W-Net: Structure and Texture Interaction for Image Inpainting

Ruisong Zhang, Weize Quan, Yong Zhang, Jue Wang, and Dong-Ming Yan

***Abstract*—Recent literature has developed two advanced tools for image inpainting: appearance propagation and attention matching. However, given the ineffective feature reorganization and vulnerable attention maps, existing works yield suboptimal results with distorted structures and inconsistent contents. Furthermore, we observe that deep sampling layers (DSL) and shallow skip connections (SSC) in U-Net separately promote image structure inference and texture synthesis. To address the above two issues, we devise a W-shaped network (W-Net), which consists of two key components: a texture spatial attention (TSA) module in SSC and a structure channel excitation (SCE) module in DSL. W-Net is a two-stage network, with coarse and refined structures derived at each stage. Meanwhile, the TSA module fills incomplete textures with reliable attention scores under the guidance of coarse structures, which effectively diminishes inconsistency from appearance to semantics. The SCE module rectifies structures according to the difference between coarse structures and refined structures enhanced by texture features. Then the module motivates them to produce more reasonable shapes. Complete textures and refined structures constitute desired inpainted images, as the output of W-Net. Experiments on multiple datasets demonstrate the superior performance of W-Net. The source code is available at https://github.com/Evergrow/W-Net.**

***Index Terms*—Image inpainting; Structure and texture; Convolutional neural network; Attention**

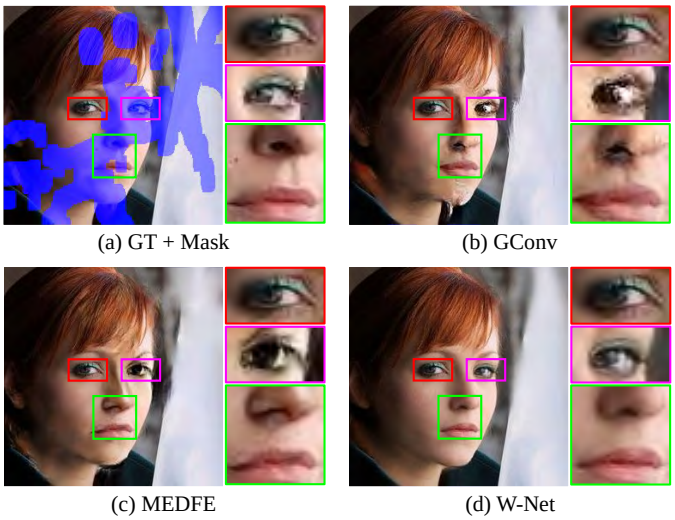
I. Introduction

THE longstanding goal of image inpainting has been to synthesize visually realistic contents in missing regions of damaged images, with contextual coherence from low-level textures to high-level semantics. Tackling this ill-posed problem presents various practical values in computer vision and graphics communities, *e.g.*, restoring deteriorated photographs [1], removing unwanted targets [2], completing occluded regions [3], [4], and editing appointed information [5].

Prior to the deep learning era, without powerful tools to mine the treasure from massive data, traditional works simply fill damaged pixels based on the low-level feature of neighbor regions, *e.g.*, propagating the appearance from borders [6], [7], [8] and copying the matching patches from contexts [9], [10], [11]. These methods perform well in repetitive textures and simple patterns but suffer from unreasonable semantics and inconsecutive structures in real-world scenarios.

In contrast, inpainting methods based on the convolutional neural network (CNN) [14], [15], [16] utilize deep generative models, such as variational autoencoder (VAE) [17] and generative adversarial network (GAN) [18], to learn the latent mapping from corrupted input images to paired ground-truth images. Recent mainstream deep works inherit traditional methods completing holes in embedded latent codes, *i.e.*, diffusion-based methods and attention-based methods. The former first extend meaningful structures, such as edges [19], [20], contours [21], and flows [22], from contexts to holes and synthesize fine textures sequentially. The later leverage the attention mechanism [23], [24], [25], [26] to search matching features in the background to enhance the feature representation of missing regions. Moreover, some methods [27], [28], [29] model and sample from the discrete distribution over the latent codes to derive diverse results.

With the involvement of a rich semantic corpus [30], [31], [32] and an adversarial training process [18], [33], deep inpainting techniques achieve impressive results. However, some issues prevail in both diffusion-based and attention-based deep methods. First, current diffusion-based [19], [21], [22] approaches have an overdependence on the supervision of prior appearance information. They rarely rearrange features in networks, *e.g.*, explicitly spread contents into spoiled regions.



(a) GT + Mask (b) GConv (c) MEDFE (d) W-Net

Fig. 1: Illustration of issues to solve. (a) The ground truth image with blue shadow as mask. (b) The result of attention-based method GConv [12]. (c) The result of diffusion-based method MEDFE [13]. (d) The result of our W-Net. From zoom-in regions, GConv and MEDFE fail to adaptively fill symmetrical eyes in red and magenta boxes, and suffer from distorted structures in green boxes.

R. Zhang, W. Quan, and D.-M. Yan are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with School of Artificial Intelligence, the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yandongming@gmail.com). *(Ruisong Zhang and Weize Quan are co-first authors. Corresponding author: Dong-Ming Yan.)*

Y. Zhang and J. Wang are with the Tencent AI Lab, ShenZhen, P.R.China.

Second, most attention-based [23], [24], [25] approaches often obtain unstable affinities between holes and contexts, because features extracted from holes without support of valid pixels are different from those of contexts in both appearance and semantics. As illustrated in Fig. 1, given the above limitations, existing methods fail to produce perfect structures and symmetrical objects consistently. Latest pioneer works [34], [13], [35], [36] have focused on these issues. Furthermore, these works have proposed some feasible strategies, *e.g.*, progressive inference [37], [34], structure and texture fusion [13], [20], and supervised attention map [35]. However, none of these proposals have solved the two abovementioned problems at the same time.

To address these two issues jointly, we analyze how U-Net [38], as the backbone of the generator, can restore impaired images. By introducing contrastive variants, *i.e.*, deleting *shallow skip connections* (SSC) and deleting *deep sampling layers* (DSL), we observe that SSC in U-Net transmit detailed textures from input images to inpainted images. Meanwhile, cascading contractions in DSL are responsible for semantic structure reasoning, which coincides with the view in [13]. Moreover, serial contractions enlarge the receptive field of U-Net, thus making it superior to ResNet [38] in the inference of deep structures. From this perspective, the ideal inpainting backbone is U-Net rather than the usual ResNet.

Based on the above insights, we present a two-stage generative network with structure-texture mutual guidance for image inpainting (see Fig. 3). The proposed network yields a W-shaped architecture called W-Net. In the first stage, damaged images undergo successive hierarchical contractions to extract coarse structure features. Moreover, expansive operations propagate valid structure information to corrupted regions. Then, multi-scale *texture spatial attention* (TSA) modules utilize coarse structure features as a proxy to calculate multi-level score maps and gather encoded features in shallow layers to stick textures in missing regions. Unlike existing attention blocks [23], [24], [25] implementing the pairwise affinity and weighted representation under the same feature maps, our TSA module borrows structure affinities to synthesize textures within SSC, which enhances the robustness of attention maps and relieves semantic inconsistency. Combined with recovered texture information, the second stage produces structure features in the same way as the first stage. At the same time, multi-scale *structure channel excitation* (SCE) modules in DSL refine structure features. The core idea of SCE module is to rectify the channel activation of structure features according to the difference between coarse and refined structure information. Plausible structures and vivid textures constitute high-quality inpainted images as the result. W-Net is trained with a pixel-wise detector [33] to eliminate visual artifacts in inpainted images. According to the result presented in Fig. 1, the TSA module and the SCE module reliably overcome distortion and incoherence.

Our contributions are summarized as follows:

- We propose a *texture spatial attention* (TSA) module, which adaptively synthesizes textures in shallow skip connections based on structure guidance. The TSA module alleviates incoherence from appearance to semantics.
- We develop a *structure channel excitation* (SCE) module that refines structures according to residual learning, thus leading to high-fidelity structures.
- These modules are assembled into W-Net with structure and texture interaction. Comparisons and ablation studies demonstrate the effectiveness.

