Image Completion with Heterogeneously Filtered Spectral Hints

Xingqian Xu[1,3], Shant Navasardyan[3], Vahram Tadevosyan[3], Andranik Sargsyan[3],
Yadong Mu[2], and Humphrey Shi[1,3]

[1]SHI Lab @ UIUC & UO, [2]Peking University, [3]Picsart AI Research (PAIR)

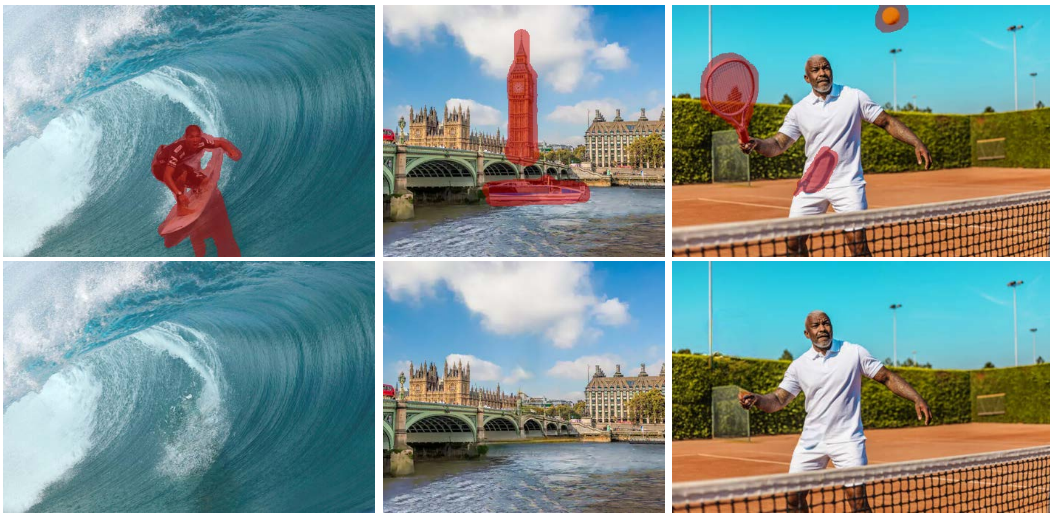


Figure 1: Results of SH-GAN on a variety of inpainting cases. SH-GAN can fill-in image content with outstanding consistency.

Abstract

*Image completion with large-scale free-form missing regions is one of the most challenging tasks for the computer vision community. While researchers pursue better solutions, drawbacks such as pattern unawareness, blurry textures, and structure distortion remain noticeable, and thus leave space for improvement. To overcome these challenges, we propose a new StyleGAN-based image completion network, Spectral Hint GAN (**SH-GAN**), inside which a carefully designed spectral processing module, Spectral Hint Unit, is introduced. We also propose two novel 2D spectral processing strategies, Heterogeneous Filtering and Gaussian Split that well-fit modern deep learning models and may further be extended to other tasks. From our inclusive experiments, we demonstrate that our model can reach FID scores of 3.4134 and 7.0277 on the benchmark datasets FFHQ and Places2, and therefore outperforms prior works and reaches a new state-of-the-art. We also prove the effectiveness of our design via ablation studies, from which one may notice that the aforementioned challenges,* i.e. *pattern unawareness, blurry textures, and structure distortion, can be noticeably resolved. Our code will be open-sourced at: https://github.com/SHI-Labs/SH-GAN.*

1. Introduction

Spectral analysis is a well-established research topic and has been intensely studied for decades [7, 8, 32, 35]. Its corresponding downstream techniques in remote sensing, telecommunication, and healthcare significantly affect our modern life. Earlier computer vision techniques largely adopted algorithms such as the Fourier transform [3, 38], wavelet transform [4, 15, 33, 50] and curvelet transform [47] for image denoising, anti-aliasing, restoration, and compression. Ever since the rapid growth of deep learning, spectral analysis on images has fallen from popularity largely due to the fact that the intriguing properties of the 2D frequency domain complicated many solutions for the content-based tasks. Nevertheless, as the progress of deep learning research goes deeper and wider, researchers start to re-focus on the image spectral analysis and its potential applications. Recent works, such as [9, 10, 40, 48, 54, 55, 57, 64], have shown that a network structure with spectral priors can be favorable in many tasks, including classification, segmentation, image synthesis, and super-resolution. These works will undoubtedly guide the future computer vision research in spectral analysis.

Despite the fast development of spectral-related deep learning methods, few works [9, 25, 48] have explored the potential of image completion with spectral priors, in which the major purpose of [9, 25] is still image synthesis. In the past few years, image completion heavily relied on feature extraction using CNNs and similarity-based patch matching techniques [36, 56, 58, 59, 61]. While these strategies have been proved useful in some scenarios, it remains a case-dependent approach due to its frequently shown structural distortion and texture artifacts. Meanwhile, the success of the StyleGAN series [24, 25, 26, 27] on generation tasks has established a robust baseline for many downstream tasks such as style transfer [1, 16, 39], GAN-inverse [37], latent space editing [17, 45, 49], and inpainting [63]. Among them, the recent image completion work CoModGAN [63] introduced the concept of co-modulation and pushed the performance to the next level.

Through our research, we compared several image completion works such as DeepFill [58, 59], LaMa [48], CoModGAN [63], *etc*. We noticed that these methods produce promising results in some cases but struggle in others. For instance, the patch-based approaches keep a good texture consistency for irregular textures with small textons (*e.g*. grass, wood, asphalt, *etc*.), but create large structural distortions, especially when the unknown masks are large. LaMa, on the other hand, utilizes a spectral transform module, namely FFC [10], and stands out in cases with strong pattern-like signals. However, it faces challenges in complex scenes and tends to create fast-repeating artifacts that are conceptually blurry. CoModGAN generates more natural-looking image content. However, CoModGAN is less bounded by image known regions; thus, it may ignore global patterns and create faulty objects.

