This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

WaveFill: A Wavelet-based Generation Network for Image Inpainting

Yingchen Yu[1,2] Fangneng Zhan[1] Shijian Lu[1]* Jianxiong Pan[2]
Feiying Ma[2] Xuansong Xie[2] Chunyan Miao[1]
[1] Nanyang Technological University [2] DAMO Academy, Alibaba Group
yingchen001@e.ntu.edu.sg, {fnzhan, shijian.lu, ascymiao}@ntu.edu.sg
{jianxiong.pjx, feiying.mfy}@alibaba-inc.com, xingtong.xxs@taobao.com

Abstract

Image inpainting aims to complete the missing or corrupted regions of images with realistic contents. The prevalent approaches adopt a hybrid objective of reconstruction and perceptual quality by using generative adversarial networks. However, the reconstruction loss and adversarial loss focus on synthesizing contents of different frequencies and simply applying them together often leads to inter-frequency conflicts and compromised inpainting. This paper presents WaveFill, a wavelet-based inpainting network that decomposes images into multiple frequency bands and fills the missing regions in each frequency band separately and explicitly. WaveFill decomposes images by using discrete wavelet transform (DWT) that preserves spatial information naturally. It applies L1 reconstruction loss to the decomposed low-frequency bands and adversarial loss to high-frequency bands, hence effectively mitigate inter-frequency conflicts while completing images in spatial domain. To address the inpainting inconsistency in different frequency bands and fuse features with distinct statistics, we design a novel normalization scheme that aligns and fuses the multi-frequency features effectively. Extensive experiments over multiple datasets show that WaveFill achieves superior image inpainting qualitatively and quantitatively.

1. Introduction

As an ill-posed problem, image inpainting is not to recover the original images for corrupted regions but to synthesize alternative contents that are visually plausible and semantically reasonable. It has been widely investigated in various image editing tasks such as object removal, old photo restoration, movie restoration, and so on. Realistic and high-fidelity image inpainting remains a challenging task especially when the corrupted regions are large and have complex texture and structural patterns.

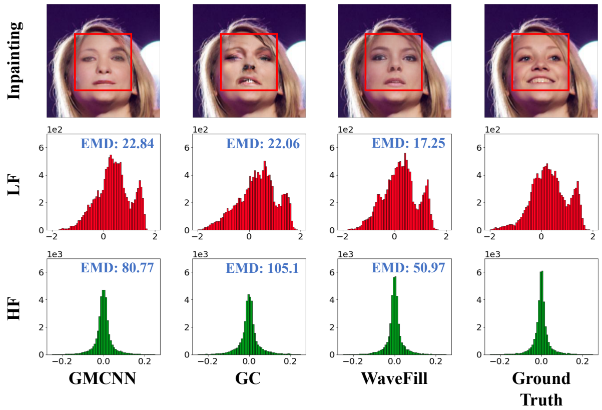


Figure 1. Image inpainting often faces a dilemma of reconstruction and perceptual quality: L1/L2 loss focuses on the reconstruction of global low-frequency structures while adversarial loss focuses on generating high-frequency texture details. State-of-the-art approaches implicitly tackle this issue by weighted summing of the two objectives (e.g. in GMCNN [35]) or employing a Coarse-to-Fine strategy (e.g. in GC [40]), but tend to produce inconsistent distributions with missing details or artifacts. The proposed *WaveFill* disentangles images into multiple frequency bands and applies relevant losses to different bands separately, which mitigates inter-frequency conflicts and produces more realistic structures and details. The distances between the ground-truth histogram and prediction histograms in both low-frequency (LF) and high-frequency (HF) are evaluated by Earth Mover's Distance (EMD) [30].

State-of-the-art image inpainting methods leverage generative adversarial networks (GANs) [10] heavily for generating realistic high-frequency details [28]. But they often face a dilemma of perceptual quality and reconstruction that share a *perception-distortion trade-off* [4]. Specifically, the adversarial loss in GANs tends to recover high-frequency texture details and improve the perceptual quality [31, 8], while the L1/L2 loss in reconstruction focuses more on recovering low-frequency global structures [28]. Concurrently optimizing the two objectives in spatial domain tends to introduce inter-frequency conflicts as illustrated in Fig. 1.

*Corresponding author

GMCNN [35] balances the two objectives by weighted sum, but it still works in spatial domain with mixed frequency and struggles to generate more realistic high-frequency details due to the inter-frequency conflicts. Gate Convolution (GC) [40] mitigates this issue by adopting a Coarse-to-Fine strategy [39, 32, 29, 20, 40] that first predicts global low-frequency structures and then refines high-frequency texture details. The coarse estimation network is generally trained with L1 loss, but the inter-frequency conflicts still exist in the refinement network. Moreover, the two-stage network often suffers from inconsistency in generated structure and texture details due to the lack of effective alignment and fusion of multi-stage features [21].

To address the aforementioned issues, we design WaveFill, an innovative image inpainting framework that employs wavelet transform to complete corrupted image regions at multiple frequency bands separately. Specifically, we convert images into wavelet domain with 2D discrete wavelet transform (DWT) [6] where the images can be disentangled into multiple frequency bands accurately without losing spatial information. The disentanglement allows us to apply adversarial (or L1) loss to the high-frequency (or low-frequency) branches explicitly and separately, which greatly mitigates the content conflicts as introduced by concurrently optimizing the two different objectives over entangled features in spatial space. In addition, we design a novel *frequency region attentive normalization (FRAN)* scheme that aggregates attention from low frequency to high frequency to align and fuse the multi-frequency features. FRAN ensures the consistency across multiple frequency bands and helps suppress artifacts and preserve texture details effectively. The separately completed features in different frequency bands are then transformed back to the spatial domain via inverse discrete wavelet transform (IDWT) to produce the final completion.

