DHG-GAN: Diverse Image Outpainting via Decoupled High Frequency Semantics

Yiwen Xu[1[0000−0002−7914−3594]], Maurice Pagnucco[1[0000−0001−7712−6646]], and Yang Song*[1[0000−0003−1283−1672]]

School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
yiwen.xu1@student.unsw.edu.au, {morri,yang.song1}@unsw.edu.au

Abstract. Diverse image outpainting aims to restore large missing regions surrounding a known region while generating multiple plausible results. Although existing outpainting methods have demonstrated promising quality of image reconstruction, they are ineffective for providing both diverse and realistic content. This paper proposes a Decoupled High-frequency semantic Guidance-based GAN (DHG-GAN) for diverse image outpainting with the following contributions. 1) We propose a two-stage method, in which the first stage generates high-frequency semantic images for guidance of structural and textural information in the outpainting region and the second stage is a semantic completion network for completing the image outpainting based on this semantic guidance. 2) We design spatially varying stylemaps to enable targeted editing of high-frequency semantics in the outpainting region to generate diverse and realistic results. We evaluate the photorealism and quality of the diverse results generated by our model on CelebA-HQ, Place2 and Oxford Flower102 datasets. The experimental results demonstrate large improvement over state-of-the-art approaches.

Keywords: Diverse image outpainting · GAN · Image reconstruction.

1 Introduction

Image outpainting (as shown in Fig. 1) aims to reconstruct large missing regions and synthesise visually realistic and semantically convincing content from a limited input content [37, 22, 8, 33]. This is a challenging task because it utilises less neighbouring reference information to extrapolate unseen areas and the regions that are outpainted should look aesthetically genuine to the human eye. Image outpainting has gained considerable interest in recent years and has broadly novel applications, including image and video-based rendering [21], image reconstruction [37] and image modification [23].

Early outpainting methods [30, 35, 2, 46] are usually patch-based, matching and stitching known pixel blocks or semantic vectors in the input to outpaint images. Afterwards, using the image reconstruction methods to infer semantics is

* Corresponding authors.

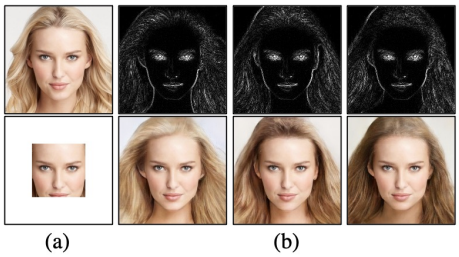

Fig. 1. Diverse image outpainting results with cropped images as inputs. (a) Original image and input image with an outpainting mask. (b) Generated high-frequency semantic image (top) and final outpainting result (down).

proven effective for image outpainting. Wang *et al.* [37] proposed a progressive semantic inference method to expand complex structures. Lin *et al.* [22] presented a coarse-to-fine network that utilises edge maps to guide the generation of structures and textures in the outpainting region. [42] and [36] achieved a large-scale expansion of landscape images and ensured semantic coherence through a long short-term memory (LSTM) [10] encoder. However, current outpainting methods are only focused on enhancing the reconstruction quality of outpainting regions. Methods for *diverse outpainting* need to be able to infer the missing area from a small area of known pixels as well as provide diverse outputs with plausible visual and semantic information.

Currently, StyleGAN [15] has made significant progress in generating diverse images. The inverse mapping method also enables StyleGAN-based models to modify the semantics of real images. [16] and [1] demonstrate that StyleGAN can modify local semantic information based on spatially varying stylemaps. In addition, a series of variational auto-encoder (VAE) [19] based methods have been proposed to generate diverse reconstruction results. Zheng *et al.* [48] combines two VAE pipelines to trade-off output diversity and reconstruction quality based on probabilistic principles. However, this also leads to a gradual deterioration in the reconstruction quality when the similarity between the reconstruction results and the ground truth decreases. Peng *et al.* [27] proposed a model based on VQ-VAE [28], which generates a variety of structural guidance via a conditional autoregressive network PixelCNN [34]. Nevertheless, due to the randomness of the generated results of PixelCNN, it is difficult to control the structural information in the output. To introduce explicit control in the reconstruction region, an intuitive approach is to utilise artificial sketches to modify the texture and structure details [36, 23].

In this paper, we focus on using *decoupled high-frequency information* to guide the diverse generation of outpainted images. There are two main challenges in this task. First, previous studies have utilised sketches as guidance for generating diverse structures, however, providing modification guidance for complex structural and texture information becomes challenging. Furthermore, it is difficult to ensure the quality, diversity and controllability of the results for outpainting.

To solve these issues, we propose Decoupled High-frequency semantic Guidance-based GAN (DHG-GAN), a diverse image outpainting network built upon a spatially varying stylemap. Our DHG-GAN is a two-stage model that starts from generating a high-frequency semantic image for the outpainting region and then completes the image reconstruction based on the guidance provided by the high-frequency semantic image. Specifically, 1) we design a StyleGAN-based model to generate for the entire image a high-frequency semantic image from a spatially varying stylemap; and, 2) the second stage utilises an encoder-decoder structure network to complete the high-frequency semantic image with low-frequency information to generate realistic outputs.