Motivated by the prior works using spectral transform to handle low-level patterns and modulated generative block to handle high-level semantic, we introduce a novel approach: *Spectral Hint GAN* (SH-GAN) along with a new module: *Spectral Hint Unit* (SHU). Our goal is to minimize the summarized problems, *i.e*., pattern unawareness, blurry textures, and structural distortions and to maintain the natural balance between pattern and semantic consistency. Moreover, we also propose two new spectral transform strategies: *Heterogeneous Filtering*, and *Gaussian Split*. Heterogeneous Filtering aims to manipulate the 2D spectrum using a learnable smooth-varying function correlated with its frequency. Meanwhile, Gaussian Split is a spectral split algorithm that distributes information to different resolution scales for image synthesis. As a result, our FID performances are 3.4134 on FFHQ [26] and 7.0277 on Places2 [65], outperforming CoModGAN and other prior works and reaching a new state-of-the-art. We also demonstrate that our model outperforms LaMa and the other works on the narrow and wide masks using a LaMa-style evaluation scheme. Lastly, we perform ablation studies on SH-GAN, from which we clearly see the performance gain using SHU and the proposed spectral transform strategies, *i.e*. Heterogeneous Filtering and Gaussian Split.

In summary, the main contributions of our work are the following:

- We propose a novel spectral-aware StyleGAN-based image completion network, *Spectral Hint GAN* (SH-GAN), in which a new module, *Spectral Hint Unit* (SHU), is introduced.
- We also bring out two new spectral processing strategies: *Heterogeneous Filtering* and *Gaussian Split*. These strategies aim to enhance the 2D spectral signals and hierarchically fuse them inside the synthesis network.
- The FID scores of SH-GAN outperform the state-of-the-art on two popular benchmark datasets: FFHQ and Places2. Meanwhile, we perform inclusive studies, through which we demonstrate the effectiveness of our new designs.

2. Related Works

2.1. Spectral Research in Computer Vision

Early image spectral research largely focused on low-level vision such as compression [2, 3], restoration [50], and denoising [4, 15, 47, 51]. In 1971, Tsai and Huang proposed Transform Coding [3] using the discrete cosine transform, which was later extended into the well-known JPEG format [2]. Huang [50] also initiated the pioneering work for image enhancement and restoration using multi-frame discrete Fourier transform (DFT) and inverse-DFT. Popular approaches of image denoising utilize fast Fourier transform (FFT) [51] or wavelet transform [4, 15]. In [47], Jean *et al*. proposed two new frequency-domain tools: ridgelet transform and curvelet transform, by which they could recover images from noise with higher perceptual quality than prior works. In recent years, researchers have shown increasing enthusiasm for spectral neural networks. [43] was one of the first works that combined spectral layers with CNN, in which it proposed spectral pooling for dimension reduction. Another work [55] proposed SyncSpecCNN in which a set of 3D features were eigen-decomposed and passed through a CNN for 3D part segmentation. For super-resolution, [64] decomposed tensors with the wavelet basis and transformed them with fully connected layers. [40] explored the inductive bias of CNNs towards low-frequency signals. A similar discovery was mentioned in [9], in which the author performed analysis on the frequency domain of a GAN network. Recently, Chi *et al*. proposed Fast Fourier Convolution (FFC) [10], in which tensors were converted between spatial and frequency domain using FFT and inverse-FFT. Chi *et al*. also showed that FFC could substitute regular

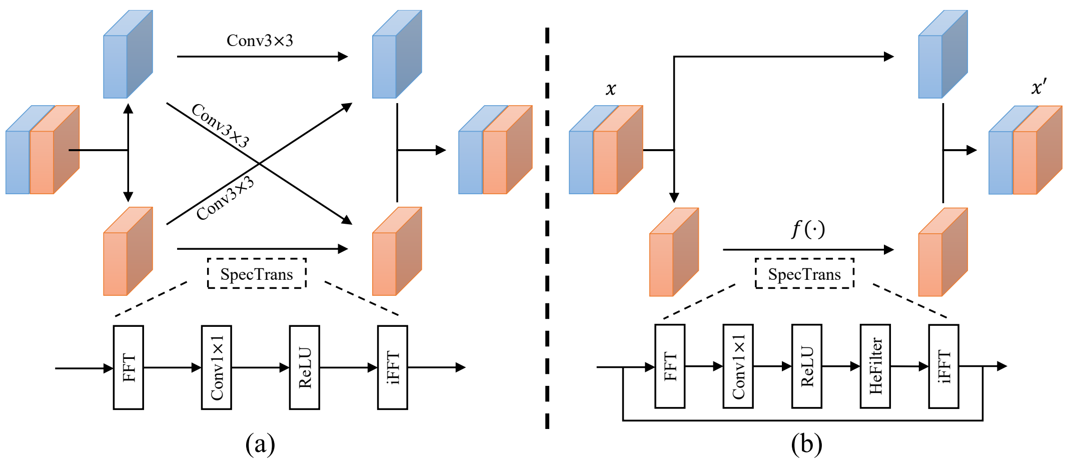


Figure 2: This figure shows the structure of FFC [10] (left) and SHU (right). Different from FFC, our SHU does not use any external convolution layers. In the spectral transform, SHU utilizes HeFilter to transform spectral tensors after ReLU, while FFC directly connects ReLU outputs to iFFT. To make things compacted, we do not include Gaussian Split in this figure.

Residual Blocks [18] and reach better performance in classification.