The contributions of this work can be summarized in three aspects. First, we propose WaveFill, an innovative image inpainting technique that synthesizes corrupted image regions at different frequency bands explicitly and separately, which effectively mitigates the inter-frequency conflicts while minimizing adversarial and reconstruction losses. Second, we design a novel normalization scheme that enables attentive alignment and fusion of the multi-frequency features with effective artifact suppression and detail preservation. Third, extensive experiments over multiple datasets show that the proposed WaveFill achieves superior inpainting as compared with the state-of-the-art.

2. Related Works

2.1. Image Inpainting

Image inpainting has been studied for years and earlier works employ diffusion and image patches heavily. Specifically, diffusion methods [3, 1] propagate neighboring information towards the corrupted regions but often fail to recover meaningful structures with little global information. Patch-based methods [2, 7] complete images by searching and transferring similar patches from the background. They work well for stationary texture but struggle while generating meaningful semantics for non-stationary data.

With the recent advance of deep learning, deep neural networks have been widely explored for the image generation and inpainting [45, 43, 44, 37, 38, 41]. In particular, generative adversarial networks [10] have been developed to complete images with both faithful structures and plausible appearance. For example, Pathak et al. [28] proposes a GAN-based method to complete large corrupted regions. Nazeri et al. [26] introduces EdgeConnect to predict salient edges without coarse estimation. Wang et al. [35] employs multiple branches with different receptive fields for inpainting. Zeng et al. [42] designs a Pyramid-Context Encoder Network to recover the missing regions across multiple scales. Liu et al. [21] recovers structures and textures by representing them with deep and shallow features. Liu et al. [20] introduces partial convolution with free-from masks for inpainting. On top of it, Yu et al. [40] presents gated convolution for inpainting.

Though the aforementioned methods address image completion in different manners, most of them work in the spatial domain where information of different frequencies is mixed and often introduces inter-frequency conflicts in learning and optimization. Our method instead decomposes images into the frequency space and applies different objectives to different frequency bands explicitly and separately, which mitigates inter-frequency conflicts and improves image inpainting quality effectively.

2.2. Wavelet-based Methods

Wavelet transforms decompose a signal into different frequency components and has shown great effectiveness in various image processing tasks [25]. Wavelet-based inpainting has been investigated far before the prevalence of deep learning. For example, Chan et al. [5] designs variational models with total variation (TV) minimization for image inpainting, and it's improved in [48] with non-local TV regularization. In addition, Dobrosotskaya et al. [9] combines diffusion with the non-locality of wavelets for better sharpness in inpainting. Zhang and Dai [46] decomposes images in the wavelet domain to generate structures and texture with diffusion and exemplar-based methods, respectively. The aforementioned methods leverage hand-crafted features which cannot generate meaningful content for large corrupted regions. We borrow the idea of wavelet-based decomposition and incorporate CNN representations and adversarial learning which mitigates this issue effectively.

Recently, incorporating wavelets into deep networks has

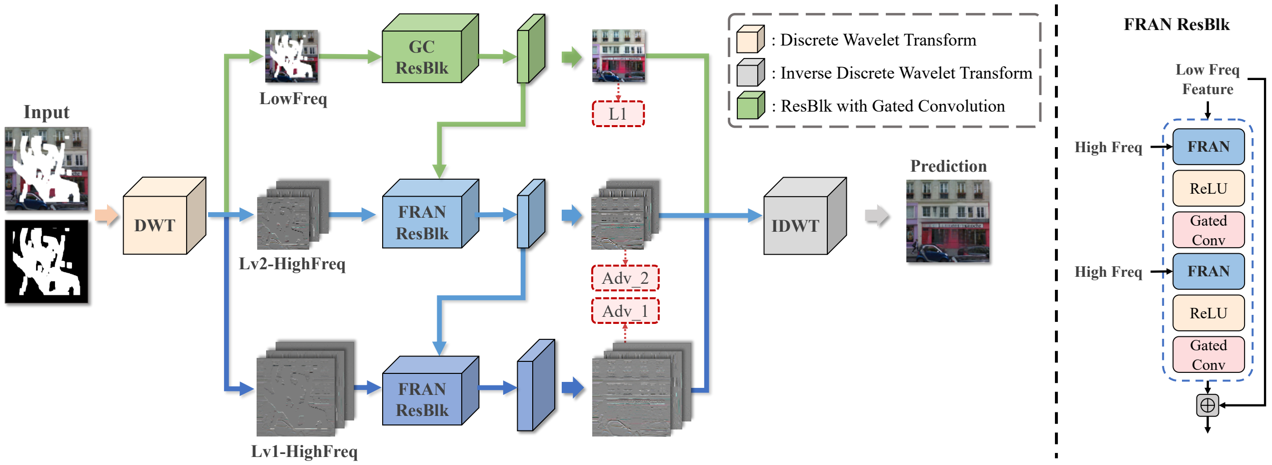


Figure 2. The architecture of the proposed WaveFill inpainting network: The WaveFill generator consists of three branches for processing information of different frequencies separately. Given an *Input Image*, we first decompose it into multiple frequency bands via *DWT* (discrete wavelet transform) and then assemble the decomposed frequency bands into three 'broadbands' *LowFreq*, *Lv2-HighFreq* and *Lv1-HighFreq*. The completion is performed in the *LowFreq* first with *GC ResBlk* and the generated features are then aligned and propagated to high-frequency branches (via a novel normalization scheme *FRAN*) for further completion. The L1 loss is explicitly applied in the low-frequency branch, and two discriminators are employed to enable adversarial training in the two high-frequency branches. The generated features in the three branches are finally transformed back to spatial domain via *IDWT* (inverse DWT) to produce the final *Prediction*.

been explored in various computer vision tasks such as super-resolution [13, 8], style transfer [23], quality enhancement [33] and image demoiréing [22]. Different from them that directly concatenate frequency bands and pass them to convolutional layers, we design separate network branches to explicitly generate contents for each group of frequency bands, and meanwhile incorporate features from other branches for better completion.