II. Related Work

In this section, we mainly summarize CNN-based works related to our method. Early deep model-based works [14], [39], [40], [15] extract semantic features and perceive plausible structures and predictable textures of corrupted images for satisfactory results. Inspired by these studies, recent deep works tend to be diversified. We roughly divide them into two categories: diffusion-based methods and attention-based methods.

Diffusion-based methods inheriting the propagation of isophotes [6], [7], [8] typically complete appearance information and then facsimile detailed textures. For instance, EdgeConnect [19] first sketched binary edge maps and combines impaired images with intact edges to paint color pixels in missing regions. As the structural representation of binary images is relatively rough, some works substitute foreground contours [21], edge-preserved flows [22], or monochromic images [41] for edge maps as prior information to guide restoration, which can acquire more accurate structures. The advantage of these two-stage methods is that they decompose a tricky problem into two feasible subproblems: structure reconstruction and texture generation. Moreover, users can make creative edits to structure images in the first stage to produce more desired results. Liu *et al.* [13] jointly restored structures and textures in a mutual encoder-decoder, where structure features and texture features benefit each other. However, when corrupted regions become larger, directly regressing preprocessing structural images becomes more difficult. Progressive methods [42], [34], [43] attempted to transmit useful contents from boundaries to the center of hole regions through iterative modules. To tackle large mask cases, RFR [34] recurrently gathered and infer the encoded features in hole boundaries. Zeng *et al.* [43] suggested a feedback mechanism in iterative inpainting, which gradually improves low-confidence pixels inside holes. Recently, Peng *et al.* [29] developed a two-stage inpainting model of the diverse structural feature generation and the synthesis of texture details.

Attention-based methods introduce the attention mechanism that uses the linear representation of correlative valid features to strengthen features in holes to maintain global consistency. This process can be regarded as the realization of patch matching [9], [10], [11] in deep feature space. Contextual attention [23] leveraged valid patches as convolutional filters to synthesize unknown patches. Inspired by [23], PEN-Net [44] filled holes by attention transfer from deep to shallow in a pyramid fashion. In addition, Yan *et al.* [24] devised a shift-connection layer, where encoder features of known regions guide to recover decoder features of missing parts. The CSA layer [25] modeled the semantic relevance in hole features for efficient image completion. However, poor reconstruction

(a) (b) (c) (d) (e) (f) (g)

Fig. 2: Visual comparisons against different baseline models and related variants. (a) GT and Mask; (b) ResNet; (c) U-Net; (d) U-Net w/o DSL; (e) U-Net w/o SSC; (f) U-Net w/ Det; (g) W-Net w/ Det. "Det" means the pixel-wise detector. To avoid the impact of model size on performance, we align the number of parameters for each model, *i.e.*, ResNet: 52.50 M, U-Net: 58.61 M, U-Net w/o DSL: 50.58 M, U-Net w/o SSC: 57.96 M, and W-Net: 46.49 M.

before attention layers inevitably leads to ineffective feature matching in score maps. To overcome this limitation, Zhou *et al.* [35] devised a dual spatial attention module using ground-truth affinities as the oracle supervision for high-fidelity face completion. Zeng *et al.* [36] modified CA modules [23] into an auxiliary branch supervised by contextual reconstruction loss to find proper reference regions. Meanwhile, partial convolution [45] renormalized the weights of convolutional filters on valid pixels, which is considered as a handcrafted attention mechanism. Other mask-aware methods like DS-Net [46] and MADF [47] designed dynamic convolution and normalization to achieve better results. Motivated by [45], Yu *et al.* [12] designed gated convolution to learn the dynamic weights for each pixel from data. Furthermore, Xie *et al.* [48] presented a bidirectional attention map for the soft mask update. Meanwhile, Zhang *et al.* [33] inserted the weight map from a dense detector into reconstruction loss for removing visual artifacts. Recent research [20] has proposed a two-stream generator jointly inferring image structures and textures with an attention-based multi-scale feature aggregation.

III. Effect Analyses of Architectures

Previous inpainting networks mainly include two types: few samplings with no skip connections represented by ResNet [38] and rich samplings with skip connections represented by U-Net [38]. Comparing the results of two backbones shown in Fig. 2b and 2c, U-Net yields complete and reasonable structures, while ResNet fails to make this information clear, especially in large missing regions. In U-Net, the original resolution is progressively scaled down to 1×1, thus providing a sufficient receptive field to cover the whole contextual regions for inpainting. Some recent works [15], [23], [19], [22] have attempted to enlarge the receptive field of ResNet or hallucinate novel contents using dilated convolution [49] or perceptual loss [50]. While these approaches are effective in some cases, the mechanisms of their networks are implicit to analyze due to the lack of hierarchical features.

To explore and analyze the workflow of U-Net, we introduce two related variants of U-Net, *i.e.*, reducing the number of the sampling layer (same as [51]) and removing skip connections between shallow layers. For the architecture without deep sampling layers (DSL), restored images in Fig. 2d also lose semantic structures, which indicates that contractions in deep layers facilitate the structure recovery. Shallow skip connections (SSC) in U-Net extract detailed textures and synthesize them in generated images. Thus, deleting SSC would cause blurry results in valid regions as shown in Fig. 2e. Ambiguous filled regions from Fig. 2c to Fig. 2e illustrate that the vanilla reconstruction loss (*e.g.*, ℓ_1 distance) always degrades to smooth inpainted contents for less error risk. We add a detector as a rival to the generator [33] to make the missing regions appear through some fine details. However, these results (Fig. 2f) still suffer from trivial textures and unsuitable structures because U-Net directly transmits incomplete textures in SSC and misses the revision to immature structures in DSL.

IV. Proposed Method

Inspired by the observations on U-Net and its shortages, we first propose W-Net in Sec. IV-A. Then, we discuss how our designed texture spatial attention (TSA) module and structure channel excitation (SCE) module improve the inpainting capability of W-Net in Sec. IV-B and IV-C. In Sec. IV-D, we describe the implementation of the whole inpainting framework, where W-Net and a dense detector [33] work together to reduce blurry artifacts with perceptual loss [50].

A. W-Net

As shown in Fig. 3, W-Net cascades two incomplete U-Net, in which features are reorganized into eight layers (from 128×128 to 1×1): three shallow layers as texture features and five deep layers as structure features. This division is based on the insight, *i.e.*, shallow skip connections synthesize textures and deep sampling layers generate structures, as mentioned in

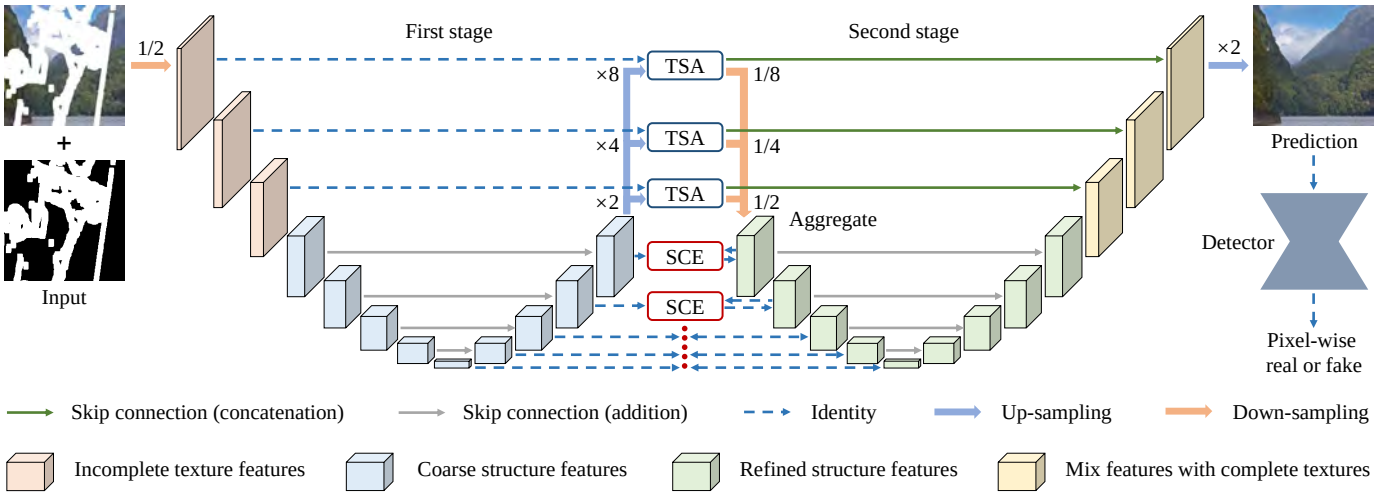


Fig. 3: The architecture of our designed W-Net with the dense detector. Note that all cubes in the figure represent features rather than convolutional blocks. The sampling rate is marked around the arrow. Arrows denoting up/down-sampling between adjacent features are omitted for simplicity. The detailed schema of TSA module and SCE module are illustrated in Fig. 4.