2.2. Image Completion

The goal of image completion is to synthesize image content for missing regions. Traditional approaches performed gray-level gradient extension [5], image quilting [13], and patch-based methods [6, 12, 14]. Notwithstanding the success of these approaches in the cases of simple and highly-textured backgrounds, they fail to recover missing semantics and complex structures. Since the popularity of deep learning, [29, 42, 53] were the first groups to design deep network architectures for inpainting. Satoshi *et al.* [21] utilized dilated convolutions and adversarial training. [30] inpainted face images using semantic maps as guidance. To address the vanilla convolutions' drawback of treating all pixels equally, Liu *et al.* [31] introduced partial convolutions in which unknown elements in the input tensor are excluded from the calculations. Yu *et al.* further improved the performance with contextual attention [58] and gated convolutions [59]. Navasardyan & Ohanyan [36] proposed the onion convolutions, in which neighboring patches could be searched and relocated simultaneously with the convolution operation. Sharing a similar spirit with [6, 58], HiFill [56] generated contextual residual to fill in higher resolution textures. Another work, CR-Fill [61] proposed the CR loss, a patch similarity loss designed to reinforce contextual consistency. Zhu *et al.* [66] introduced the MADF module and cascaded refinement decoders. Zeng *et al.* [60] introduced the AOT block and utilized soft mask-guided PatchGAN [22] for the network training. Suvorov *et al.* introduced LaMa, which is a U-shape structure with FFCs [10]. Zhao *et al.* [63] proposed CoModGAN along with the co-modulation idea on top of the StyleGAN [24, 26, 27]. All these works have achieved plausible results on faces and natural images using free-form masks. The concurrent works in diffusion models [20, 46] can also be extended to inpainting tasks in which LDM [44] and DALL-E2 [41] shows promising results at a higher computational cost during inference time.

3. Method

In this section, we illustrate the key designs of our work, including Spectral Hint Unit (SHU), Heterogeneous Filtering Layer (HeFilter), Gaussian Split, and the overall GAN architecture. The definition of image completion is to restore an RGB image I_{fake} from a masked color image $I = I_{real} \odot M$, where I_{real} is the ground-truth image and M is the mask, in which the known pixels have value 1 and the unknown pixels are represented with zeros.

3.1. Spectral Transform with SHU

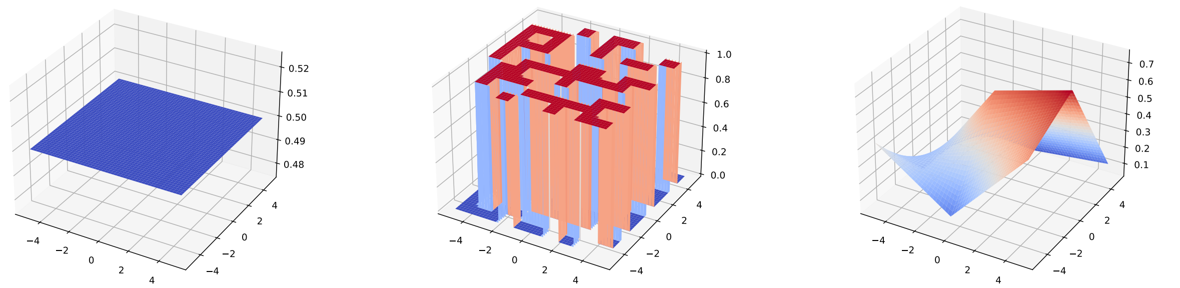
Spectral Hint Unit (SHU) transforms tensors using a neat FFT−Network−iFFT pipeline (see Fig. 2). The recent work FFC [10] also suggested a similar structure in which the author blended spectral transform inside a densely connected convolutional network. Different from FFC, SHU is light-weighted because it has no extra convolutions outside the spectral transform. More precisely, let $x \in \mathbb{R}^{N\times H\times W}$ be an N-channeled tensor with height and width equal to H and W respectively. Then the output of SHU is the tensor x' with the same dimentionality formed by the following way:

$$x' = \text{concat}\left(x_{[0\ldots N-K]},\ x_{[N-K\ldots N]} + f\left(x_{[N-K\ldots N]}\right)\right)$$

$$f = \text{iFFT} \circ g \circ \text{FFT}$$

$$g = \text{HeFilter} \circ \text{ReLU} \circ \text{Conv1} \times 1$$

Our design fits GAN training for the following reasons: a) local operations such as a convolution should be well-



(a) Conv1×1 (b) ReLU (c) HeFilter

Figure 3: Graphic explanations on manipulating spectral tensors using different types of layers. Conv1×1 is a homogeneous operation that treats the complex vector-space equally over the 2D frequency domain. ReLU works as a band-pass filter by changing negative components to zeros. Our HeFilter applies a smooth varying mapping function on the frequency domain and manipulates complex vectors based on their spectral location.

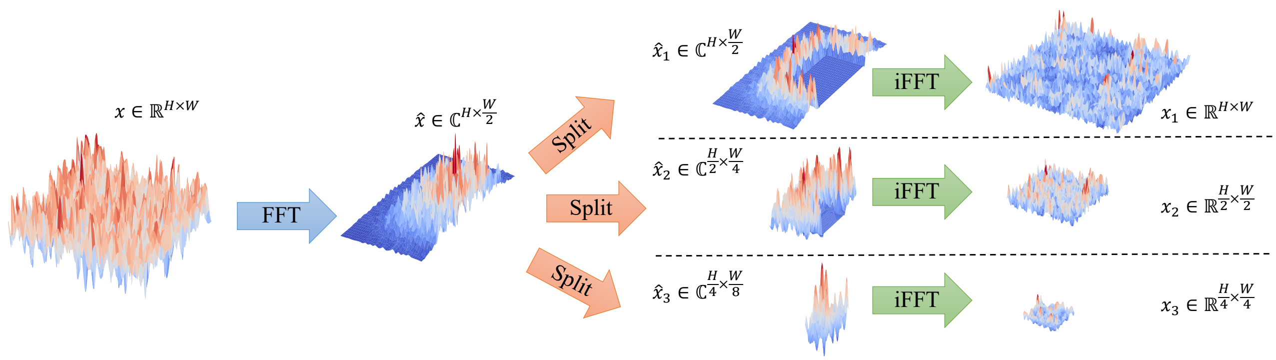


Figure 4: A graphic illustration on the Vanilla Split. Our Gaussian Split is an extended version adding a Gaussian mask on each split level for better anti-aliasing. Due to the limited space, we only show the Vanilla Split here.

handled by the synthesis network with modulated convolutions; b) GAN training should maintain a subtle balance between the spectral and spatial transforms, and it should not overwhelm tensors with spectral information.