3. Proposed Method

3.1. Overview

The overview of our proposed inpainting network is illustrated in Fig. 2. An input image is first decomposed and assembled into 3 frequency bands *LowFreq*, *Lv2HighFreq* and *Lv1HighFreq*, which are then fed to three network branches for respective completion. We apply L1 reconstruction loss to *LowFreq* and adversarial loss to *Lv2HighFreq* and *Lv1HighFreq* to mitigate the inter-frequency conflicts. In addition, we design a novel normalization scheme *FRAN* that aligns and fuses features from the three branches to enforce the completion consistency across the three frequency bands. The generation results in the three branches are finally transformed back to the spatial domain to complete the inpainting, more details to be described in the ensuing subsections.

3.2. Wavelet Decomposition

The key innovation of our work is to disentangle images into multiple frequency bands and complete the images in different bands separately in the wavelet domain. We adopt 2D discrete wavelet transform (DWT) to first decompose images into multiple wavelet sub-bands with different frequency contents. For each iteration of the decomposition, the DWT applies low-pass and high-pass wavelet filters alternatively along image columns and rows (followed by downsampling), which produces 4 sub-bands including LL, LH, HL, and HH. The decomposition continues iteratively on LL^{n-1} to produce LL^n, LH^n, HL^n, and HH^n until the target level of decomposition N_w is reached. Hence, a total number of $3N_w + 1$ wavelet sub-bands will be finally produced including $LL^{N_w}, \{LH^n\}_{n=1}^{N_w}, \{HL^n\}_{n=1}^{N_w}$, and $\{HH^n\}_{n=1}^{N_w}$. Here LL^{N_w} captures low-frequency information at the N_w-th level, LH^n, HL^n and HH^n capture the horizontal, vertical and diagonal high-frequency information at the n-th level, respectively. Note that the sizes of sub-bands at the n-th level are down-sampled with a factor of $1/2^n$.

In this work, we adopt the Haar wavelet filter as the basis for the wavelet transform, where the high-pass filter is $h_{high} = (1/\sqrt{2}, 1/\sqrt{2})$ and the low-pass filter is $h_{low} = (1/\sqrt{2}, -1/\sqrt{2})$. The level of wavelet decomposition N_w is empirically set to 2, we treat LL^2 as low-frequency, concatenate LH^n, HL^n and HH^n in the channel dimension as n-th level high-frequency. Given a input image of size $H \times W \times 3$, we will thus obtain three inputs in the wavelet domain, namely, *LowFreq* with size of $H/4 \times W/4 \times 3$, *Lv2-HighFreq* with size of $H/4 \times W/4 \times 9$ and *Lv1-HighFreq* with size of $H/2 \times W/2 \times 9$.

3.3. Frequency Region Attentive Normalization

It is a vital step to align and fuse the low-frequency and high-frequency features for generating consistent and realistic contents across different frequency bands.An effective fusion of low-frequency and high-frequency features has two major challenges. First, the statistics of low-frequency and high-frequency bands have clear differences, direct summing or concatenating them could greatly suppress high-frequency information due to its high sparsity. Second, the different branches are trained with their explicit loss terms, and the learning capacity (No. of CNN layers and kernel sizes) also varies among the branches. Thus, when inpainting different branches independently without inter-branch alignment, a network branch may generate contents that are reasonable in its own frequency bands but inconsistent across frequency bands of other branches (in object shapes or sizes). Both issues could lead to various blurs and artifacts in the completion results. We design a novel Frequency Region Attentive Normalization (FRAN) technique that aligns and fuses low-frequency and high-frequency features for more realistic inpainting.

For the issue with the statistical difference, we propose to align the low-frequency features with the target high-frequency features so as to fuse them effectively and alleviate the difficulty of generating target high-frequency bands. Inspired by the spatially-adaptive normalization (SPADE) [27], we achieve the feature alignment by injecting the learnable modulation parameters $\boldsymbol{\gamma}_H$ and $\boldsymbol{\beta}_H$ of high-frequency features $X_H = \{x_H^1, ..., x_H^N\}$ to the low-frequency features $X_L = \{x_L^1, ..., x_L^N\}$, where N is the number of spatial positions, i.e. $N = H \times W$.