Previous research shows that high-frequency information can improve image reconstruction performance [40, 25, 22]. In our method, we decouple the high-frequency semantic information through Fourier high-pass filtering, which becomes the ground truth for the first stage. This decoupled high-frequency semantics can provide rich texture details and structural information for the outpainting region. By interpolating the spatially varying stylemaps, it is feasible to generate a variety of high-frequency semantic images for the outpainting region that allow the semantic completion network to synthesise diverse results (as shown in Fig. 1). We compare with Canny, Sobel, Prewitt and Laplacian edge maps and determine that our decoupled high-frequency semantic image provides the best performance in terms of quality and diversity of image outpainting. There are three main contributions in this paper:

- We present the first diverse image outpainting model utilising decoupled high-frequency semantic information, which demonstrates state-of-the-art performance on CelebA-HQ [24], Places2 [49] and Oxford Flower102 [26] datasets which are commonly used in outpainting studies.
- We propose a two-stage diverse outpainting model DHG-GAN that consists of a high-frequency semantic generator and semantic completion network. The first stage generates images to provide guidance of high-frequency semantics for the outpainting region. The second stage performs semantic completion in the outpainting region to generate realistic images.
- We design a StyleGAN-based high-frequency semantic image generator for modifying the structure and texture information in the outpainting region via a spatially varying stylemap. Ablation experiments show that our method can achieve editing of complicated structures and textures.

2 Related Work

2.1 Image Outpainting

Early outpainting models expand input images by retrieving appropriate patches from a predefined pool of patch candidates [20, 46, 35, 32]. The performance of such methods depends on the retrieval ability and the quality and quantity of candidate images. Later inspired by generative adversarial networks (GANs) [6], semantic regeneration network (SRN) [37] incorporates a relative spatial variant

loss to gradually infer semantics, which improves the performance for repairing contours and textures. In addition, Yang *et al.* [41] proposed an outpainting method that can synthesis association relationships for scene datasets. Besides, [22] and [17] utilise an edge map generation module to provide richer textural and structural guidance, thereby improving the outpainting performance. However, these GAN-based outpainting methods rely on pixel-wise reconstruction loss and do not employ random variables or control information, and hence the outpainted images are limited in diversity. Wang *et al.* [36] developed a method to generate controllable semantics in the outpainted images based on artificial sketches. However, it is difficult to use sketches to provide complex textural and structural information such as hair and petal texture. Our method is based on a high-frequency semantic image that can present more detailed guidance information for outpainting regions.

2.2 Diverse Image Reconstruction

To generate diverse reconstruction results on cropped images, some methods train VAE-based encoder-decoder networks to condition a Gaussian distribution and sample the diversity result at test time [47, 48]. Zheng *et al.* [48] proposed a framework with two parallel paths; one reconstructs the original image to provide the prior distribution for the cropped regions and coupling with the conditional distribution in the other path. Peng *et al.* [27] proposed to utilise the hierarchical architecture of VQ-VAE [28] to disentangle the structure and texture information and use vector quantisation to model the discrete distribution of the structural features through auto-regression to generate a variety of structures. However, the sampled distribution constrains the diversity of outputs generated by these methods. In contrast to these methods, our model trains high and low frequency features independently in two stages and modifies the structure and texture details using an encoded stylemap to produce diverse and realistic outputs.

2.3 GAN-based Image Editing

In order to make the reconstruction results controllable, various methods inject sketches as guidance to edit the content of the image. DeFLOCNet [23] and SC-FEGAN [13] perform sketch line refinement and colour propagation in the convolutional neural network (CNN) feature space for injected sketches. In addition, recent studies have demonstrated that GANs are capable of learning rich semantics from latent space and manipulating the corresponding features of the output images by modifying the latent code. For instance, BicycleGAN [50] constructs an invertible relationship between the latent space and the generated image, which assists in decoupling the semantic information contained in the latent code in order to achieve semantic editing. Kim *et al.* [16] and Alharbi *et al.* [1] proposed spatially varying stylemaps that enable semantic modification of generated images locally. Nevertheless, because these approaches are not designed for image reconstruction, the output images usually contain artifacts. In