Like the spectral transform of FFC [10], Conv1×1 maps between two complex vector-spaces uniformly in the frequency domain. ReLU is the non-linearity operation that filters out all negative components in the vector-space. Lastly, HeFilter performs heterogeneous filtering in which the mapping is a smooth-varying function over the 2D spectral location. We will describe HeFilter with more details in the next subsection. In summary, the spectral transform of SHU is as follows:

a) $\mathbb{R}^{K\times H\times W} \rightarrow \mathbb{C}^{K\times H\times\frac{W}{2}}$ (FFT)
b) $\mathbb{C}^{K\times H\times\frac{W}{2}} \rightarrow \mathbb{C}^{K\times H\times\frac{W}{2}}$ (HeFilter $\circ$ ReLU $\circ$ Conv)
c) $\mathbb{C}^{K\times H\times\frac{W}{2}} \rightarrow \mathbb{R}^{K\times H\times W}$ (iFFT)

3.2. Heterogeneous Filtering and Gaussian Split

As mentioned earlier, one of the contributions of this work is to introduce two novel spectral processing strategies: Heterogeneous Filtering and Gaussian Split, that well-fit the deep learning training scheme.

Heterogeneous Filtering. Recall that FFC [10] transforms spectral tensors with Conv1×1 and ReLU, which can also be viewed as a homogeneous operation and a band-pass filter. In many cases, ReLU is a necessary step but not a recommended end operation for spectral transform with the following reasons: a) it deactivates a frequency band with no recovery; b) it introduces aliasing effects due to non-smoothness; and c) it responses according to magnitude instead of location (*i.e.* requency band). Therefore, we create HeFilter, inside of which a heterogeneous filtering strategy is introduced transforming the complex vector-space via a smooth-varying function over the frequency domain. More precisely, HeFilter learns several weight matrices scattered on an even-spaced 2D frequency domain. During propagation, HeFilter linearly interpolates these weights and multiplies them with the corresponding spectral vector. Figure 3 explains the characteristic of Conv1×1, ReLU, and HeFilter in the spectral transform. We prepare a total of 3×2 weights on the 2D spectral domain. The asymmetry along each dimension is because the FFT of $\mathbb{R}^{K\times H\times W}$ is $\mathbb{C}^{K\times H\times\frac{W}{2}}$, and the skipped half is the reflected complex conjugate. We do not impose constraints on learning these weights, thus HeFilter can be low-pass, band-pass or high-pass depend-

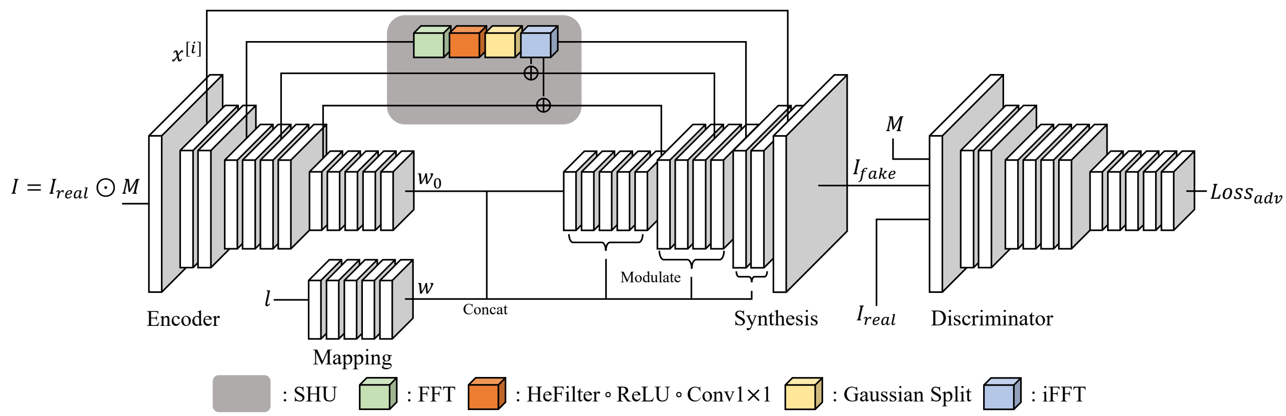


Figure 5: This figure shows the overall structure of our SH-GAN in which our SHU is highlighted in the gray area.

ing on the weights it learns. This also explains the name *heterogeneous* because it only guarantees the discrepancy of the transform on various spectral locations. A simple extension of our HeFilter is to use a larger grid size such as 5×3 or 7×4. Nevertheless, we notice sufficient improvements on model performance using grid size 3×2. HeFilter with larger grid size is left for further studies.

Gaussian Split. Popular generative models [23, 24, 25, 26, 27] adopt the progressive structure in which low-resolution features are gradually elaborated into high-resolution features. We fit our SHU in this structure by adding a unique Gaussian Split, segregating spectral tensors into multi-resolution sub-tensors before the iFFT operation. A fundamental property of the Fourier Transform $\mathcal{F}$ is linearity, in which the FT of the addition of the functions f_1 and f_2 equals the addition of the individual FTs of f_1 and f_2 (see Eq. 1).

$$\begin{aligned} f(x) &\overset{\mathcal{F}}{\longleftrightarrow} \hat{f}(\omega) \\ \alpha f_1(x) + \beta f_2(x) &\overset{\mathcal{F}}{\longleftrightarrow} \alpha \hat{f}_1(\omega) + \beta \hat{f}_2(\omega) \end{aligned} \tag{1}$$