To align the contents in the missing regions, we aggregate the self-attention score of low-frequency features to high-frequency features. Since the attention map depicts the correlation between low-frequency feature patches, the misaligned high-frequency features of corrupted regions can be reconstructed by collectively aggregating features from uncorrupted regions. Another advantage of applying attention aggregation is to leverage complementary features of distant regions by establishing long-range dependencies. As shown in Fig. 3, the attention scores $W_{j,i}$ are computed from low-frequency features $X_L \in R^{C\times N}$ (C is the channel number) which are firstly transformed to two features space for key and query respectively, i.e. $K = f(X_L), Q = g(X_L)$, f and g are the 1×1 convolutions. For efficiency, we employ max-pooling to obtain a spatial dimension of $N = 1024$ (32×32) for attention calculation and aggregation.

$$W_{j,i} = \frac{\exp(s_{i,j})}{\sum_{i=1}^{N}\exp(s_{i,j})}, \text{ where } s_{i,j} = f(x_L^i)^T g(x_L^j) \ . \quad (1)$$

The high-frequency features X_H is then mapped to the feature space with the same hidden dimension by $V = h(X_H)$ where h is the transformation function by convolution. The aggregation of X_H at position i is defined by:

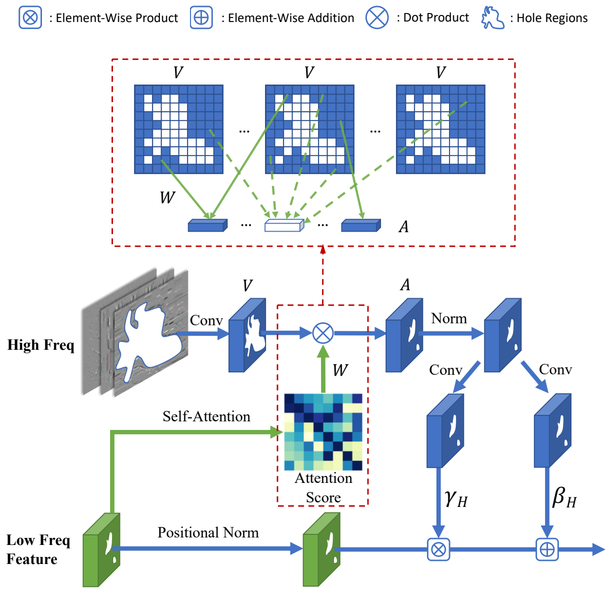


Figure 3. The structure of the proposed Frequency Region Attentive Normalization (FRAN): The irregular regions in each feature map denote corrupted regions. After projecting high-frequency information to a feature space, FRAN aligns the corrupted regions of low-frequency and high-frequency features by aggregating the attention score of low-frequency to high-frequency. The aligned high-frequency features are then convolved to produce modulation parameters $\boldsymbol{\gamma}_H$ and $\boldsymbol{\beta}_H$ that are injected into the normalized low-frequency features.

$$A_i = \sum_{j=1}^{N} W_{j,i} h(x_H^i) \ . \quad (2)$$

Since the high-frequency features are significantly sparse, the magnitude of resultant aggregation is relatively small. We adopt a parameter-free positional normalization [19] to normalize it and meanwhile preserve structure information. The same normalization is also applied to low-frequency features before the modulation. Finally, the aggregation output A is convolved to produce the modulation parameters $\boldsymbol{\gamma}_H$ and $\boldsymbol{\beta}_H$ to modulate the normalized low-frequency features:

$$H = \boldsymbol{\gamma}_H \frac{X_L - \mu_L}{\sigma_L} + \boldsymbol{\beta}_H \ , \quad (3)$$

where H is the modulated features, μ_L and σ_L is the mean and standard deviation of X_L along the channel dimension.

3.4. Network Architecture

Our network consists of one generator and 2 discriminator as illustrated in Fig. 2.

Generation Network. The generation network consists of 3 branches *LowFreq*, *Lv2-HighFreq* and *Lv1-HighFreq* that recover corrupted regions separately. The *LowFreq* branch consists of a completion module *GC ResBlk* that adopts gated convolution [40] and residual connection [11]. Specifically, *GC ResBlk* consists of several consecutive residual blocks with growing dilation rates up to 16 to increase the receptive field. Meanwhile, it replaces all convolutions by gated convolution to dynamically handle missing regions. The generated low-frequency features will be propagated to a decoder that has two gated convolutions to predict the completion of low-frequency sub-bands. Besides, they will also be transferred to two high-frequency branches for guiding and aligning with their generation.

The high-frequency branch *Lv2-HighFreq* consists of a new residual block *FRAN ResBlk* that is introduced with FRAN as illustrated in Fig. 2 (right). As the learned modulation parameters have encoded high-frequency information, we directly feed the high-frequency bands to the FRAN without additional encoding. After injecting the high-frequency information to low-frequency features, we propagate the acquired high-frequency features to a separate decoder which also consists of two gated convolutions. Another high-frequency branch *Lv1-HighFreq* shares similar structures with *Lv2-HighFreq*, except that it concatenates the well-aligned and normalized features from the previous two branches and up-sampling them to the current spatial dimension. The generation network thus predicts the inpainting of all 3 frequency bands, and finally converts them back to the spatial domain via inverse Discrete Wavelet Transform (IDWT). As DWT and IDWT are both differentiable, the network can be trained end-to-end.

Discrimination Network. To synthesize high-frequency information, we adopt two discriminators of the same structure to predice *Lv2-HighFreq* and *Lv1-HighFreq*, respectively. Motivated by PatchGAN [15] and global and local GANs [14], we adopt global and local sub-networks on top of PatchGAN to ensure the generation consistency. Additionally, we append a self-attention layer [47] after the last convolutional layer to assess the global structure and enforce the geometric consistency.

3.5. Loss Functions

We denote the finally completed image by I_{out}, the predictions in the wavelet domain by $\{L_{out}^{N_w}, H_{out}^1, ..., H_{out}^{N_w}\}$ (N_w is number of levels in wavelet decomposition), the ground-truth image by I_{gt} and its corresponding wavelet coefficients by $\{L_{gt}^{N_w}, H_{gt}^1, ..., H_{gt}^{N_w}\}$. D_n is the discriminator for the n-th level high-frequency wavelet coefficients in the wavelet domain.