Sec. III. In the first stage, W-Net employs sequential down-sampling convolutions to extract semantic structure features of the corrupted image. Embedded features are added to corresponding previous features and expanded to higher resolutions for propagating structure information from contexts to holes. This simple encoder-decoder can effectively recover rough structures in missing regions, as demonstrated in Fig. 2e. Multi-scale TSA modules, guided by the topmost structure feature, fill holes in each texture feature layer successively. Specifically, TSA modules utilize incomplete texture features and structure features with necessary up-sampling (blue solid arrows in Fig. 3) as the input to obtain complete texture features, which are conveyed to the second stage through skip connections (three dark green arrows in Fig. 3). The aggregation of filled texture features from all TSA modules is encoded into deep structure features. Then, multi-scale SCE modules enhance these structure features. Activated structure features produced by the SCE module, which replace original structure features (five two-way blue dotted arrows), are down-sampled and input into the next-level SCE module. Finally, complete textures and refined structures are concatenated to form a high-quality image that serves as the output of W-Net.

W-Net arranges texture features and structure features into two flows, respectively. The texture flow, with corrupted textures as the input, synthesizes complete textures with coarse structures as an intermediary guidance. Meanwhile, the structure flow, with coarse structures as the start, refines precise structures with complete textures as a middle supplement. This interaction between the media can enhance the capability of both flows. Different from the previous two-stage networks [23], [12], [25], [19] that supervise coarse inpainting results, W-Net without semi-finished results only regresses the final results to ground-truth images. Compared with different targets at each stage, end-to-end training with a uniform objective is beneficial in reorganizing features in networks. Unlike most diffusion-based approaches [19], [22], [13], W-Net does not require any prior information as supervision. The principle of W-Net is based on the intrinsic property that textures displayed as pixels are sensitive to contractions while structures represented as patches are the opposite.

B. Texture Spatial Attention Module

Shallow skip connections in U-Net lack the requisite filling operation, thus resulting in blurry textures and inconsistent contents in filled images, as shown in Fig. 2f. To solve this issue, we design a novel TSA module to establish the long-term dependencies of features in the spatial domain so as to enhance corrupted features and ensure global consistency. Original non-local blocks [52], [53] catch the self-attention of entire images for feature self-activation, while the TSA module obtains auxiliary information from structures to intensify texture features in missing regions, as shown in Fig. 4a.

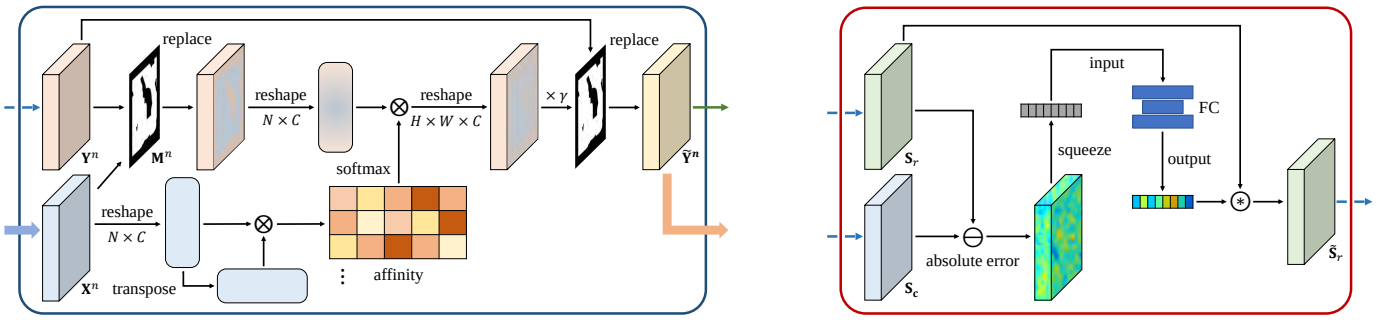
Formally, $\mathbf{X}$ is the topmost structure feature, and $\mathbf{Y}^n (n \in \{1,2,3\})$ denotes the n-th texture layer. For instance, when restoring $\mathbf{Y}^n$, $\mathbf{X}$ is up-sampled by n times to $\mathbf{X}^n$ with the same resolution as $\mathbf{Y}^n$. The corresponding $\mathbf{X}^n$ replaces corrupted features in the mask $\mathbf{M}^n$, *i.e.*, $\mathbf{Y}^n = (\mathbb{1} - \mathbf{M}^n) \odot \mathbf{Y}^n + \mathbf{M}^n \odot \mathbf{X}^n$, where $\odot$ denotes the pixel-wise product. Subsequently, $\mathbf{X}^n$ and $\mathbf{Y}^n$ are reshaped into $N \times C$, where N, C are the number of feature locations and channels, respectively. The affinity $a^n_{i,j} \in \mathbb{R}^{N \times N}$ between the i^{th} location and j^{th} location in $\mathbf{X}^n$ is written as

$$a^n_{i,j} = \frac{\exp(s_{ij})}{\sum_{k=1}^{N} \exp(s_{ik})}, \text{ where } s_{ij} = \langle \mathbf{X}^n_i, \mathbf{X}^n_j \rangle. \tag{1}$$

The computation of the score map directly uses raw features to retain the robustness of the TSA modules. According to the attention map computed from $\mathbf{X}^n$, $\mathbf{Y}^n$ is reinforced by

$$\tilde{\mathbf{Y}}^n = \gamma \cdot \mathbf{M}^n \odot \mathcal{F}(\mathbf{Y}^n \otimes \mathbf{A}^n) + \mathbf{Y}^n, \tag{2}$$

where $\mathcal{F}(\cdot)$ is the reshape operation from $N \times C$ to $H \times W \times C$ (H, W are the height and width of original features), $\otimes$ is the



(a) TSA for texture synthesis (b) SCE for structure rectification

Fig. 4: Overview of TSA module (a) and SCE module (b). Let $\otimes$ be matrix multiplication, $\ominus$ be pixel-wise absolute error, and $\circledast$ be channel-wise multiplication. The rest symbols are the same as those in Fig. 3.

matrix multiplication. γ is a learnable weight to merge features in corrupted regions. Filled texture features $\tilde{\mathbf{Y}}^n$ is output to the shallow skip connection and down-sampled for aggregation.

Essentially, the TSA module uses the correlation between structure features to repair texture features. After a comprehensive inference, holes and contexts in the structure features follow the trend of the similar modality from appearance to semantics in Fig. 2e, which make the calculating score map more reliable. Most attention blocks in inpainting works such as CA [23] and CSA [25], which are parallel or within the deep reasoning process, may be disturbed by incongruent features. During the enhancement of invalid features, we consider global features instead of contextual features [23], [12] only, and the self-representation in mask regions makes the completed contents more consistent.