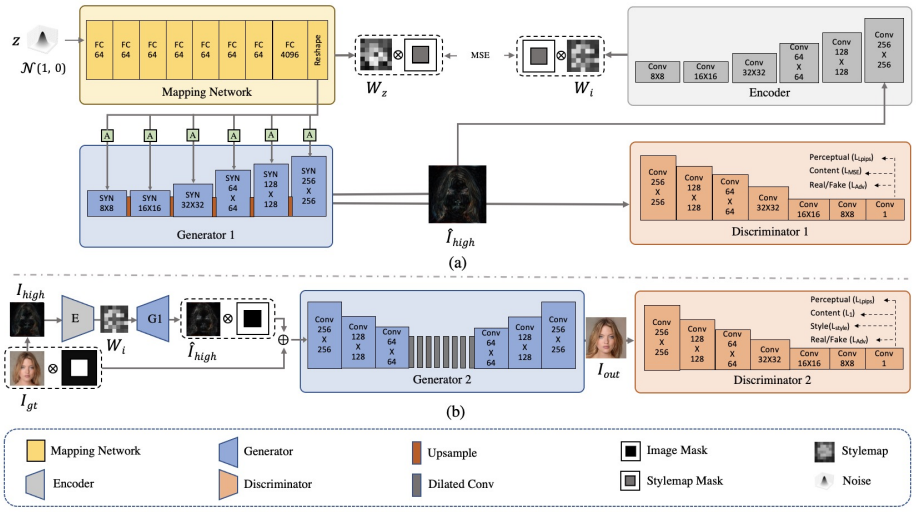

Fig. 2. Our framework illustration. (a) The high-frequency semantic image generator (HFG) consists of a mapping network, encoder, generator, and discriminator. The generator synthesises high-frequency semantic images (HSIs) based on the stylemap of generated by the mapping network. The encoder constructs an accurate inverse mapping of the synthesised HSIs through MSE supervision between the encoder and the mapping network. (b) The semantic completion network (SCN) consists of an encoder-decoder structure generator and an SN-PatchGAN discriminator. The input of the generator is a combination of the input image and a generated HSI. The semantic competition network outputs a realistic image after completing semantics.

addition, Cheng *et al.* [4] proposed a method that searches multiple suitable latent codes through the inversion process of GAN to generate diverse outpainting results. However, this method is only designed for landscape images.

3 Methodology

In this paper, we propose the DHG-GAN method for reconstructing images with diverse structures and textures in the outpainting region. Our proposed model consists of two stages: a high frequency semantic generator (HFG) and a semantic completion network (SCN). Given an original (ground truth) image I_{gt} of size 256×256 pixels, we first obtain a cropped image $\bar{I}$ of size 128×128 pixels by $\bar{I} = I_{gt} \times M_i$, where M_i is a binary mask indicating the known region to be kept in I_{gt}. The outpainting process is to reconstruct I_{gt} from $\bar{I}$ and diversity is introduced via the HFG module. The overall framework is shown in Fig. 2.

The first stage is inspired by StyleGAN [15] and the goal of this stage is to generate a high-frequency semantic image (HSI) to provide guidance for the second-stage semantic completion. Our generated HSI contains rich textural and structural information, which is sufficiently similar to the decoupled high-

frequency information from the ground truth. Accordingly, this enables the second stage to establish a strong correlation between the completed low-frequency semantic components and the associated HSI when reconstructing the image. Moreover, in the first stage, we construct an accurate mapping between the spatially varying stylemap and HSI, and then modifying the stylemap for the outpainting region enables the generator to synthesise diverse HSIs.

Existing research has shown that high-frequency information such as edge maps can be used to improve the quality of reconstructed images, particularly to sharpen contours and textures [40, 25, 22]. However, common edge maps, such as Canny, Sobel and Prewitt, lose a lot of texture information, whereas Laplacian edge maps contain grain noise. Therefore, we choose to decouple the high-frequency information of the RGB the channels of the ground truth image through a Fourier high-pass filter bank, which is used as the ground truth for learning the HSI. Such decoupled high-frequency information contains more comprehensive structural and textural information and less noise.

3.1 Revisting StyleGAN

We first briefly introduce StyleGAN, on which our proposed DHG-GAN is based. In the first stage, the aim of using a StyleGAN-based network is to encode samples to obtain a mapping of stylemap-to-image. Then, we can change the style code in the stylemap to get diverse images. StyleGAN proposes a style-based generator, which consists of a mapping network and synthesis network. To generate images using the generator, StyleGAN first randomly samples a latent vector Z with a Gaussian distribution $\mathcal{N}(1, 0)$, then projects it to a stylemap W by the mapping network. The stylemap W is then fed into each convolutional layer in the synthesis network through an affine transformation and Adaptive Instance Normalization (AdaIN) [11]. The discriminator then distinguishes the authenticity of the images.

Unlike StyleGAN, we provide an encoder to create the inverse mapping of HSIs to stylemaps. The mean squared error (MSE) is used to minimize the difference between HSIs and stylemaps in order to construct a more precise inverse mapping. Our encoder and mapping network outputs are 3D stylemaps, which allow direct editing of structural information and textural detials for the outpainting regions. The encoder network structure is similar to the discriminator in StyleGAN.

3.2 DHG-GAN

High Frequency Semantic Generator (HFG) To generate diverse structures and textures for the outpainting region, we design an HFG module. HFG generates a high-frequency semantic image (HSI) that provides texture and structure information to guide the second stage to complete and improve the quality of image outpainting. In addition, modifying the style code in the stylemap for the outpainting region can enable HFG to synthesise various HSIs, thus providing diverse guidance for the second stage to generate diverse outputs.

Our HFG utilises a similar network structure to StyleGAN. As shown in Fig. 2, HFG consists of a mapping network F, encoder E, generator G_1 and discriminator D_1. We map the input latent code Z onto a spatially varying stylemap W_z via the mapping network. Such additional mapping has been demonstrated to learn more disentangled features [31, 15]. The high-frequency information I_{high} is decoupled from the orginal image I_{gt} using Fourier high-pass filters, which is used as the ground truth of HFG. The generator learns to synthesise I_{high} from W_z and output the HSI $\hat{I}_{high}$. Besides, we use the encoder E to build a reverse mapping from HSI to spatially varying stylemaps W_z. This inverse mapping is to enable semantic modification for outpainting regions.