Utilizing the property mentioned above, we can split any spectral signal $\hat{x} = \mathcal{F}(x)$ into several sub-signals $\hat{x}_i, i \in \{1 \ldots n\}$. As long as $\sum_i \hat{x}_i = \hat{x}$, we expect no information loss on $x = \sum_i \mathcal{F}^{-1}(\hat{x}_i)$, and this property holds for FFT on discrete tensors. For convenience, we use the same set of symbols $\hat{x}, \hat{x}_i$ representing spectral tensors. The graphic explanation of the split is highlighted in Figure 4, from which one may notice that our decomposition, like the Wavelet Transform, automatically segregates signals based on their frequency bands. For example, a two-level Vanilla Split is formulated in Eq. 2 and 3, in which we migrate all low frequency values from the large tensor to the small one.

$$\hat{x} \in \mathbb{C}^{H \times \frac{W}{2}} \overset{\text{split}}{\longleftrightarrow} \left(\hat{x_1} \in \mathbb{C}^{H \times \frac{W}{2}}, \hat{x_2} \in \mathbb{C}^{\frac{H}{2} \times \frac{W}{4}} \right) \tag{2}$$

$$\begin{aligned} \hat{x_1}_{[i,j]} &= \begin{cases} \hat{x}_{[i,j]} & (i,j) \notin (\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}) \\ 0 & (i,j) \in (\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}) \end{cases} \quad \text{(Vanilla)} \\ \hat{x_2} &= \hat{x}_{[\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}]} \end{aligned} \tag{3}$$

Our Gaussian Split is an upgraded version of Vanilla Split, in which we smooth each splits with Gaussian weight maps $\mathcal{N}$ to minimize the aliasing effect. (see Eq. 4). The center of $\mathcal{N}$ is at $(\frac{H}{2}, 0)$ and the standard deviation σ is proportional to the corresponding resolution. Since the Fourier transform of a Gaussian function is another Gaussian function, applying Gaussian maps in the frequency domain is equivalent to applying a Gaussian blur filter on the spatial domain that we usually do before downsampling. One may also notice that the multi-level Gaussian Split serves as the well-known Difference of Gaussian (DoG) [34] in the spatial domain. We will demonstrate the effectiveness of our Heterogeneous Filtering and Gaussian Split in section 4.

$$\begin{aligned} \hat{x_1}_{[i,j]} &= \begin{cases} \hat{x}_{[i,j]} & (i,j) \notin (\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}) \\ \hat{x}_{[i,j]} \times (1 - \mathcal{N}_{[i,j]}) & (i,j) \in (\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}) \end{cases} \\ \hat{x_2} &= \hat{x}_{[\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}]} \times \mathcal{N}_{[\frac{H}{4} \ldots \frac{3H}{4}, 0 \ldots \frac{W}{4}]} \\ &\text{(Gaussian)} \end{aligned} \tag{4}$$

3.3. Network Architecture

Similar to CoModGAN [63], our model is a U-shape architecture containing an encoder and a synthesis network. As Figure 5 shows, we first pass the masked input image I into the encoder, inside which I is encoded into a set of feature maps $x^{[i]}$ with resolution i, and a global vector w_0. We then split K channels from $x^{[i]}$ ($K = 32, i = 64$), and pass it to SHU for spectral transform. Inside SHU, we create a series of wavelet-like feature maps $x_i, i \in \{4, ..., 64\}$ using Gaussian Split, in which the low-frequency information is encoded in the low-resolution feature maps (*e.g.* x_4)

Method	FFHQ 256				Places2 256			
	FID(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	FID(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)
CoModGan (small)	5.0184	0.2579	16.31	0.5892	9.5159	0.3995	14.49	0.4914
CoModGan (official)	4.7755	0.2568	16.24	0.5913	9.3621	0.3990	14.50	0.4923
LaMa	32.7035	0.2590	**17.58**	0.6277	23.7409	**0.3679**	16.58	**0.5448**
DeepFillV2	50.9323	0.3204	16.11	0.5569	46.2012	0.4166	14.97	0.4913
CR-Fill	-	-	-	-	40.9690	0.3957	15.28	0.4925
Onion-Conv	-	-	-	-	42.4625	0.4360	15.03	0.5046
MADF	33.6207	0.2800	17.54	**0.6279**	66.2659	0.3889	**16.61**	0.5360
AOT-GAN	73.7962	0.4270	15.60	0.5956	90.6184	0.5139	14.87	0.4790
(ours - small)	4.8225	0.2558	16.36	0.5891	8.3078	0.3969	14.50	0.4918
(ours - regular)	**4.3459**	**0.2542**	16.37	0.5911	**7.5036**	0.3940	14.58	0.4958

Table 1: This table compares the performance of prior models with our SH-GAN on datasets FFHQ and Places2 under resolution 256.

Method	FFHQ 512				Places2 512			
	FID(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	FID(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)
CoModGan (small)	3.9420	0.2497	18.38	0.6911	8.8390	0.3464	15.96	0.5925
CoModGan (official)	3.6996	0.2469	18.46	0.6956	7.9735	0.3420	16.00	0.5953
LaMa	19.5577	0.2871	18.99	**0.7178**	12.6721	**0.3158**	17.12	**0.6521**
DeepFillV2	32.8696	0.3283	18.29	0.6886	29.7345	0.3802	15.91	0.5953
CR-Fill	-	-	-	-	26.6398	0.3593	16.52	0.6038
Onion-Conv	-	-	-	-	25.7480	0.3999	15.02	0.6061
MADF	17.1962	0.2688	**19.62**	0.7196	29.4928	0.3299	**17.77**	0.6239
AOT-GAN	36.1344	0.3403	18.14	0.7131	46.7640	0.3976	16.84	0.6029
(ours - small)	3.7460	0.2491	18.36	0.6897	7.6122	0.3455	15.95	0.5926
(ours - regular)	**3.4134**	**0.2447**	18.43	0.6936	**7.0277**	0.3386	16.03	0.5973

Table 2: This table compares the performance of prior models with our SH-GAN on datasets FFHQ and Places2 under resolution 512.

and the high-frequency information is encoded in the high-resolution feature maps (*e.g.* x_{64}). We then add back x_i to the corresponding $x^{[i]}$. For $i > 64$, we directly pass $x^{[i]}$ to synthesis blocks. Recall that we adopt StyleGAN2 [24] as our synthesis network. We then modulate the synthesis network using the concatenation of w and w_0, in which w is the projected vector of the latent code l using the mapping network, and connecting $x^{[i]}$ with each synthesis block via addition.