C. Structure Channel Excitation Module

After making deep inferences twice, repaired results, especially in large impaired regions, sometimes appear with ambiguous semantics and unsharp structures, which affects visual realism. Hence, we propose an SCE module, which is based on the disparity between coarse and refined structure features, and adaptively recalibrates channel responses of refined structure features to tackle the above scenarios. Concretely, we take $\mathbf{S}_c$ as the coarse structure features, and use $\mathbf{S}_r$ to denote refined structure features. $\mathbf{S}_r$ extracted from filled texture features are more accurate than $\mathbf{S}_c$ to represent structure information. The SCE module reinforces correct structures that are added after refined reasoning through residual learning so that hole structures have the same delicate degree as background structures. Inspired by the SE block [54], we utilize three fully-connected (FC) layers to capture channel-wise dependencies and rescale $\mathbf{S}_r$, which are formulated as

$$\tilde{\mathbf{S}}_r = \mathrm{FC}(\mathcal{G}(|\mathbf{S}_r - \mathbf{S}_c|)) \circledast \mathbf{S}_r. \tag{3}$$

Here, $\mathcal{G}(\cdot)$ denotes squeezing global spatial information into a channel descriptor, and $\circledast$ means the channel-wise multiplication. As shown in Fig. 4b, the above transformation is performed at each level of structure features successively, and outputs corresponding activated structure features $\tilde{\mathbf{S}}_r$.

D. Implementation

The input of W-Net is an image with white pixels as impaired regions and the corresponding binary mask, where black pixels indicate valid regions and white pixels indicate missing regions. Practically, a pair of a ground-truth image $\mathbf{I}_{gt}$ and a mask $\mathbf{M}$ blends to obtain an impaired image $\mathbf{I}_{im}$, *i.e.*, $\mathbf{I}_{im} = \mathbf{I}_{gt} \odot (\mathbb{1} - \mathbf{M}) + \mathbf{M}$. W-Net fills holes and outputs the prediction image $\mathbf{I}_{pred}$ with the same resolution as $\mathbf{I}_{gt}$. The prediction image merges with the impaired image as the final completed result $\mathbf{I}_{re}$, written as $\mathbf{I}_{re} = (\mathbb{1}-\mathbf{M})\odot\mathbf{I}_{im}+\mathbf{M}\odot\mathbf{I}_{pred}$. As mentioned in Sec. III, taking the vanilla reconstruction loss as the objective to train W-Net would generate over-smooth results. To prevent this degradation, we leverage a pixel-wise detector to localize visual artifacts, which frequently occur in missing regions. Naturally, the detector is trained by the focal loss [55] (considering the sample imbalance problem) with $\mathbf{M}$ as a weak label, which can be formulated as

$$\begin{aligned}\mathcal{L}_{focal}(\mathbf{V}, \mathbf{M}) = -\frac{1}{N}\sum_{i=1}^{N}(1-\alpha)(1-\mathbf{V}_i)^{\gamma}\mathbf{M}_i \log \mathbf{V}_i \\ + \alpha \mathbf{V}_i^{\gamma}(1-\mathbf{M}_i)\log(1-\mathbf{V}_i),\end{aligned} \tag{4}$$

where α is the weighting factor equal to the mask ratio of damaged images, γ is tunable focusing parameter set to 2, and $\mathbf{V}$ is the detector output for the valuation of inpainted images. A exponential function transforms the valuation to the weight, which combines with the reconstruction loss to train W-Net. To further enhance the nontrivial details, $\mathbf{I}_{re}$ and $\mathbf{I}_{gt}$ are measured in the feature space [50], where the pre-trained VGG16 model [56] on ImageNet [57] extracts abundant semantics. The loss is written as

$$\mathcal{L}_{perc} = \frac{1}{L}\sum_{i=1}^{L}||\Phi_i(\mathbf{I}_{re}) - \Phi_i(\mathbf{I}_{gt})||_1, \tag{5}$$

where $\Phi_i(\cdot)$ is the output of the i-th layer of VGG16. To this end, the final training objective can be written as

$$\begin{cases} \min\limits_{G} \ ||x^{D(G(\mathbf{I}_{im},\mathbf{M}))} \odot (G(\mathbf{I}_{im},\mathbf{M}) - \mathbf{I}_{gt})||_1 + \lambda\mathcal{L}_{perc}, \\ \max\limits_{D} \ -\mathcal{L}_{focal}(D(G(\mathbf{I}_{im},\mathbf{M})),\mathbf{M}). \end{cases} \tag{6}$$

Here, we define G as the generator, and define D as the detector. x is a base number of exponential function set to 10, λ is the trade-off weight set to 0.05.

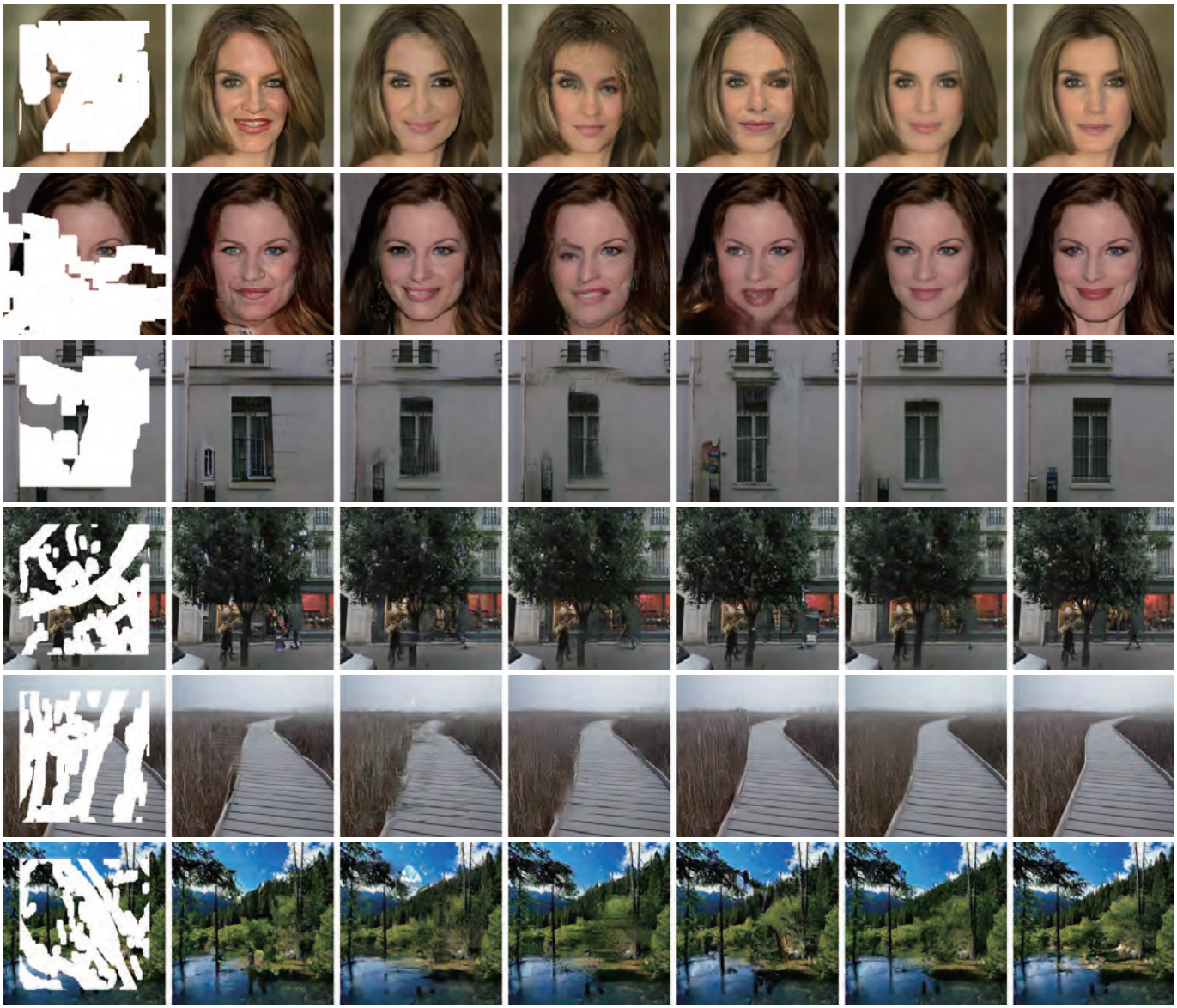


Fig. 5: Qualitative comparisons of our W-Net with GConv [12], MEDFE [13], CTSDG [20], and DSI [29] for irregular masks. From top to bottom: CelebA-HQ, Paris StreetView, and Places2. More comparisons are provided in supplemental material.