Specifically, given a latent code Z with Gaussian distribution, the mapping network $F : Z \rightarrow W_z$ produces a 3D stylemap W_z. Then, we use AdaIN operations to transfer the encoded stylemap W_z to each convolution layer in the generator. Here, AdaIN is defined as:

$$\mathbf{x}_{i+1} = a_i \frac{\mathbf{x}_i - \mu\left(\mathbf{x}_i\right)}{\sigma\left(\mathbf{x}_i\right)} + b_i \tag{1}$$

where μ and σ compute the mean and standard deviation of each feature map $\mathbf{x}_i$. a_i and b_i are the style code computed from W_z. This enables stylemaps to be added to every synthesis module in the generator. As shown in Fig. 3, the generator contains synthesis modules of different resolution scales, and the last image-scale synthesis module generates the HSI $\hat{I}_{high}$. Then, the discriminator distinguishes I_{high} and $\hat{I}_{high}$.

The encoder E is used to establish an inverse mapping from HSI $\hat{I}_{high}$ to stylemap W_i. Then we crop W_i and W_z to keep the outpainted regions $\bar{W}_i = W_i \times (1 - M_s)$ and $\bar{W}_z = W_z \times (1 - M_s)$, where M_s is a binary mask corresponding to the outpainting region. We minimise the difference between $\bar{W}_i$ and $\bar{W}_z$ by MSE loss to train the encoder. This supervision aims to make the stylemaps $\bar{W}_i$ and $\bar{W}_z$ close in the latent space, so that E can generate stylemaps that are more suitable for semantic modification of the outpainting region. This MSE loss at the stylemap level is formulated as:

$$L_{mse_s} = \|\bar{W}_i - \bar{W}_s\|_2^2 \tag{2}$$

We also use a combination of MSE, learned perceptual image patch similarity (LPIPS) and hinge adversarial losses to train the generator and discriminator. MSE loss measures the pixel-level similarity which can be formulated as:

$$L_{mse_i} = \|(I_{high} - \hat{I}_{high}\|_2^2 \tag{3}$$

LPIPS loss [45] is used to measure perceptual differences and improve the perceptual quality of the HSI. Inspired by [12, 7], we observe that images generated with LPIPS loss have less noise and richer textures than using perceptual loss [14]. This is due to the VGG-based perceptual loss being trained for image classification, whereas LPIPS is trained to score image patches based on human

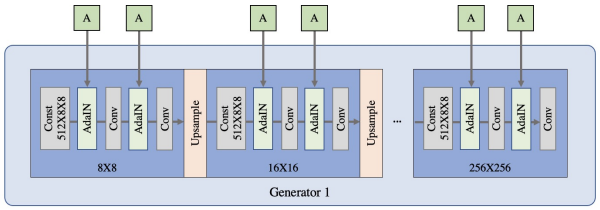

Fig. 3. Generator G_1 in HFG. The generator consists of affine transformation modules, adaptive instance normalization (AdaIN) operation and multiple scale convolution layers. After obtaining the spatially varying stylemaps, we add it to the generator using affine transformations and AdaIN.

perceptual similarity judgments. LPIPS loss is defined as:

$$L_{lpips1} = \sum_{l} \tau^{l} \left(\phi^{l} \left(\hat{I}_{high} \right) - \phi^{l} \left(I_{high} \right) \right) \tag{4}$$

where ϕ is a feature extractor based on a pre-trained AlexNet. τ is a transformation from embeddings to a scalar LPIPS score, which is computed from l layers and averaged.

In addition, the hinge adversarial loss for generator G_1 and discriminator D_1 in this stage is defined as:

$$L_{G_1} = -\mathbb{E}_{I_{high}} \left[D_1 \left(G_1 \left(A(W_z) \right), I_{high} \right) \right] \tag{5}$$

$$\begin{aligned} L_{D_1} =& \mathbb{E}_{(I_{gt}, I_{high}))} \left[\max \left(0, 1 - D_1 \left(I_{gt}, I_{high} \right) \right) \right] \\ &+ \mathbb{E}_{I_{high}} \left[\max \left(0, 1 + D_1 \left(G(A(W_z), I_{high} \right) \right) \right] \end{aligned} \tag{6}$$

where A is the affine transformation function. The overall objective function of HFG can be written as:

$$L_{total_1} = L_{lpips1} + L_{G_1} + L_{mse_s} + 0.1 \cdot L_{mse_i} \tag{7}$$

The pixel-level MSE loss improves the outpainting quality but also affects the diversity of HSI. Here we assign a smaller weight to L_{mse_i} based on our empirical evaluations.

Semantic Completion Network (SCN) In the second stage, we design a SCN to utilise the HSI generated in the first stage to complete the outpainting region semantically and create a realistic image. SCN has an encoder-decoder structure with a generator G_2 and discriminator D_2. To train SCN, first, based on the ground truth image I_{gt}, we obtain the high-frequency information I_{high} and generate the stylemap W_i by inverse mapping I_{high} through the first stage encoder E. The generator G_1 then uses W_i to synthesise a HSI $\hat{I}_{high}$. We then combine the cropped input image $\hat{I}$ (to be outpainted) and $\hat{I}_{high}$ into $\bar{I} = \hat{I} +$

$\hat{I}_{high} \times (1 - M_i)$ as the input to the generator G_2. The generator G_2 then performs semantic completion on this input to generate realistic results I_{out}.