Low-Frequency L1 Loss. We explicitly employ the L1 loss on the low-frequency subbands in the wavelet domain, which can be defined by:

$$\mathcal{L}_{LF} = ||L_{out}^{N_w} - L_{gt}^{N_w}||_1 \ . \tag{4}$$

Adversarial Loss. For the 2 discriminators of high-frequency branches, we apply the same adversarial losses to them using hinge loss [15]. The adversarial loss for discriminator D_n is defined as:

$$\begin{aligned}\mathcal{L}_{D_n} &= \mathbb{E}_{H_{gt}^n}[\mathit{ReLU}(1 - D_n(H_{gt}^n)] \\ &\quad + \mathbb{E}_{H_{out}^n}[\mathit{ReLU}(1 + D_n(H_{out}^n)] \ .\end{aligned} \tag{5}$$

For the generator, we sum up the adversarial loss of each discriminator to obtain the final loss $\mathcal{L}_G$ as below:

$$\mathcal{L}_G = -\sum_{n=1}^{N_w} \mathbb{E}_{H_{out}^n}[D_n(H_{out}^n)] \ . \tag{6}$$

Feature Matching Loss. As the training could be unstable due to the sparsity of high-frequency bands, we adopt the feature matching loss following pix2pixHD [34] on both the two discriminators to stabilize the training process.

$$\mathcal{L}_{FM} = \sum_{n=1}^{N_w} \mathbb{E}\left[\sum_{i=1}^{L} \frac{1}{N_i}||D_n^i(H_{out}^n) - D_n^i(H_{gt}^n)||_1\right] , \tag{7}$$

where L is the last layer of the discriminator, D^i and N_i are the activation map and its number of elements in the i-th layer of the discriminator, respectively.

Perceptual Loss. To penalize the perceptual and semantic discrepancy, we employ the perceptual loss [16] using a pertrained VGG-19 network:

$$\begin{aligned}\mathcal{L}_{perc} = &\sum_i \lambda_i ||\Phi_i(I_{out}) - \Phi_i(I_{gt})||_1 \\ &+ \lambda_l ||\Phi_l(I_{out}) - \Phi_l(I_{gt})||_2 \ ,\end{aligned} \tag{8}$$

where λ_i are the balancing weights. Φ_i is the activation of i-th layer of the VGG-19 model which corresponds to the activation maps from layers *relu1_2*, *relu2_2*, *relu3_2*, *relu4_2* and *relu5_2*. Φ_l represents the activation maps of *relu4_2* layer, and we select this specific layer to emphasize the high-level semantics.

Full Objective. With the linear combination of the aforementioned losses, the network is optimized by the following objective:

$$\mathcal{L}_\theta = \min_G \max_{D_1, D_2} (\lambda_l \mathcal{L}_{LF} + \mathcal{L}_G + \lambda_f \mathcal{L}_{FM} + \lambda_p \mathcal{L}_{perc}) \ , \tag{9}$$

where we empirically set $\lambda_l = 2$, $\lambda_f = 5$, and $\lambda_p = 10$ in our experiments for balancing the objectives.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct experiments on three public datasets that have different characteristics:

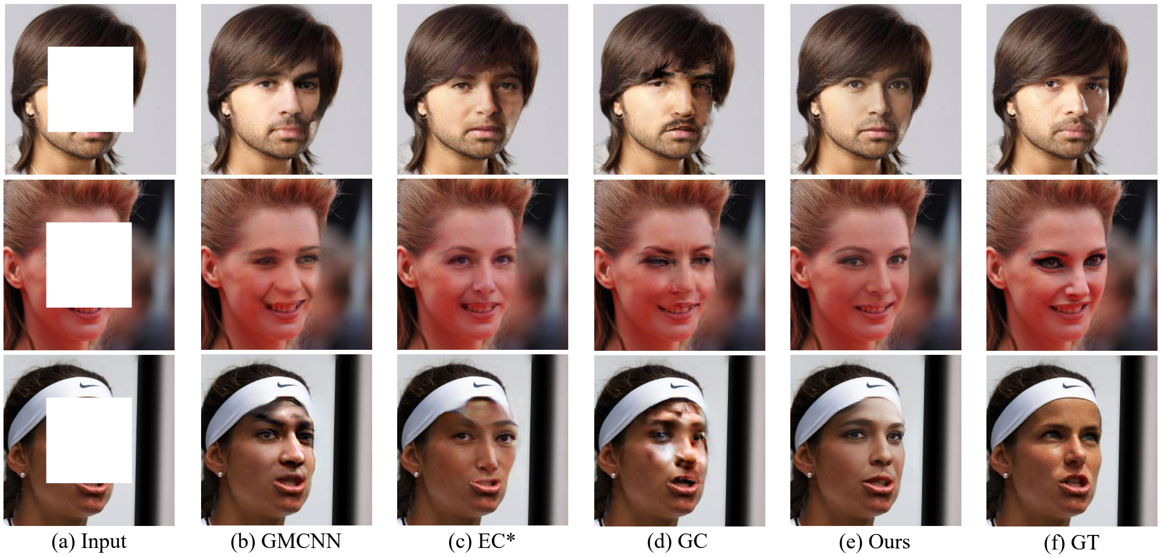


Figure 4. Qualitative comparison of WaveFill with the state-of-the-art: WaveFill generates more realistic inpainting with much less artifacts (over the dataset CelebA-HQ[17] with central square masks). $*$ means that the model is trained with official implementation.