V. Experiment

In this section, we first evaluate our inpainting model via quantitative and qualitative comparisons with several advanced works. We also conduct ablation studies about the TSA module and the SCE module to validate their effectiveness. Finally, we present the real-world applications of our method.

A. Experimental Settings

All experiments are performed on three extensively adopted datasets: CelebA-HQ [30], [58], Places2 [31], and Paris StreetView [32]. We split these three datasets into training and testing set according to the method in [33], and utilize QD-IMD[1] (human drawings) to construct a large-scale irregular mask set for training models. For diversified evaluation, test mask sets include irregular masks (from Liu *et al.* [45]) and a center square mask. All images and masks in training and testing are resized to 256×256.

Inpainting models are optimized by the Adam optimizer [59], where $\beta_1 = 0.0$ and $\beta_2 = 0.9$. We separately retain the learning rate of the generator and detector at 10^{-4} and 10^{-5}. On a single NVIDIA TITAN RTX GPU (24GB), we train our model for 200 epochs with a batch size of 4.

B. Comparisons with State-of-the-arts

We qualitatively and quantitatively compare our method with several advanced methods: GConv [12], MEDFE [13], CTSDG [20], DSI [29], DS-Net [46], and MADF [47].

[1]https://github.com/karfly/qd-imd

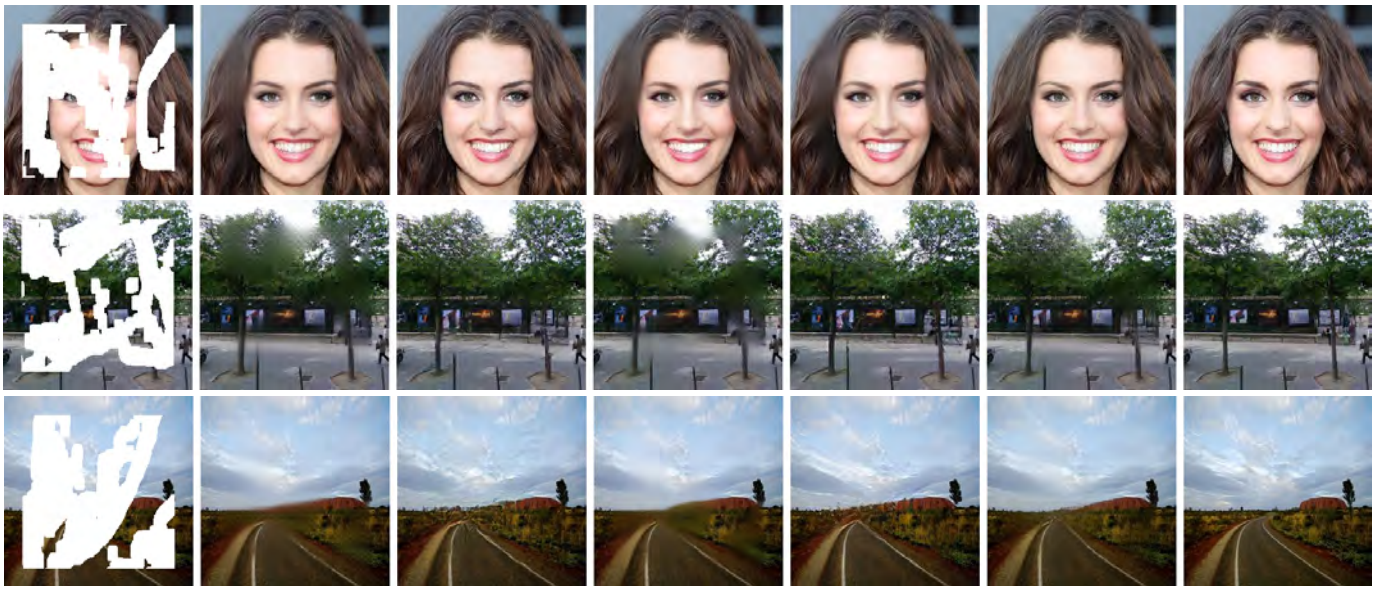


Fig. 6: Qualitative comparisons with DS-Net [46], MADF [47], and their versions without the style loss (DS-Net* and MADF*).

Fig. 7: Qualitative comparisons with GConv [12], MEDFE [13], CTSDG [20], and DSI [29] for center square masks.

Qualitative Comparisons. Fig. 5 shows the results for filling irregular holes. Given the rich deep samplings and designed SCE modules, W-Net has powerful capability to reconstruct facial shapes and infer window structures. GConv, as a ResNet-based network, is insufficient to handle them, and CTSDG is worse due to the lack of effective structure refinement in deep samplings. These phenomena are consistent with the viewpoint in Sec. III. Moreover, both MEDFE and DSI fail to synthesize reasonable textures on correct structures due to the insufficient organization between structure and texture features in their generators. Only W-Net effectively utilizes contextual information to synthesize contents, *e.g.*, facial symmetry and similar pattern, as shown in the second and fifth rows of Fig. 5. This adaptive filling content based on contexts is crucial for visual realism. However, our method tends to produce smooth results when masked regions with large ratio cover critical regions (the neck in the second row) or missing regions lack rich structural descriptors (the grass in the fifth row).

In Fig. 6, we also compare W-Net with DS-Net [46], MADF [47], and their versions without the style loss, which are aligned with the training loss of W-Net, *i.e.*, only using the reconstruction loss and the perceptual loss. Our method is superior to these two methods, where the style loss seems to play a decisive role in fine texture recovery.

TABLE I: Quantitative comparisons. "DS-Net*" and "MADF*" represent the results of training corresponding models without the style loss. † Lower is better. ¶ Higher is better. **Bold** font indicates the best score.