As shown in Fig. 2, G_2 contains 3 encoder and decoder layers, and there are 8 dilation layers and residual blocks in the intermediate layers. This generator structure is inspired by [25], and we using dilation blocks in the middle in order to promote a wider receptive field at the output neuron. The discriminator follows the SN-PatchGAN structure [44]. In order for SCN to generate realistic results, we utilise several loss functions, including style loss [29], $L1$ loss, LPIPS loss and adversarial loss. We measure pixel-wise differences between the output of G_2 and I_{gt} by $L1$ loss and take into account high-level feature representation and human perception by LPIPS loss. Besides, style loss compares the difference between the deep feature maps of I_{out} and the I_{gt} from pre-trained and fixed VGG-19, and has shown effectiveness in counteracting "checkerboard" artifacts produced by the transpose convolution layers. Style loss is formulated as follows:

$$L_{\text{style}} = \mathbb{E}\left[\sum_j \left\| Gram_j^{\phi}\left(I_{gt}\right) - Gram_j^{\phi}\left(I_{out}\right)\right\|_1\right] \tag{8}$$

where the $Gram_j^{\phi}$ is the Gram matrix of of the $j\text{-}th$ feature layer. The adversarial loss over the generator G_2 and discriminator D_2 are defined as:

$$L_{G_2} = -\mathbb{E}_{I_{\text{gt}}}\left[D_2\left(I_{out}, I_{\text{gt}}\right)\right] \tag{9}$$

$$L_{D_2} = \mathbb{E}_{I_{gt}}\left[\text{ReLU}\left(1 - D_2\left(I_{gt}\right)\right)\right] + \mathbb{E}_{I_{out}}\left[\text{ReLU}\left(1 + D_2\left(I_{out}\right)\right)\right] \tag{10}$$

Finally, the total function is defined as a weighted sum of different losses with the following coefficients:

$$L_{total_2} = L_1 + 250 \cdot L_{style} + 0.1 \cdot L_{lpips2} + 0.1 \cdot L_{G_2} \tag{11}$$

4 Experiments

4.1 Datasets and Implementation Details

We evaluate our model on three datasets, including CelebA-HQ [24], Places2 [49] and Oxford Flower102 [26]. The CelebA-HQ dataset includes 30000 celebrity face images at 1024 × 1024 pixels. Places2 contains 1,803,460 images with 434 different scene categories. Oxford Flower102 comprises 102 flower categories and a total of 8189 images. The official training, testing and validation splits are used for these three datasets. We resize the images to 256×256 pixels with data augmentation and use the center outpainting mask to reserve the middle region of images for testing.

Our model is implemented in Pytorch v1.4 and trained on an NVIDIA 2080 Ti GPU. The model is optimized using the Adam optimizer [18] with $\beta_1 = 0$ and

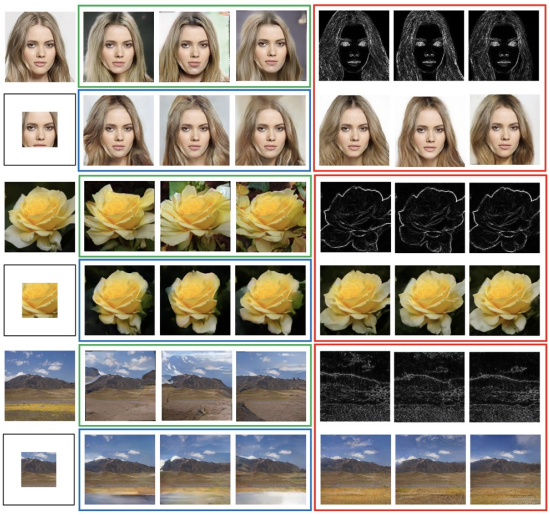

Fig. 4. Qualitative comparison results with diverse-solution methods on CelebA-HQ [24], Places2 [49] and Oxford Flower102 [26] datasets. For each dataset, from top to bottom, from left to right, the images are: original image, input image, results of PIC [48] (with green box), results of DSI [27] (with blue box), results of our method showing both HSIs and the generated images (with red box).

$\beta_2 = 0.9$. The two stages are trained with learning rates of 0.001 and 0.0001, respectively, until the loss plateaus, with a batch size of 4. In addition, in order to train the HFG, we adjusted the size scale of the stylemap. Through our experiments, we found that stylemaps of 64×8×8 can produce satisfactory HSIs.

During the testing phase, we traverse the images $\{I_{gt}\}$ in the validation set to obtain a set of HSIs $\{I_{high}\}$ using Fourier high-pass filtering and utilise the encoder E to map these HSIs to obtain a variety of stylemaps. We use W_i to represent the stylemap of the current test image and $\{W_{ref}\}$ to represent the stylemaps of other images in the validation set. In addition, W_z is generated by the mapping network F through random vectors Z. We crop $\{W_{ref}\}$ and W_z through the outpainting mask, retaining only the style code for the outpainting region. For the test image I, we crop its stylemap W_i to retain the style code only for the centre kept region, corresponding to the input image that will be outpainted, so that ground truth information is not used during outpainting. Then, we can get various complete stylemap $\hat{W}_i$ by combine W_i, W_{ref} and W_z with the following operations:

$$\hat{W}_i = (W_i \times M_s) + (W_{ref} \times (1 - M_s)) \cdot 0.2 + (W_z \times (1 - M_s)) \cdot 0.8 \qquad (12)$$

where $W_i \times M_s$ denotes the cropped stylemaps of the test image, $W_{ref} \times (1 - M_s)$ and $W_z \times (1 - M_s)$ denote the cropped stylemap of a sampled reference image in the validation set and cropped stylemap generated from the random vector,