- CelebA-HQ [17]: It is a high-quality version of the human face dataset CelebA [24] with 30,000 aligned face images. We follow the split in [40] that produces 28,000 training images and 2,000 validation images.
- Places2 [49]: It consists of more than 1.8M natural images of 365 different scenes. We randomly sampled 10,000 images from the validation set in evaluations.
- Paris StreetView [28]: It is a collection of street view images in Paris, which contains 14,900 training images and 100 validation images.

Compared Methods. We compare our method with a number of state-of-the-art methods as listed:

- GMCNN [35]: It is a generative model with different receptive fields in different branches.
- GC [40]: It is also known as DeepFill v2, a two-stage method that leverages gated convolution.
- EC [26]: It is a two-stage method that first predicts salient edges to guide the generation.
- MEDFE [21]: It is a mutual encoder-decoder that treats features from deep and shallow layers as structures and textures of an input image.

Evaluation Metrics. We perform evaluations by using four widely adopted evaluation metrics: 1) Fréchet Inception Score (FID) [12] that evaluates the perceptual quality by measuring the distribution distance between the synthesized images and real images; 2) mean ℓ_1 error; 3) peak signal-to-noise ratio (PSNR); and 4) structural similarity index (SSIM) [36] with a window size of 51.

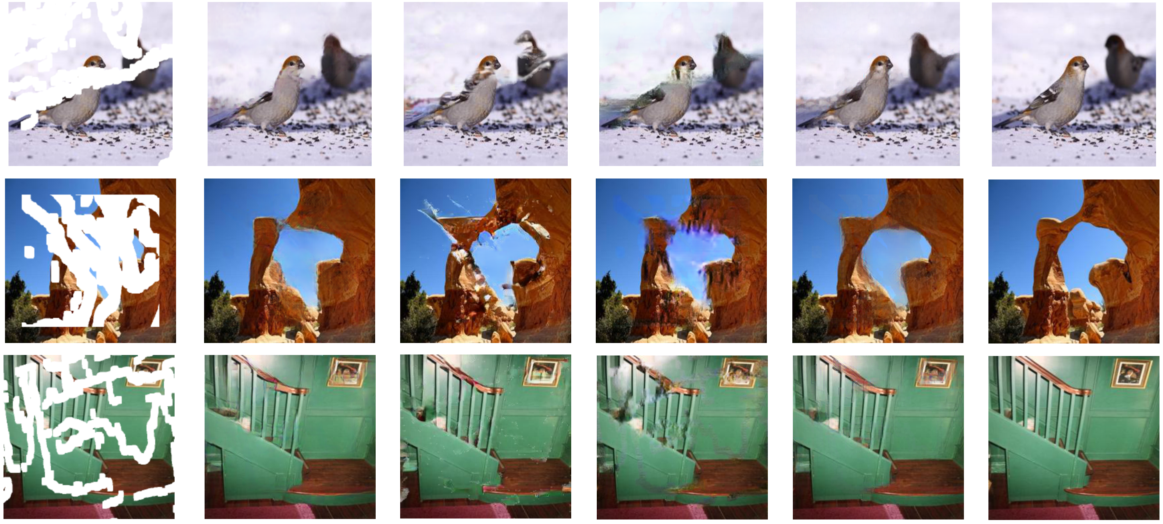
Implementation Details. The proposed method is implemented in PyTorch. The network is trained using 256 × 256 images with random rectangle masks or irregular masks [20]. We use Adam optimizer [18] with $\beta_1 = 0$ and $\beta_2 = 0.9$, and set the learning rate at 1e-4 and 4e-4 for the generator and discriminators, respectively. The experiments are conducted on 4 NVIDIA(R) Tesla(R) V100 GPU. The inference is performed in a single GPU, and our full model runs at 0.138 seconds per 256 × 256 image.

	GMCNN [35]	EC$*$ [26]	GC [40]	**Ours**
FID$\downarrow$	8.17	8.04	7.39	**6.48**
$\ell_1(\%)\downarrow$	2.38	2.31	2.53	**2.26**
PSNR$\uparrow$	25.86	25.64	25.37	**26.53**
SSIM$\uparrow$	0.905	0.896	0.894	**0.911**

Table 1. Quantitative comparison of WaveFill with state-of-the-art methods over CelebA-HQ [17] validation images (2,000) with square masks. $*$ denotes that we trained the model based on official implementation.

4.2. Quantitative Evaluation

We perform extensive quantitative evaluations over data with central square masks and irregular masks [20]. For inpainting with central square masks, we use the mask size of 128 × 128, and benchmark with GMCNN [35], EC [26] and GC [40] over the validation images of CelebA-HQ [17]. For inpainting with irregular masks, we conducted experiments over Places2 [49] and Paris StreetView [27] and benchmarked with GC [40], EC [26] and MEDFE [21]. The irregular masks in the experiments are categorized based the ratios of the masked regions over the image size. Performance of the compared methods was acquired by running publicly available pre-trained. The only exception is EC [26] which was trained with the official implementation on



(a) Input (b) EC (c) GC (d) MEDFE (e) Ours (f) GT

Figure 5. Qualitative comparison of WaveFill with the state-of-the-art: WaveFill generates more realistic inpainting with much less artifacts as compared with the state-of-the-art (over dataset Places2 [49] with irregular masks).

CelebA-HQ [17] with random rectangle masks.