Dataset		CelebA-HQ			Paris StreetView			Places2		
Mask Ratio		1%-20%	20%-40%	40%-60%	1%-20%	20%-40%	40%-60%	1%-20%	20%-40%	40%-60%
PSNR¶	GConv	34.08	26.30	21.66	31.84	24.68	20.77	31.16	23.92	20.02
	MEDFE	33.55	26.56	22.16	31.61	25.08	21.64	30.89	24.13	20.23
	CTSDG	34.78	26.94	22.31	34.07	27.54	23.39	31.47	25.85	21.84
	DSI	34.66	27.14	22.40	33.58	26.37	22.15	31.83	24.39	20.29
	DS-Net	35.04	27.77	23.42	34.19	27.24	23.18	31.74	25.18	21.41
	DS-Net*	35.38	28.36	24.08	**34.93**	**28.09**	**24.11**	32.45	26.12	22.37
	MADF	35.05	27.97	23.66	34.22	27.58	23.39	32.46	25.72	21.77
	MADF*	**35.63**	**29.18**	**24.52**	34.29	27.66	23.67	**33.01**	**26.53**	**22.59**
	W-Net	35.49	28.42	24.04	34.29	27.72	23.75	32.66	25.92	22.11
SSIM¶	GConv	0.981	0.922	0.803	0.960	0.855	0.693	0.957	0.842	0.662
	MEDFE	0.981	0.926	0.820	0.959	0.853	0.685	0.957	0.844	0.656
	CTSDG	0.983	0.929	0.811	0.973	0.910	0.775	0.961	0.880	0.728
	DSI	0.984	0.934	0.827	0.971	0.889	0.737	0.961	0.853	0.674
	DS-Net	0.985	0.942	0.857	0.974	0.903	0.770	0.962	0.867	0.710
	DS-Net*	0.987	0.948	0.872	**0.977**	**0.916**	**0.792**	0.967	0.887	0.744
	MADF	0.984	0.945	0.868	**0.977**	0.912	0.786	0.968	0.885	0.736
	MADF*	**0.989**	**0.953**	**0.880**	0.975	0.908	0.776	**0.971**	**0.894**	**0.746**
	W-Net	0.987	0.949	0.871	**0.977**	0.912	0.788	0.968	0.884	0.736
FID†	GConv	1.42	4.37	9.74	14.05	40.71	67.42	8.44	22.15	38.44
	MEDFE	1.58	4.59	9.66	11.39	35.37	66.09	6.18	18.64	35.57
	CTSDG	1.52	5.73	10.25	9.21	33.09	63.77	3.28	13.28	33.74
	DSI	1.13	3.97	8.18	9.63	32.67	62.98	3.31	12.06	29.55
	DS-Net	**1.06**	**3.44**	**6.82**	8.80	27.76	51.16	3.22	9.86	20.04
	DS-Net*	1.27	4.99	12.07	10.36	37.18	94.48	4.44	19.51	50.16
	MADF	1.10	4.36	11.15	**7.50**	**24.78**	**49.10**	**2.51**	**7.77**	**16.82**
	MADF*	1.22	4.53	11.57	11.86	42.56	91.99	3.97	17.27	45.63
	W-Net	1.11	3.89	9.32	9.59	31.07	61.31	3.42	11.91	26.80
LPIPS†	GConv	0.023	0.076	0.153	0.037	0.113	0.235	0.044	0.130	0.243
	MEDFE	0.023	0.068	0.129	0.033	0.103	0.206	0.049	0.135	0.247
	CTSDG	0.023	0.083	0.175	0.024	0.088	0.190	0.031	0.111	0.232
	DSI	0.018	0.061	0.132	0.027	0.091	0.188	0.033	0.107	0.216
	DS-Net	0.017	**0.052**	**0.104**	0.026	0.082	0.166	0.037	0.104	0.194
	DS-Net*	0.026	0.069	0.144	0.031	0.106	0.220	0.045	0.139	0.271
	MADF	0.017	0.063	0.139	**0.022**	**0.074**	**0.161**	**0.027**	**0.084**	**0.178**
	MADF*	0.019	0.069	0.153	0.039	0.125	0.261	0.045	0.145	0.268
	W-Net	**0.016**	0.054	0.115	0.027	0.085	0.179	0.033	0.103	0.206

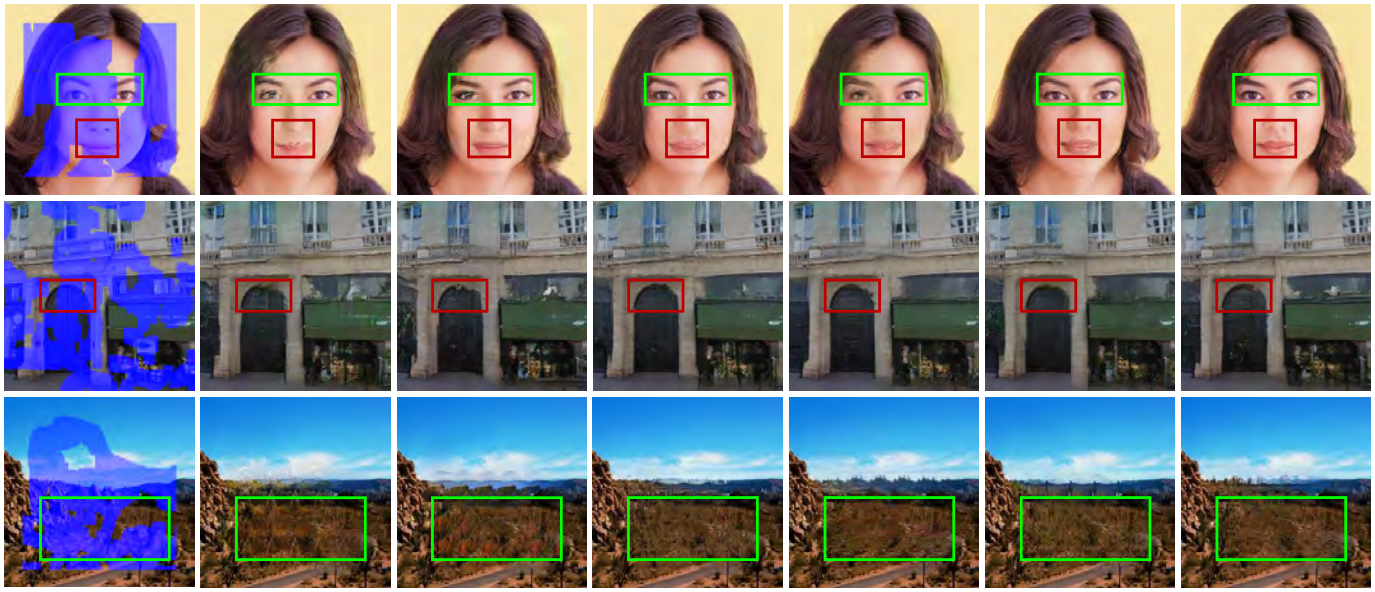
To assess the generality of mask shapes, we test inpainting models on the center square mask without extra training, and the results are shown in Fig. 7. Compared with the other four methods, our inpainting results have fewer visual artifacts.

Quantitative Comparisons. We utilize PSNR, SSIM [60], FID [61], and LPIPS [62] to measure the performance quantitatively. As shown in Table I, our method performs well for all metrics on the CelebA-HQ dataset because we consider facial symmetry and the results are closer to ground-truth images. The interaction between structures and textures enables W-Net to achieve superior results on the other two datasets. DSI samples from latent codes for diverse results, which makes higher FID/LPIPS and lower PSNR/SSIM, especially in large mask ratios. DS-Net and MADF perform very well in FID and LPIPS, while DS-NET* and MADF* perform the opposite, mainly due to the style loss favoring perception measure.

C. Ablation Studies

To verify the effectiveness of the TSA module and the SCE module, we conduct ablation studies on all three datasets without perceptual loss. The experiment involves three variants, *i.e.*, removing the TSA module and the SCE module separately and removing both modules simultaneously (baseline) for comparison. Moreover, we utilize self-attention (SA) modules and squeeze-and-excitation (SE) blocks to replace the TSA modules and the SCE modules, respectively, to demonstrate the improvement of our designed modules over existing modules. The attention modules from previous inpainting works including CA [23] and SLA [27] are also used to demonstrate the superiority of the TSA module.

Effect of TSA module. As shown in Fig. 8, models with the TSA module (d, f, and g) repair corrupted regions with high-fidelity symmetry and harmonious textures (please see green boxes in the first and second rows). The results of b, e, and g show that the SA module has a slight improvement in texture synthesis but is far inferior to the TSA module. Numerical metrics in Table III agree with the above observation because the TSA module explores correlations in the topmost structure features, where missing regions and valid regions have similar forms of descriptor. Therefore, the TSA module can employ

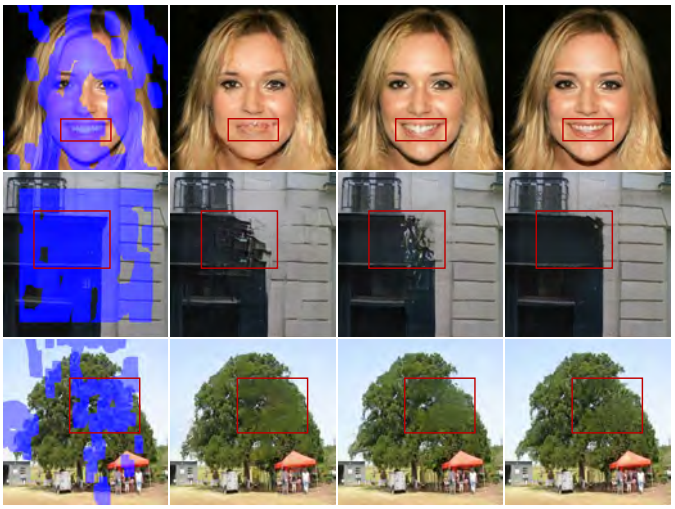


(a) GT + Mask (b) Baseline (c) W/o TSA (d) W/o SCE (e) TSA → SA (f) SCE → SE (g) Full W-Net

Fig. 8: Ablation studies of TSA module and SCE module. (a) Ground truth images with masks; (b) Baseline; (c) W-Net without TSA module; (d) W-Net without SCE module; (e) Replacing TSA modules with SA modules; (f) Replacing SCE modules with SE modules; (g) W-Net. Green and red boxes highlight effects of TSA module and SCE module, respectively.