During the training time, we use a StyleGAN2 discriminator for our adversarial loss. We also use path length regularization on the generator with $w_{pl} = 2$, and R_1 regularization on the discriminator with $\gamma = 10$. Other training details can be found in Section 4.2.

3.4. Mask generation

We use the same free-form mask generation algorithm as DeepFillV2 [59] and CoModGAN [63]. These masks are drawn using multiple brush strokes and rectangles. The width of the brush stroke is sampled from $U(12, 48)$ and the number of strokes is chosen randomly from $U(0, 20)$, where $U(\cdot)$ represents the discrete uniform distribution. Meanwhile, we sample $U(0, 5)$ rectangles up to the full size of the input image, and $U(0, 10)$ rectangles up to the half-size. For more details, please see supplementary.

4. Experiments

This section goes through the details about our dataset, metrics, settings, experiments, and other studies. We provide comprehensive comparisons between our SH-GAN and other prior works through scores and images.

4.1. Datasets and Metrics

We use three datasets: FFHQ [26], Places2 [65], and DTD [11]. FFHQ contains 70,000 high-resolution well-aligned face images, in which we split out 60,000 images for training and use the remaining 10,000 images for validation. Places2 contains 8,026,628 images in its training set and 36,500 images in its validation set. The contents of Places2 are regular scenes and objects. We maintain the original train-validation split for our experiments. DTD contains 5,640 categorized texture images, among which 3,760 are from the train and validation set, and 1,880 are from the test set. We train our models using the train and validation sets and evaluate them using the test set.

We use *Fréchet Inception Distance* (FID) [19] as our primary metric. FID is a statistical score that calculates the

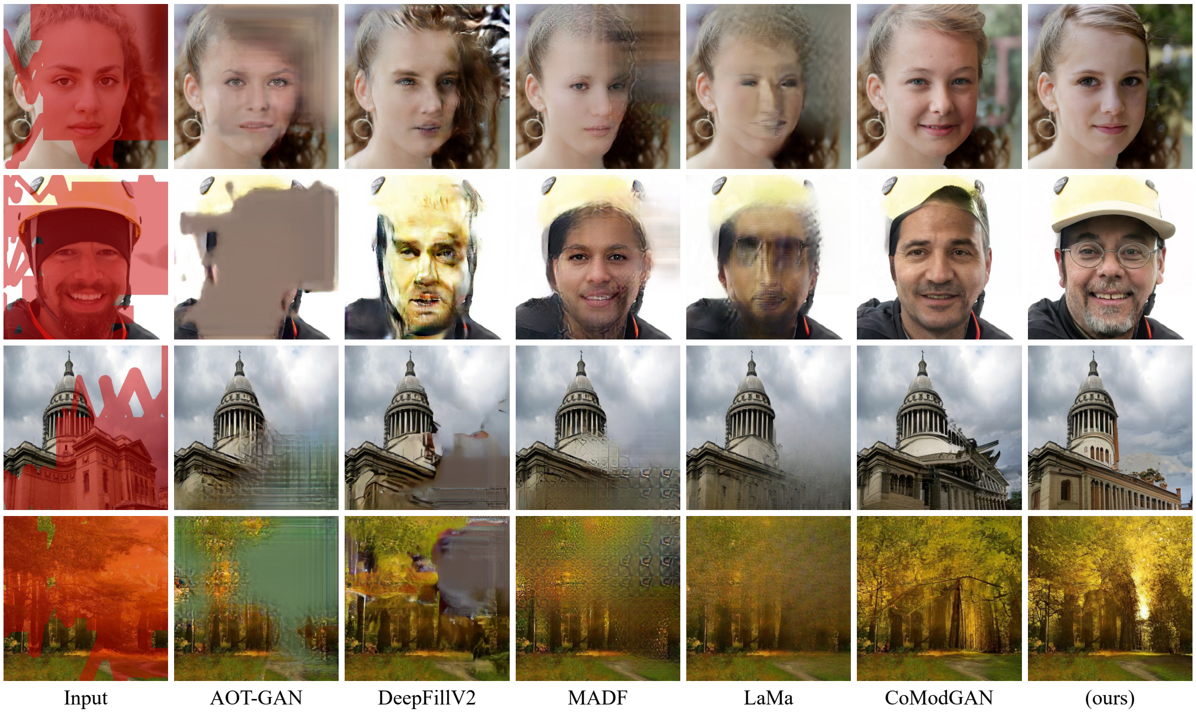


Figure 6: The qualitative comparison between prior approaches and our approach using free-form masks. For other types of masks, please see supplementary.

distance between the distributions of real and synthetic features. Besides, we also adopt *Learned Perceptual Image Patch Similarity* (LPIPS) [62], *Peak Signal-to-Noise Ratio* (PSNR), and *Structural Similarity Index* (SSIM) [52] to gauge models from different angles. We will show all the metric scores in Section 4.3.