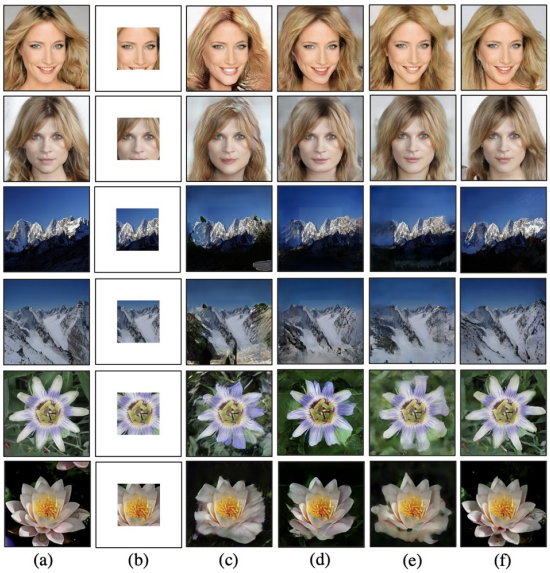

Fig. 5. Qualitative comparison results with single-solution methods on CelebA-HQ (top), Place2 (middle) and Oxford Flower102 (down) datasets. For each dataset, from left to right, the image are: (a) Original image. (b) Input images. (c) Results of CA [43]. (d) Results of EdgeConnect [25]. (e) Results of SRN [37]. (f) Results of our method.

respectively. Then, by passing the various $\hat{W}_i$ to the generator G_1 in HFG, we can obtain diverse $\hat{I}_{high}$, so that SCN can produce various outpainting results I_{out}. The coefficients assigned to the stylemaps force HSIs to trade-off between being similar to the original image or the other reference images. Following previous diverse-solution methods [48, 47], we also select the top samples closest to the original image through the discriminator D_2 for qualitative and quantitative evaluation. For compared approaches, we use their official implementations and train on the same dataset with identical configurations to ensure a fair comparison. We use the centre mask to compare the results, which has the advantage of showing the generation effect of different regions in a balanced manner.

4.2 Qualitative Results

Fig. 4 shows the qualitative comparison results of our model and the state-of-the-art diverse solution reconstruction models: pluralistic image completion (PIC) [48] and diverse structure inpainting (DSI) [27], on CelebA-HQ, Places2 and Oxford Flower102 datasets for the outpainting task. We did not choose models based on artifact-injected edge maps for comparison, as these models are strongly influenced by artifact input [23, 44, 13]. PIC and DSI are able to generate diverse results without human intervention, which are thus more directly comparable with our method. PIC is based on a probabilistic principle with

Table 1. Comparison of different methods on CelebA-HQ, Oxford Flower102 and Place2. S denotes the single-solution methods, D denotes the diverse-solution methods.

		CelebA-HQ				Place2				Flower102			
		PSNR	SSIM	MIS	FID	PSNR	SSIM	MIS	FID	PSNR	SSIM	MIS	FID
S	CA	15.22	0.52	0.017	31.62	14.29	0.51	0.018	29.09	14.16	0.49	0.018	33.46
	EC	15.26	0.55	**0.024**	27.31	16.84	0.57	0.019	29.17	16.49	0.57	0.021	29.43
	SRN	16.65	0.56	0.023	28.29	16.94	0.60	**0.023**	**27.12**	15.97	0.50	0.021	29.73
	Ours w/o HSI	15.24	0.51	0.019	30.18	16.13	0.53	0.018	29.19	15.37	0.51	0.020	32.17
D	PIC	14.77	0.45	0.019	33.12	15.10	0.51	0.020	29.19	13.68	0.48	0.017	39.11
	DSI	15.00	0.51	0.021	31.45	15.33	0.53	0.021	29.71	16.68	0.56	0.020	29.38
	Ours w/ HSI	**16.71**	**0.59**	0.023	**27.14**	**16.97**	**0.61**	**0.023**	28.59	**17.05**	**0.62**	**0.024**	**28.76**

two parallel reconstruction paths, while DSI uses an auto-regressive network to model a conditional distribution of discrete structure and texture features and samples from this distribution to generate diverse results. It can be seen that PIC generates reasonable results on CelebA-HQ but fails to achieve inner-class diversity in multi-class datasets (Oxford Flower102). DSI performs better on inner-class diversity due to the distribution obtained through an auto-regressive model. The results also show that our method can generate more realistic outputs, e.g., the generated shapes of flowers and mountains are more plausible. In addition, since our results are generated with guidance of a high-frequency semantics image, they show finer texture and structural details.

Table 2. Classification result on the Oxford Flower102 dataset.