Table 1 shows experimental results for dataset CelebA-HQ with central square masks. It can be observed that WaveFill outperforms all existing methods under different evaluation metrics consistently. In addition, experiments with irregular masks show that WaveFill achieves superior inpainting under different mask ratios as shown in Table 2. The effectiveness of WaveFill largely attributes to the wavelet-based frequency decomposition and the proposed normalization scheme. Specifically, disentangling frequency information in the wavelet domain helps mitigate the conflicts in generating low-frequency and high-frequency contents effectively, and it improves the inpainting quality in PSNR and SSIM as well. With the proposed normalization scheme, the low and high frequency information can be aligned for consistent generations in different frequency bands. Moreover, it allows the model to establish long-range dependencies which help generate more semantically plausible contents with better perceptual quality in FID. Quantitative results for Paris StreetView [27] are provided in the supplementary materials due to space limit.

	Mask	EC [26]	GC [40]	MEDFE [21]	**Ours**
FID↓	10-20%	2.55	5.18	2.81	**1.96**
	20-30%	5.36	10.06	7.51	**4.08**
	30-40%	9.28	15.67	15.84	**7.33**
	40-50%	15.17	22.69	28.98	**12.68**
$\ell_1(\%)\downarrow$	10-20%	1.55	2.19	1.42	**1.39**
	20-30%	2.71	3.73	2.62	**2.32**
	30-40%	3.97	5.34	4.13	**3.42**
	40-50%	5.42	7.05	5.97	**4.73**
PSNR↑	10-20%	27.23	24.96	28.48	**28.72**
	20-30%	24.30	22.02	24.76	**25.87**
	30-40%	22.31	20.03	22.05	**23.74**
	40-50%	20.67	18.54	19.87	**21.99**
SSIM↑	10-20%	0.942	0.906	0.954	**0.956**
	20-30%	0.890	0.833	0.902	**0.918**
	30-40%	0.830	0.758	0.833	**0.867**
	40-50%	0.758	0.679	0.749	**0.803**

Table 2. Quantitative comparison of WaveFill with state-of-the-art methods over Places2 [49] validation images (10,000) with irregular masks [20].

4.3. Qualitative Evaluations

Figs. 4 and 5 show qualitative experimental results over the validation set of CelebA-HQ [17] and Places2 [49], respectively. As demonstrated in Fig. 4, the inpainting by GMCNN [35] and EC [26] suffers from unreasonable semantics and inconsistency near edge regions clearly, while the inpainting by GC [40] contains obvious artifacts and blurry textures. As a comparison, the inpainting by WaveFill are more semantically reasonable and has less artifacts but more texture details. For dataset Places2 [49], the inpainting by GC [40] and MEDFE [21] contains undesired artifacts and distorted structures as shown in Figs. 5b and 5c. Though EC [26] produces more visually appealing contents with less artifacts, its generated semantics are still short of plausibility. Thanks to the frequency disentanglement and FRAN, WaveFill achieves superior inpainting for both central square masks and irregular masks.

4.4. User Study

We performed user studies over datasets Paris StreetView[28], Places2[49] and CelebA-HQ[17]. Specifically, we randomly sampled 25 test images from each

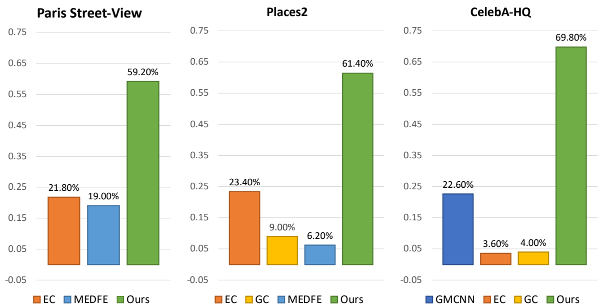


Figure 6. Inpainting evaluation by user study: The percentages tell the amount of images that are judged as the most realistic among images inpainted by all compared methods.

Models	FID↓	$\ell_1(\%)\downarrow$	PSNR↑	SSIM↑
Spatial + Concat	33.95	2.45	28.37	0.898
DCT + Concat	100.93	4.71	23.56	0.765
Wavelet + Concat	32.73	2.46	28.46	0.899
Wavelet + SPADE	32.14	2.38	28.81	0.901
Wavelet + FRAN	**31.02**	**2.34**	**28.94**	**0.904**

Table 3. Ablation study of WaveFill over Paris StreetView [27] validation set (100) with irregular masks [20]. Model in the last row is the standard WaveFill.

dataset with no idea of inpainting results, which leads to 75 multiple choice questions in the survey. We recruited 20 volunteers with image processing backgrounds and each subject is asked to vote for the most realistic inpainting in each question. As Fig. 6 shows, the proposed WaveFill outperforms state-of-the-art methods by large margins.

4.5. Ablation Study

We study the individual contributions of our technical designs by several ablation studies over Paris StreetView [27] as shown in Table 3. In the ablation studies, we trained four network models including: 1) *Spatial + Concat* (Baseline) that adopts the typical encoder-decoder network with gated convolution [40]. Different from WaveFill, L1 and adversarial losses are applied together, multi-level features are directly concatenated; 2) *DCT + Concat* that adopts discrete cosine transform (DCT) to compare with wavelet transformation. Similar to WaveFill, we split the frequency bands into three groups and feed them to the three generation branches; 3) *Wavelet + Concat* that replaces FRAN by concatenation of multi-frequency features; 4) *Wavelet + SPADE* that replace FRAN by SPADE [27].