4.2. Training Details

Many of the training settings of this work closely follow CoModGAN [63] and StyleGAN2 [24]. We use Adam [28] optimizer with $\beta = (0, 0.99)$ for our generator and discriminator. The learning rates are 0.001 for FFHQ/Places2, and 0.002 for DTD. Besides, the training lengths are 25 million images on FFHQ, 50 million on Places2, and 10 million on DTD. We apply both path length regularization and R_1 regularization during the training, in which we set $w_{pl} = 2$ and $\gamma = 10$. The batch size is 32 for all models on all datasets. Like StyleGAN2, we compute the exponential moving average of our generator with momentum 0.99993 (*i.e.* half-decay at 20,000 images with batch size 32). As shown in Tables 1 and 2, we evaluate SH-GAN with small and standard settings under resolutions 256 and 512. SH-GAN (small) is a reduced version of SH-GAN with base channel decreased from 32,768 to 16,384 (See supplementary). Such small and standard settings match the small and official versions of CoModGAN in terms of the model size. We train the small model with 4 GPUs and the standard model with 8 GPUs. Besides, we use 2080Ti for resolution 256 and A100 for resolution 512.

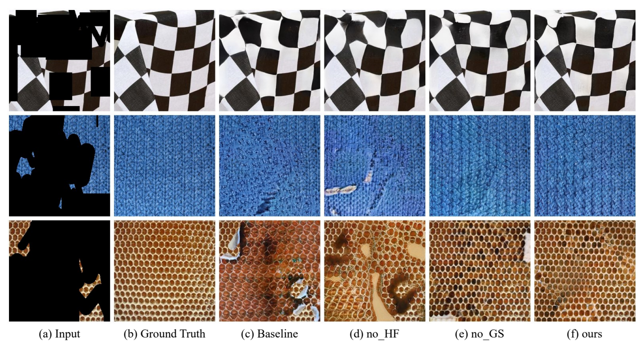


Figure 7: This figure shows the qualitative comparison of various settings in our ablation study. The performance is gradually improved with our SHU and our spectral processing strategies.

4.3. Results Comparison

Tables 1 and 2 compare the performances of SH-GAN with other prior works [48, 59, 60, 61, 63, 66] on FFHQ and Places2. As mentioned in Section 3.4, we adopt the free-from masks originated from the CoModGAN paper. Among the four metrics, FID/LPIPS gauge perception quality, and PSNR/SSIM gauge pixel accuracy. Note that these metrics reveal image quality in a different way so they may disagree in numbers. For most of the prior works, except for CoModGAN, we have downloaded the official code and models for evaluation. We re-implemented CoModGAN in

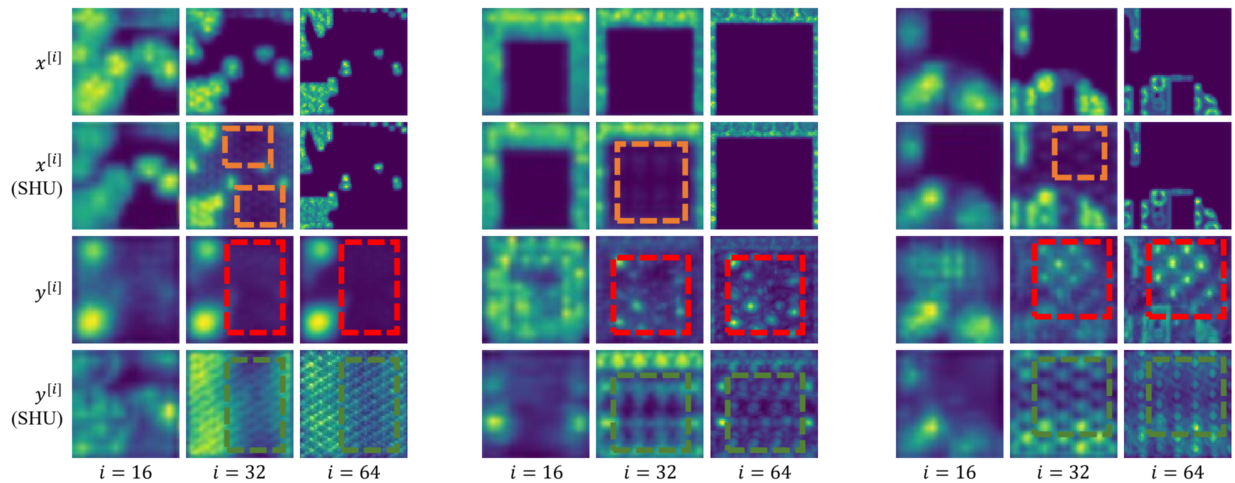


Figure 8: This figure shows three examples comparing intermediate features with SHU and without SHU. The orange dashed boxes highlight the spectral hints added by SHU. The green and red dashed boxes compare the locations where features are generated with and without fidelity due to the spectral hints.

Pytorch, and trained and tested the replicated version. The FID scores on the replicated CoModGAN match the FID scores in the original paper. As a result, SH-GAN reaches 4.3459 and 7.5036 on FFHQ and Places2 datasets respectively for the resolution 256, and 3.5713 and 7.8482 for the resolution 512. SH-GAN surpasses all other approaches in terms of FID and becomes the new state-of-the-art.

In addition to the free-form mask experiments, we also tried other types of masks such as the LaMa-style [48] narrow and wide masks. These detailed performance can be found in our supplementary material.

4.4. Extended Studies

For this section, we carry out extended experiments to justify our new designs (*i.e.* SHU, Heterogeneous Filtering, and Gaussian Split) over the prior challenges such as pattern unawareness, blurry texture, and structure distortion.

Our first experiment is an ablation study using the DTD dataset, in which we trained several models and excluded SHU, Heterogeneous Filter, or Gaussian Split, respectively. We focused on the DTD dataset because texture images are highly structured images that could make obvious cases for comparison. In Table 3, we show that the full version of our model performed FID 48.58, lower than the model without Gaussian Split (*i.e.* ours - no_GS) by 1.48, lower than the model without Heterogenerous Filtering (*i.e.* ours - no_HF) by 1.83, and lower than the baseline (*i.e.* CoModGAN [63]) by 3.35. The LPIPS, PSNR and SSIM scores agree with our FID score. Moreover, we clearly show in Figure 7 that our model generates sharp and robust patterns even when the mask coverage is large. Our HeFilter is very helpful in cases with complex structures and our Gaussian Split helps to remove the aliasing effect to make the pattern sharper.