TABLE II: Quantitative comparisons of different attention modules. † Lower is better. ¶ Higher is better. **Bold** font indicates the best score.

Dataset		CelebA-HQ	Paris SV	Places2
PSNR¶	TSA	**28.52**	**28.19**	**26.71**
	SLA	28.48	27.91	26.33
	CA	26.75	26.56	25.61
SSIM¶	TSA	**0.935**	**0.886**	**0.868**
	LSA	0.933	0.885	0.862
	CA	0.908	0.852	0.842
FID†	TSA	**4.13**	**32.73**	**11.42**
	SLA	4.18	33.44	12.22
	CA	6.61	44.41	16.55
LPIPS†	TSA	**0.072**	**0.107**	**0.123**
	SLA	0.073	0.109	0.125
	CA	0.100	0.138	0.164



(a) GT + Mask (b) CA (c) SLA (d) TSA

Fig. 9: Ablation studies of different attention modules.

robust score maps to recover incomplete textures with semantic and apparent consistencies, thus significantly promoting the visual realism. As the innovation of the TSA module lies on its ability to determine where to calculate affinities, rather than how to calculate affinities. CA [23] adopts the convolution to reproduce self-attention module, and SLA [27] progressively computes the attention maps on the previous level of output. Both numerical comparisons in Table II and visual effects in Fig. 9 show that TSA is superior to CA and SLA.

Effect of SCE module. In the first and third rows of Fig. 8, comparisons in the visual effect of red boxes among d, f, and g present that the SCE module (superior to SE block) activates to generate fine-grained structures. When the SCE module lacks entire textures from the TSA module, the SCE module also derives relatively clear structures based on the difference between two inferences, comparing b and c. This scenario proves that the SCE module can work independently from the TSA module.

D. Model Computational Performance

In this subsection, we evaluate the computational performance of W-Net and its competitors and select the computational complexity (in FLOPs), number of parameters, and inference time as statistics. Inference time is the time of a forward pass of networks. As shown in Table IV, W-Net has relatively low computational complexity, a moderate number of parameters, and moderate inference time among all models. Moreover, the computational complexity of three TSA

TABLE III: Quantitative analysis of ablation studies. † Lower is better. ¶ Higher is better. **Bold** font indicates the best score.

Dataset		CelebA-HQ				Paris StreetView			
Mask Ratio		10%-20%	20%-30%	30%-40%	40%-50%	10%-20%	20%-30%	30%-40%	40%-50%
PSNR¶	full W-Net	**32.29**	**29.12**	**26.77**	**24.93**	**31.45**	**28.37**	**26.27**	**24.57**
	SCE → SE	32.11	28.97	26.65	24.81	31.35	28.34	26.21	24.50
	TSA → SA	31.22	28.58	26.52	24.79	30.47	27.79	25.83	24.17
	w/o SCE	31.74	28.78	26.53	24.74	31.35	28.31	26.19	24.48
	w/o TSA	30.60	28.34	26.38	24.64	29.96	27.58	25.69	24.12
	baseline	30.43	28.12	26.19	24.47	29.88	27.44	25.60	24.01
SSIM¶	full W-Net	**0.977**	**0.957**	**0.930**	**0.897**	**0.963**	**0.928**	**0.885**	**0.832**
	SCE → SE	**0.977**	0.955	0.928	0.895	0.962	0.927	0.884	0.830
	TSA → SA	0.972	0.952	0.926	0.892	0.956	0.921	0.877	0.821
	w/o SCE	0.976	0.955	0.928	0.894	0.962	0.927	0.883	0.830
	w/o TSA	0.970	0.952	0.925	0.891	0.951	0.917	0.875	0.820
	baseline	0.969	0.950	0.924	0.889	0.951	0.916	0.872	0.816
FID†	full W-Net	**1.88**	**3.43**	**5.05**	**7.25**	**15.19**	**27.16**	**39.83**	**53.08**
	SCE → SE	**1.88**	3.45	5.13	7.39	15.47	27.72	39.94	53.29
	TSA → SA	2.34	3.78	5.69	8.37	18.17	30.51	43.81	58.84
	w/o SCE	1.91	3.50	5.25	7.63	15.62	28.03	40.93	54.27
	w/o TSA	3.06	4.05	5.74	8.66	20.87	32.11	44.87	60.58
	baseline	3.10	3.95	5.93	8.86	19.99	31.88	45.20	60.02
LPIPS†	full W-Net	**0.031**	**0.052**	**0.079**	**0.108**	**0.044**	**0.079**	**0.119**	**0.163**
	SCE → SE	0.032	0.054	0.080	0.109	0.045	0.080	0.120	0.164
	TSA → SA	0.041	0.061	0.087	0.117	0.055	0.090	0.131	0.178
	w/o SCE	0.033	0.054	0.081	**0.108**	0.046	0.081	0.120	0.164
	w/o TSA	0.042	0.065	0.089	0.119	0.061	0.093	0.131	0.179
	baseline	0.045	0.067	0.093	0.123	0.062	0.094	0.133	0.181

TABLE IV: Model computational performance statistics.

Model	FLOPs	Params	Infer. time
GConv [12]	55.57 G	4.05 M	13.98 ms
MEDFE [13]	137.93 G	130.32 M	113.91 ms
CTSDG [20]	17.65 G	52.15 M	19.75 ms
DSI [29]	220.46 G	70.32 M	40.20 s
DS-Net [46]	9.47 G	33.09 M	62.34 ms
MADF [47]	51.77 G	85.14 M	15.59 ms
W-Net	25.19 G	46.49 M	29.58 ms

modules in W-Net is 1.84 G FLOPs, which is only 8.96% of W-Net. However, TSA modules (especially the top one) are heavy on memory, which visibly increases time costs. Memory limitations prevent W-Net from directly repairing high-resolution images. Meanwhile, we can use methods similar to CRA [63], where inpainted images in low-resolution combined with aggregated residuals produce high-resolution results. Most of the parameter cost of W-Net is in deep sampling layers (DSL), where the convolution with 512 channels takes up about 4M parameters. Reusing the deep sampling layers of coarse and refined structure inference greatly reduces the number of model parameters and does not decrease either quantitative or qualitative performance.

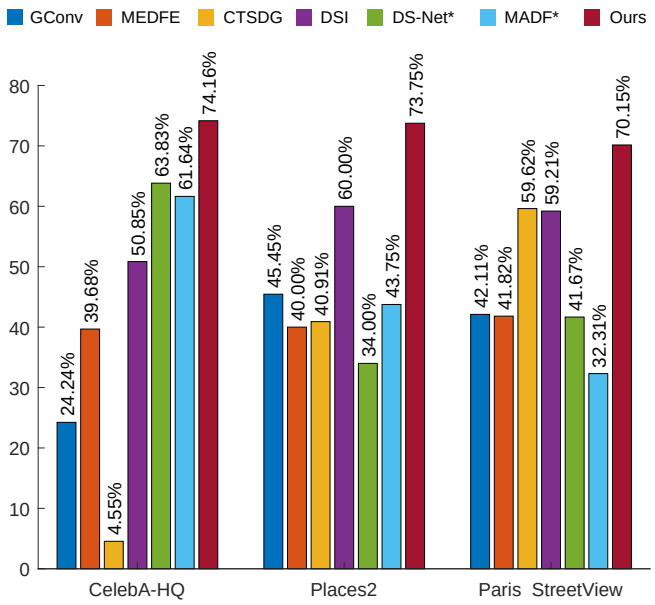


Fig. 10: The statistical results of user study. The height of the bar refers to the percentage of being selected as the more natural one, recorded at the top of the corresponding bar.

