8758	0.8063	0.7371
	MEDFE	0.9043	0.8362	0.7738	0.7738	0.7831	0.7040
	RFR	0.9162	0.8577	0.8037	0.8790	0.8111	0.7420
	CTSDG	0.9132	0.8510	0.7952	0.8677	0.7953	0.7264
	Ours	**0.9174**	**0.8594**	**0.8058**	**0.8797**	**0.8123**	**0.7434**
PSNR↑	EC	28.9555	25.6320	23.8351	26.1639	23.6358	21.7106
	MEDFE	28.6145	25.0635	23.1334	23.1334	22.1756	20.3530
	RFR	29.6431	26.2295	24.3853	26.2626	23.8289	21.7845
	CTSDG	29.4489	26.0593	24.2276	25.8047	23.4214	21.7728
	Ours	**29.7249**	**26.2853**	**24.4466**	**26.3085**	**23.8717**	**21.8137**
FID↓	EC	24.3491	37.9633	54.7269	41.2519	57.7532	73.2970
	MEDFE	28.0274	43.0151	62.1008	51.2478	71.6308	93.7418
	RFR	20.1282	29.7673	45.2115	40.2128	55.4988	70.3306
	CTSDG	25.7223	39.8158	57.4709	50.9831	72.4844	93.3651
	Ours	**20.0351**	**29.7407**	**45.0302**	**39.2793**	**54.2814**	**69.0963**

Table 2. The impact of the DPFM module

	Baseline	Our(w/o DPFM)	Our(w/ DPFM)
Mean L1	0.01170	0.01161	**0.01160**
SSIM	0.92027	0.92086	**0.92088**
PSNR	29.32795	29.39742	**29.40792**
FID	24.95860	25.05563	**24.50298**

error and Fréchet inception distance (FID) [24] as evaluation metrics. Table 1 shows the quantitative comparison results of various image inpainting methods. As can be seen, the proposed method outperforms the other state-of-the-art approaches.

Qualitative Comparison. Fig. 4 shows two examples of visual comparison. It can be seen that the EdgeConnect [9], MEDFE [19] and CTSDG [20] produced obvious artifacts and suffer from distorted structures. RFR [22] generates competitive results, but it brings repetitive abnormal texture. In contrast, our method generates more realistic texture.

4.3. Effectiveness of Modules

We verify the effectiveness of the proposed module. The fist model (dubbed baseline) moves the texture branch and the DPFM. The second model moves the DPFM. The third model is the complete model. As shown in Table 2, the Mean L1, SSIM, PSNR and Fid demonstrate effectiveness of the texture branch and the DPFM module.

5. CONCLUSION

In this paper, we propose the detail generation and fusion network (DGFNet), which includes the dual-stream texture generation network and the multi-scale difference perception fusion network. The texture branch in the dual-stream texture generation network explicitly models missing texture areas and improves the performance of the whole model on detail repair. The fusion network integrates coarse information and texture information in multi-scale features. Experiments on common dataset validate the feasibility of our method.

6. REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24, 2009.

[2] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg, "Transformation-grounded image generation network for novel 3d view synthesis," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 3500–3509.

[3] Alexei A Efros and Thomas K Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*. IEEE, 1999, vol. 2, pp. 1033–1038.

[4] Alexei A Efros and William T Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 341–346.

[5] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu, "High-resolution image inpainting with iterative confidence feedback and guided upsampling," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17.

[6] 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*, 2019, pp. 4170–4179.

[7] Xue Zhou, Tao Dai, Yong Jiang, and Shu-Tao Xia, "Bishift-net for image inpainting," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2470–2474.

[8] Zhilin Huang, Chujun Qin, Ruixin Liu, Zhenyu Weng, and Yuesheng Zhu, "Semantic-aware context aggregation for image inpainting," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2465–2469.

[9] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[10] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo, "Foreground-aware image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5840–5848.

[11] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 181–190.

[12] Liang Liao, Ruimin Hu, Jing Xiao, and Zhongyuan Wang, "Edge-aware context encoder for image inpainting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3156–3160.

[13] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.

[15] 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/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[16] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[17] 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 (ECCV)*, 2018, pp. 85–100.

[18] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao, "Progressive reconstruction of visual structure for image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5962–5971.

[19] 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*. Springer, 2020, pp. 725–741.

[20] Xiefan Guo, Hongyu Yang, and Di Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14134–14143.

[21] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia, "Structure extraction from texture via relative total variation," *ACM transactions on graphics (TOG)*, vol. 31, no. 6, pp. 1–10, 2012.

[22] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7760–7768.

[23] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros, "What makes paris look like paris?," *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.