Method	Original	CA	EdgeConnect	SRN	PIC	DSI	Ours
VGG-S	**0.9213**	0.7945	0.8285	0.8317	0.7807	0.8310	**0.8458**
Inception-v3	**0.9437**	0.7419	0.8230	0.8593	0.7719	0.8683	**0.8731**
EffNet-L2	**0.9814**	0.7385	0.8194	0.8494	0.7355	0.8511	**0.8770**

For comparison with single-solution reconstruction methods, we select the top-1 results by discriminator D_2. As shown in Fig. 5, due to the lack of prior structural information, CA [43] has difficulties generating reasonable structures, whereas EdgeConnect [25] generates better textures and structures due to the use of edge maps as semantic guidance. SRN [37] infers semantics gradually via the relative spatial variant loss [37], enabling it to generate adequate structural and textural information. Since our results are guided by richer structural and semantic information in the outpainted regions, the semantic completion network is able to better infer and complete the images.

4.3 Quantitative Results

Following the previous image reconstruction methods, we utilise the common evaluation metrics including Peak Signal-to-Noise Ratio (PSNR) and Structural

Similarity (SSIM) [38] to determine the similarity between the outpainting results and ground truth. Additionally, we use Modified Inception Score (MIS) [47] and Fréchet Inception Distance (FID) [9] as perceptual quality indicators. Furthermore, FID is capable of detecting the GAN model's mode collapse and mode dropping [24]. For multiple-solution methods, we sample 100 output images for each input image and report the average result of top 10 samples. Table 1 shows that, although our method is lower than SRN and EdgeConnect in some metrics, it still outperforms other single and diverse solution reconstruction methods. Moreover, the MIS and FID measurements demonstrate that our method is more effective in generating visually coherent content.

Furthermore, we analyse the quality of outpainted images on the Oxford Flower102 dataset by classifying the generated images for the official validation set into 102 categories as labelled in the dataset. We use pre-trained VGG-S [3], Inception-v3 [39], EffNet-L2 [5] to evaluate our generated results, which are state-of-the-art classification methods on the Oxford Flower102 dataset. As shown in Table 2, our method has a higher classification accuracy than other methods and is closer to the classification result on original images. This can be attributed to the fact that our method can generate more reasonable and realistic semantics.

Table 3. Comparison of diversity scores on the CelebA-HQ dataset.

	LPIPS - I_{out}	LPIPS - $I_{out(m)}$
PIC [48]	0.032	0.085
DSI [27]	0.028	0.073
Ours	**0.035**	**0.089**

Also, we calculate the diversity score for comparisons with diverse-solution methods using the LPIPS [51] metric on the CelebA-HQ dataset. The average score is calculated based on 50K output images generated from 1K input images. We compute the LPIPS scores for the complete images I_{out} and then only the outpainting regions $I_{out(m)}$. As shown in Table 3, our method achieves relatively higher diversity scores than other existing methods.

4.4 Ablation Study

Alternative high-frequency semantic feature. We conduct an ablation study on CelebA-HQ to show the impact of using various types of high-frequency semantic features as a guide for outpainting. As shown in Fig. 6, the facial contours generated without using the high-frequency semantic map as a guide show some shrinkage and the texture is not clear enough. The high-frequency semantic information provided by utilizing Canny, Prewitt and Sobel edge detection operators can solve the contour shrinkage to an extent. However, we observed that such high-frequency semantic features cannot provide specific texture information. Although the Laplacian edge map can provide more texture details, it

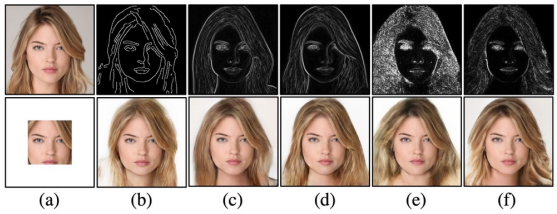

Fig. 6. Results of image outpainting using different high-frequency information. (a) Original image and input image. (b) Results of Canny. (c) Results of Sobel. (d) Results of Prewitt. (e) Results of Laplacian. (f) Results of our method.

contains a lot of grain noise, which adversely affects the quality of image restoration and texture modification. Compared to these techniques, our method can generate overall more detailed structural and texture information.

Quality of high-frequency semantic image for diverse outpainting quality. We notice that the dimension of stylemaps affects the capacity of decoupling latent semantics in the stylemap and thus the fineness of diverse outpainting. To choose a suitable size for stylemaps, we consider that for 256×256 sized images, we need at least a size of 4×4×1 for stylemaps to identify the outpainting region, thus performing semantic editing for the outpainting area. Furthermore, as illustrated in the supplementary material (Figure 1), we determined that, if the latent vector dimension in the stylemap is small, the encoder will fail to establish a high quality mapping of HSI. We found that using an 8×8×64 stylemap provided a sufficiently accurate mapping to avoid compromising the quality of the images generated.

5 Conclusion and Future Work

In this paper, we propose a diverse-solution outpainting method, DHG-GAN, for generating diverse and high-quality images guided by decoupled high-frequency semantic information. Our method first generates diverse high-frequency semantic images and then complete their semantics to produce the outpainted images. The proposed high-frequency semantic generator based on spatial varying stylemaps helps introduce diversity in the outpainted images. Extensive qualitative and quantitative comparisons show the superiority of our method in terms of diversity and quality of outpainting. However, a limitation of our method is that when the generated HSI is considerably different from the ground truth HSI of the outpainting region, in a few cases, the regions around the object contours can be blurry, and the structural and textural information in the generated HSI can be quite different from the ground truth in these cases. This makes it challenging for the second-stage semantic completion network to generate reasonable semantics. In our future work, we plan to improve the robustness of our method to generate diverse results. Also, a promising future direction could be exploiting HSI for outpainting complex scene images.