Models	SHU	HF	GS	FID(↓)	LPIPS(↓)
baseline				51.9289	0.3628
ours - no_HF	✓		✓	50.4074	0.3614
ours - no_GS	✓	✓		50.0634	0.3573
ours	✓	✓	✓	**48.5814**	**0.3519**

Table 3: The ablation study on DTD [11] with SHU, Heterogeneous Filtering (HF), and Gaussian Split (GS) removed respectively.

In our second experiment, we extract the skip feature maps $x^{[i]}$ generated by the encoder, and the intermediate feature maps $y^{[i]}$ generated by the synthesis blocks, $i \in \{16, 32, 64\}$. We then compute the 2-norm of each feature map along the channel axis. To make clear comparisons, our SHU is only connected on resolution 32 without splitting. Figure 8 shows the impact of SHU over these features, in which readers may notice that SHU provides critical hints on patterns to its downstream synthesis blocks for texture generation.

5. Conclusions

We introduce SH-GAN, a novel image completion approach that transforms deep features with spectral hints in the modulated GAN framework. Throughout this paper, we reveal the details of our newly designed module: SHU, and introduce our new spectral transform strategies: Heterogeneous Filtering and Gaussian Split. With inclusive experiments, we show that all our designs are very useful in solving challenging inpainting cases with large-scale free-form missing regions. We believe that our SHU and spectral transform strategies are worth exploring further in other computer vision tasks.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 2

[2] Nasir Ahmed, T₋ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 2

[3] G. Anderson and T. Huang. Piecewise fourier transformation for picture bandwidth compression. *IEEE Transactions on Communication Technology*, 19(2):133–140, 1971. 1, 2

[4] S Arivazhagan, S Deivalakshmi, K Kannan, BN Gajbhiye, C Muralidhar, Sijo N Lukose, and MP Subramanian. Performance analysis of image denoising system for different levels of wavelet decomposition. *International Journal of Imaging Science and Engineering*, 1(3):104–107, 2007. 1, 2

[5] 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. 3

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

[7] Reinhard Beer. *Remote sensing by Fourier transform spectrometry*, volume 170. John Wiley & Sons, 1992. 1

[8] E Oran Brigham. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988. 1

[9] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. Ssd-gan: Measuring the realness in the spatial and spectral domains. *arXiv preprint arXiv:2012.05535*, 2020. 1, 2

[10] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 1, 2, 3, 4

[11] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6, 8

[12] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. 3

[13] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001. 3

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

[15] Burhan Ergen. *Signal and image denoising using wavelet transform*. InTech London, UK, 2012. 1, 2

[16] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2

[17] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 3

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[24] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 2, 3, 5, 6, 7

[25] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 5

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3, 5, 6

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3, 5

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[29] Rolf Köhler, Christian Schuler, Bernhard Schölkopf, and Stefan Harmeling. Mask-specific inpainting with deep neural networks. In *German conference on pattern recognition*, pages 523–534. Springer, 2014. 3

[30] Avisek Lahiri, Arnav Jain, Prabir Kumar Biswas, and Pabitra Mitra. Improving consistency and correctness of sequence inpainting using semantically guided generative adversarial network. *arXiv preprint arXiv:1711.06106*, 2017. 3

[31] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 3

[32] Dimitris G Manolakis and Vinay K Ingle. *Applied digital signal processing: theory and practice*. Cambridge university press, 2011. 1

[33] Josiah Manson and Scott Schaefer. Wavelet rasterization. In *Computer Graphics Forum*, volume 30, pages 395–404. Wiley Online Library, 2011. 1

[34] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 5

[35] Michael C Martin, Charlotte Dabat-Blondeau, Miriam Unger, Julia Sedlmair, Dilworth Y Parkinson, Hans A Bechtel, Barbara Illman, Jonathan M Castro, Marco Keiluweit, David Buschke, et al. 3d spectral imaging with synchrotron fourier transform infrared spectro-microtomography. *nature methods*, 10(9):861–864, 2013. 1

[36] Shant Navasardyan and Marianna Ohanyan. Image inpainting with onion convolutions. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 3

[37] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[38] Chun-Su Park. 2d discrete fourier transform on sliding windows. *IEEE Transactions on Image Processing*, 24(3):901–907, 2015. 1

[39] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[40] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 1, 2

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[42] Jimmy S Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. *Advances in neural information processing systems*, 28, 2015. 3

[43] Oren Rippel, Jasper Snoek, and Ryan P Adams. Spectral representations for convolutional neural networks. *Advances in neural information processing systems*, 28, 2015. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[45] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[47] Jean-Luc Starck, Emmanuel J Candès, and David L Donoho. The curvelet transform for image denoising. *IEEE Transactions on image processing*, 11(6):670–684, 2002. 1, 2

[48] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 1, 2, 7, 8

[49] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[50] R. Tsai and T. Huang. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 1, 2

[51] Jin Wang, Yanwen Guo, Yiting Ying, Yanli Liu, and Qunsheng Peng. Fast non-local algorithm for image denoising. In *2006 International Conference on Image Processing*, pages 1429–1432. IEEE, 2006. 2

[52] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 7

[53] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25, 2012. 3

[54] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021. 1

[55] Li Yi, Hao Su, Xingwen Guo, and Leonidas J Guibas. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2282–2290, 2017. 1, 2

[56] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 2, 3

[57] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 1

[58] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 5505–5514, 2018. 2, 3

[59] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2, 3, 6, 7

[60] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *arXiv preprint arXiv:2104.01431*, 2021. 3, 7

[61] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14164–14173, 2021. 2, 3, 7

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[63] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2, 3, 5, 6, 7, 8

[64] Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. *Advances in neural information processing systems*, 31, 2018. 1, 2

[65] 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*, 2017. 2, 6

[66] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021. 3, 7