As shown in Table 3, using DCT degrades the inpainting greatly due to the lack of spatial information. Wavelet transformation preserves spatial information which improves inpainting by large margins. In addition, using wavelet outperforms the baseline especially in FID, largely because wavelet-based model disentangles multi-frequency information and recovers corrupted regions in different frequency bands separately. Visual evaluation is well aligned with quantitative experiments in Fig. 7. We can see that DCT-based model fails to synthesize meaningful structures as shown in (c). Spatial-based model instead introduces unreasonable semantics and clear artifacts as shown in (b). Our wavelet-based model fills the missing regions with much less artifacts as shown in (d). Further, concatenation and SPADE do not align the features of different frequencies for better content consistency. FRAN addresses this issue effectively as shown in Table 3 and Fig. 7. More ablation studies are included in the supplementary materials.

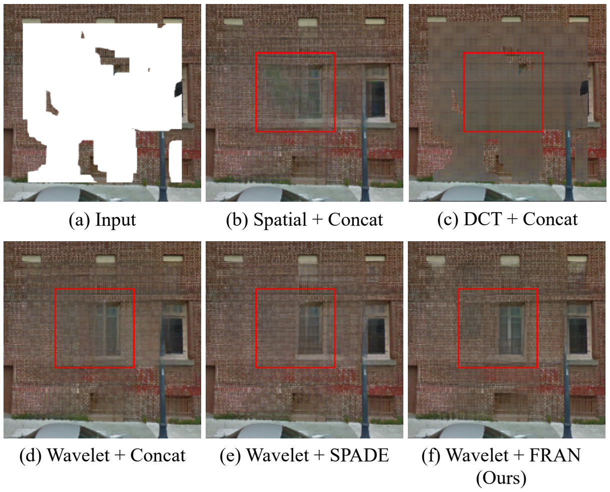

(a) Input
(b) Spatial + Concat
(c) DCT + Concat
(d) Wavelet + Concat
(e) Wavelet + SPADE
(f) Wavelet + FRAN (Ours)

Figure 7. Ablation study of the proposed WaveFill: Our 'Wavelet' and 'FRAN' designs both help suppress artifacts and synthesize plausible semantics effectively. The study is performed over Paris StreeView [28] with irregular masks. The red boxes are used to highlight the main differences across different approaches.

5. Conclusion

This paper presents WaveFill, a novel image inpainting framework that disentangles low and high frequency information in the wavelet domain and fills the corrupted regions explicitly and separately. To ensure the inpainting consistency across multiple frequency bands, we propose a novel frequency region attentive normalization (FRAN) that effectively aligns and fuses the multi-frequency features especially those within the missing regions. Extensive experiments show that WaveFill achieves superior image inpainting for both rectangle and free-form masks. Moving forward, we will study how to adapt the idea of wavelet decomposition and separate processing in different frequency bands to other image recovery and generation tasks.

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8):1200–1211, 2001. 2

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (TOG)*, 28(3):24, 2009. 2

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 2

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1

[5] Tony F Chan, Jianhong Shen, and Hao-Min Zhou. Total variation wavelet inpainting. *Journal of Mathematical imaging and Vision*, 25(1):107–125, 2006. 2

[6] Fergal Cotter. *Uses of Complex Wavelets in Deep Convolutional Neural Networks*. PhD thesis, University of Cambridge, 2020. 2

[7] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 2

[8] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *International Conference on Computer Vision*, pages 3076–3085, 2019. 1, 3

[9] Julia A Dobrosotskaya and Andrea L Bertozzi. A wavelet-laplace variational technique for image deconvolution and inpainting. *IEEE Transactions on Image Processing*, 17(5):657–663, 2008. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[12] 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. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 6

[13] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1689–1697, 2017. 3

[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 5

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 5

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018. 6, 7

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pages 1622–1634, 2019. 4

[20] 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 *European Conference on Computer Vision*, pages 85–100, 2018. 2, 6, 7, 8

[21] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, 2020. 2, 6, 7

[22] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Ales Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *European Conference on Computer Vision*, 2020. 3

[23] Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11877–11886, 2019. 3

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, pages 3730–3738, 2015. 6

[25] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999. 2

[26] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2, 6, 7

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 4, 6, 7, 8

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 1, 2, 6, 7, 8

[29] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *International Conference on Computer Vision*, pages 181–190, 2019. 2

[30] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 1

[31] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *International Conference on Computer Vision*, pages 4491–4500, 2017. 1

[32] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In *European Conference on Computer Vision*, pages 3–19, 2018. 2

[33] Jianyi Wang, Xin Deng, Mai Xu, Congyong Chen, and Yuhang Song. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video. In *European Conference on Computer Vision*, 2020. 3

[34] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 5

[35] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 331–340, 2018. 1, 2, 6, 7

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[37] Rongliang Wu and Shijian Lu. Leed: Label-free expression editing via disentanglement. In *European Conference on Computer Vision*, pages 781–798. Springer, 2020. 2

[38] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020. 2

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 2

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 2, 5, 6, 7, 8

[41] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[42] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019. 2

[43] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[44] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Kaiwen Cui, Aoran Xiao, Shijian Lu, and Ling Shao. Bi-level feature alignment for versatile image translation and manipulation. *arXiv preprint arXiv:2107.03021*, 2021. 2

[45] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 2

[46] Hongying Zhang and Shimei Dai. Image inpainting based on wavelet decomposition. *Procedia Engineering*, 29:3674–3678, 2012. 2

[47] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 5

[48] Xiaoqun Zhang and Tony F Chan. Wavelet inpainting by nonlocal total variation. *Inverse Problems & Imaging*, 4(1):191, 2010. 2

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 6, 7