E. User Study

We conducted a user study to compare the visual quality of our method with six state-of-the-art image inpainting methods. For each of three public datasets, we randomly selected 16 inpainted images with different mask ratios (amounting to a total of 48 images). Then, we invited 15 volunteers to select the more natural one from two inpainted images generated by different methods without showing the mask and ground truth. Finally, we collected 675 votes. Fig. 10 shows the corresponding statistical results, which reveals that our W-Net has the highest probability of being selected.

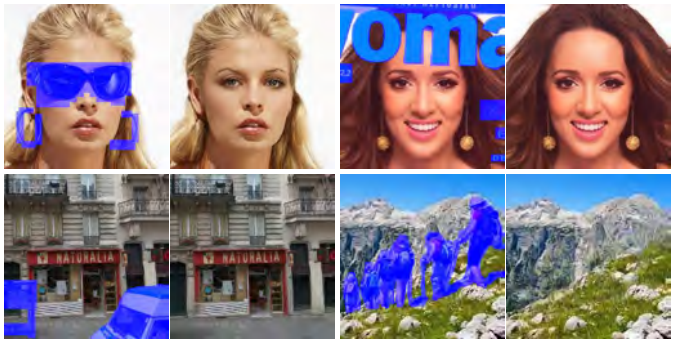


Fig. 11: Examples of daily applications using our W-Net. From top left to bottom right: face attribute editing, watermark removing, occlusion completing, and object removal.

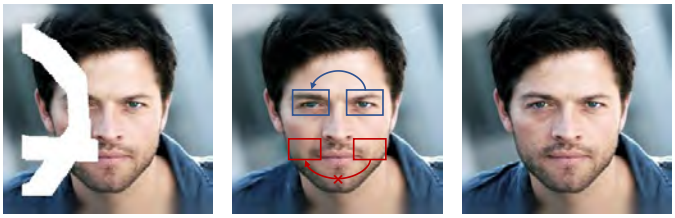


Fig. 12: Limitation. Our method cannot copy texture patterns from regions lacking structural representation.

F. Applications

Fig. 11 demonstrates the potential applications of W-Net in real-world scenarios. The examples from top left to bottom right are face attribute editing, watermark removal, occlusion completion, and object removal. Users interactively draw masks over editing regions or undesired objects to manipulate images. All results are produced by models trained on benchmarks without specific tuning. W-Net generates plausible textures and reasonable structures for required regions.

VI. Conclusion and Extension

We propose W-Net with attention modules and excitation modules for image inpainting. From the perspective of U-Net workflow, W-Net infers coarse and refined structures in two stages. Furthermore, the TSA module synthesizes the entire texture based on coarse structures, and the SCE module recalibrates structures according to their difference. Extensive comparisons and ablation studies demonstrate the superiority of W-Net in inpainting performance. However, W-Net may fail to sustain semantic coherence on regions without rich structural information, as shown in Fig. 12. The TSA module copies textures under the guidance of structural affinities, and red boxes with fewer structural descriptors cannot match as the blue boxes. Hence, a powerful structural representation with a supervision signal may be able to address this limitation.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (62102418 and 62172415), the Tencent AI Lab Rhino-Bird Focused Research Program (No: JR202127), the Open Project Program of National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University (No.2021SCUVS002), and the Open Research Fund Program of State key Laboratory of Hydroscience and Engineering, Tsinghua University (sklhse-2022-D-04).

References

[1] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, and F. Wen, "Bringing old photos back to life," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2747–2757.

[2] X. Li, B. Zhang, J. Liao, and P. V. Sander, "Let's see clearly: Contaminant artifact removal for moving cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2011–2020.

[3] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3784–3792.

[4] X. Zeng, Z. Wu, X. Peng, and Y. Qiao, "Joint 3d facial shape reconstruction and texture completion from a single image," *Computational Visual Media*, vol. 8, p. 239–256, 2022.

[5] X. Bian, C. Wang, W. Quan, J. Ye, X. Zhang, and D.-M. Yan, "Scene text removal via cascaded text stroke detection and erasing," *Computational Visual Media*, vol. 8, p. 273–287, 2022.

[6] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of ACM SIGGRAPH*, 2000, pp. 417–424.

[7] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of ACM SIGGRAPH*, 2001, pp. 341–346.

[8] 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*, vol. 10, no. 8, pp. 1200–1211, 2001.

[9] 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, p. 24, 2009.

[10] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: combining inconsistent images using patch-based synthesis," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, vol. 31, no. 4, pp. 1–10, 2012.

[11] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, vol. 33, no. 4, pp. 1–10, 2014.

[12] 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/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4471–4480.

[13] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[14] 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 (CVPR)*, 2016, pp. 2536–2544.

[15] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, vol. 36, no. 4, pp. 1–14, 2017.

[16] H. Wu, J. Zhou, and Y. Li, "Deep generative model for image inpainting with local binary pattern learning and spatial attention," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] 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*, 2014, pp. 2672–2680.

[19] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: Structure guided image inpainting using edge prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

[20] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 134–14 143.

[21] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5840–5848.

[22] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 181–190.

[23] 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 (CVPR)*, 2018, pp. 5505–5514.

[24] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-Net: Image inpainting via deep feature rearrangement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 1–17.

[25] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4170–4179.

[26] H. Sun, W. Li, Y. Duan, J. Zhou, and J. Lu, "Learning adaptive patch generators for mask-robust image inpainting," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[27] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1438–1447.

[28] A. Lahiri, A. K. Jain, S. Agrawal, P. Mitra, and P. K. Biswas, "Prior guided gan based semantic inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 696–13 705.

[29] J. Peng, D. Liu, S. Xu, and H. Li, "Generating diverse structure for image inpainting with hierarchical vq-vae," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 775–10 784.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.

[31] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[32] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?" *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, vol. 31, no. 4, pp. 101:1–101:9, 2012.

[33] R. Zhang, W. Quan, B. Wu, Z. Li, and D.-M. Yan, "Pixel-wise dense detector for image inpainting," *Computer Graphics Forum*, vol. 39, no. 7, pp. 471–482, 2020.

[34] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7760–7768.

[35] T. Zhou, C. Ding, S. Lin, X. Wang, and D. Tao, "Learning oracle attention for high-fidelity face completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7680–7689.

[36] Y. Zeng, Z. Lin, H. Lu, and V. M. Patel, "Cr-fill: Generative image inpainting with auxiliary contextual reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 164–14 173.

[37] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, "Progressive image inpainting with full-resolution residual network," in *ACM International Conference on Multimedia*, 2019, p. 2496–2504.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[39] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6721–6729.

[40] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5485–5493.

[41] T. Wang, H. Ouyang, and Q. Chen, "Image inpainting with external-internal learning and monochromic bottleneck," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5120–5129.

[42] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, "Semantic image inpainting with progressive generative networks," in *ACM International Conference on Multimedia*, 2018, p. 1939–1947.

[43] Y. Zeng, Z. Lin, J. Yang, J. Zhang, E. Shechtman, and H. Lu, "High-resolution image inpainting with iterative confidence feedback and guided upsampling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 1–17.

[44] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1486–1494.

[45] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[46] N. Wang, Y. Zhang, and L. Zhang, "Dynamic selection network for image inpainting," *IEEE Transactions on Image Processing*, vol. 30, pp. 1784–1798, 2021.

[47] M. Zhu, D. He, X. Li, C. Li, F. Li, X. Liu, E. Ding, and Z. Zhang, "Image inpainting by end-to-end cascaded refinement with mask awareness," *IEEE Transactions on Image Processing*, vol. 30, pp. 4855–4866, 2021.

[48] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8858–8867.

[49] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[50] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 694–711.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015, pp. 234–241.

[52] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[53] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[55] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

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

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

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[63] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7508–7517.