References

1. Alharbi, Y., Wonka, P.: Disentangled image generation through structured noise injection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5134–5142 (2020)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
3. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: British Machine Vision Conference. pp. 1–12 (2014)
4. Cheng, Y.C., Lin, C.H., Lee, H.Y., Ren, J., Tulyakov, S., Yang, M.H.: Inout: Diverse image outpainting via gan inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11431–11440 (2022)
5. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. pp. 2672–2680 (2014)
7. Guan, S., Tai, Y., Ni, B., Zhu, F., Huang, F., Yang, X.: Collaborative learning for faster StyleGAN embedding. arXiv preprint arXiv:2007.01758 (2020)
8. Guo, D., Liu, H., Zhao, H., Cheng, Y., Song, Q., Gu, Z., Zheng, H., Zheng, B.: Spiral generative network for image extrapolation. In: European Conference on Computer Vision. pp. 701–717. Springer (2020)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in neural Information Processing systems **30** (2017)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
12. Jo, Y., Yang, S., Kim, S.J.: Investigating loss functions for extreme super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 424–425 (2020)
13. Jo, Y., Park, J.: SC-FEGAN: Face editing generative adversarial network with user's sketch and color. In: 2019 IEEE/CVF International Conference on Computer Vision. pp. 1745–1753 (2019). https://doi.org/10.1109/ICCV.2019.00183
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711. Springer (2016)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
16. Kim, H., Choi, Y., Kim, J., Yoo, S., Uh, Y.: Exploiting spatial dimensions of latent in GAN for real-time image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 852–861 (2021)

17. Kim, K., Yun, Y., Kang, K.W., Kong, K., Lee, S., Kang, S.J.: Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2122–2130 (2021)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Kopf, J., Kienzle, W., Drucker, S., Kang, S.B.: Quality prediction for image completion. ACM Transactions on Graphics (ToG) **31**(6), 1–8 (2012)
21. Lin, C.H., Lee, H.Y., Cheng, Y.C., Tulyakov, S., Yang, M.H.: InfinityGAN: Towards infinite-resolution image synthesis. arXiv preprint arXiv:2104.03963 (2021)
22. Lin, H., Pagnucco, M., Song, Y.: Edge guided progressively generative image outpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 806–815 (2021)
23. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J., Jiang, B., Liu, W.: Deflocnet: Deep image editing via flexible low-level controls. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10765–10774 (2021)
24. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are GANs created equal? a large-scale study. arXiv preprint arXiv:1711.10337 (2017)
25. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: EdgeConnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
26. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 6th Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
27. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical VQ-VAE. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10775–10784 (2021)
28. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with VQ-VAE-2. Advances in Neural Information Processing Systems **32** (2019)
29. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: EnhanceNet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4491–4500 (2017)
30. Shan, Q., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Photo uncrop. In: European Conference on Computer Vision. pp. 16–31. Springer (2014)
31. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9243–9252 (2020)
32. Sivic, J., Kaneva, B., Torralba, A., Avidan, S., Freeman, W.T.: Creating and exploring a large photorealistic virtual space. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8. IEEE (2008)
33. Teterwak, P., Sarna, A., Krishnan, D., Maschinot, A., Belanger, D., Liu, C., Freeman, W.T.: Boundless: Generative adversarial networks for image extension. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10521–10530 (2019)
34. Van Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International Conference on Machine Learning. pp. 1747–1756. PMLR (2016)

35. Wang, M., Lai, Y.K., Liang, Y., Martin, R.R., Hu, S.M.: BiggerPicture: data-driven image extrapolation using graph matching. ACM Transactions on Graphics **33**(6) (2014)
36. Wang, Y., Wei, Y., Qian, X., Zhu, L., Yang, Y.: Sketch-guided scenery image outpainting. IEEE Transactions on Image Processing **30**, 2643–2655 (2021)
37. Wang, Y., Tao, X., Shen, X., Jia, J.: Wide-context semantic image extrapolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1399–1408 (2019)
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
39. Xia, X., Xu, C., Nan, B.: Inception-V3 for flower classification. In: 2nd International Conference on Image, Vision and Computing. pp. 783–787 (2017)
40. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5840–5848 (2019)
41. Yang, C.A., Tan, C.Y., Fan, W.C., Yang, C.F., Wu, M.L., Wang, Y.C.F.: Scene graph expansion for semantics-guided image outpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15617–15626 (2022)
42. Yang, Z., Dong, J., Liu, P., Yang, Y., Yan, S.: Very long natural scenery image prediction by outpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10561–10570 (2019)
43. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5505–5514 (2018)
44. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480 (2019)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
46. Zhang, Y., Xiao, J., Hays, J., Tan, P.: FrameBreak: Dramatic image extrapolation by guided shift-maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1171–1178 (2013)
47. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: UCTGAN: Diverse image inpainting based on unsupervised cross-space translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5741–5750 (2020)
48. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1438–1447 (2019)
49. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2017)
50. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Multimodal image-to-image translation by enforcing bi-cycle consistency. In: Advances in Neural Information Processing Systems. pp. 465–476 (2017)
51. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. Advances in Neural Information Processing Systems **30** (2017)