ReGO: Reference-Guided Outpainting for Scenery Image

Yaxiong Wang, Yunchao Wei, Xueming Qian, *Member, IEEE*, Li Zhu, and Yi Yang, *Senior Member, IEEE*

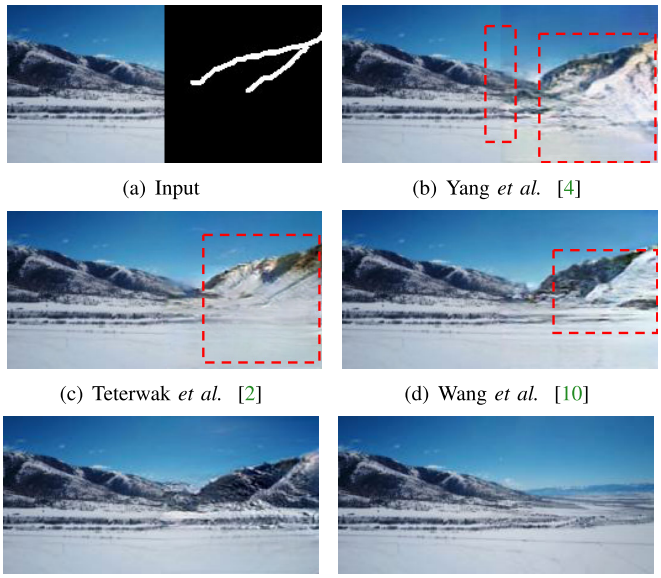
***Abstract*— We present ReGO (Reference-Guided Outpainting), a new method for the task of sketch-guided image outpainting. Despite the significant progress made in producing semantically coherent content, existing outpainting methods often fail to deliver visually appealing results due to blurry textures and generative artifacts. To address these issues, ReGO leverages neighboring reference images to synthesize texture-rich results by transferring pixels from them. Specifically, an Adaptive Content Selection (ACS) module is incorporated into ReGO to facilitate pixel transfer for texture compensating of the target image. Additionally, a style ranking loss is introduced to maintain consistency in terms of style while preventing the generated part from being influenced by the reference images. ReGO is a model-agnostic learning paradigm for outpainting tasks. In our experiments, we integrate ReGO with three state-of-the-art outpainting models to evaluate its effectiveness. The results obtained on three scenery benchmarks, *i.e.* NS6K, NS8K and SUN Attribute, demonstrate the superior performance of ReGO compared to prior art in terms of texture richness and authenticity. Our code is available at https://github.com/wangyxxjtu/ReGO-Pytorch.**

***Index Terms*— Image outpainting, GAN, generation model, adversarial learning.**

I. Introduction

THE task of image outpainting involves generating plausible visual content beyond the boundaries of the input image. Traditional approaches such as [5], [6], [7], and [8] rely on a simple pipeline that involves searching and stitching of image patches to the original input image for extrapolation. However, these solutions are inflexible and may not meet practical requirements. In recent times, researchers have turned to generative adversarial networks (GANs) [15], [18] for the synthesis of unseen content outside the input image boundaries through adversarial learning [1], [2], [4], [9], [39], [50]. For instance, Zongxin et al. [4] propose a recurrent framework that predicts new content for the given image patch, while Teterwak et al. [2] translate the input image to a larger picture with new content beyond the boundary. Wang et al. [10] take image outpainting a step further by introducing sketch-based clues to control the synthesis procedure.

In comparison with the random outpainting, sketch-guided image outpainng is a challenging yet meaningful task. While current methods are capable of producing coherent content for a given image patch, the results are not always satisfactory due to the lack of texture details. As illustrated in Fig. 1(b)-1(d), the outpainting results generated by the state-of-the-art methods [2], [4], [10] generally succeed in synthesizing the desired images to match the guided sketches. However, closer inspection reveals poor quality of the generated regions, including pixels with fewer texture particulars and blurry boundaries between different semantic regions. Consequently, the overall outpainting results lack authenticity.



(a) Input (b) Yang *et al.* [4] (c) Teterwak *et al.* [2] (d) Wang *et al.* [10] (e) Ours (f) original image

Fig. 1. Comparisons of sketch-guided image outpainting. All methods except ours suffer from the lack of the texture details and blurry boundaries of different semantic regions. The dashed boxes indicate the blurry regions.

Manuscript received 2 June 2023; revised 28 October 2023; accepted 16 January 2024. Date of publication 1 February 2024; date of current version 14 February 2024. This work was supported in part by NSFC Project under Grant 62302140; in part by NSFC under Grant 62272380 and Grant 62103317; in part by the Science and Technology Program of Xi'an, China, under Grant 21RGZN0017; and in part by the National Key Research and Development Program of China under Grant 2022ZD0118201. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiantao Zhou. *(Corresponding authors: Yunchao Wei; Xueming Qian; Li Zhu.)*

Yaxiong Wang is with the School of Computer and Information Science, Hefei University of Technology, Hefei 230000, China (e-mail: wangyx15@stu.xjtu.edu.cn).

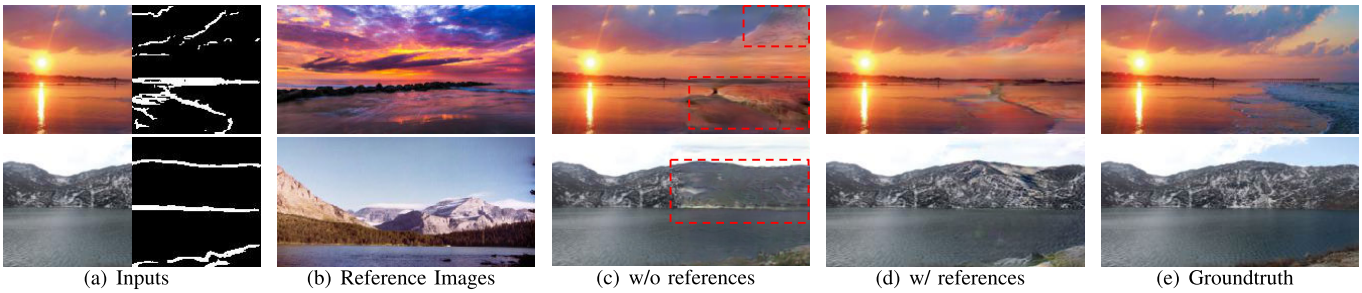
Yunchao Wei is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100000, China (e-mail: wychao1987@gmail.com).

Xueming Qian is with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, and the SMILES Laboratory, Xi'an Jiaotong University, Xi'an 710049, China, and also with Zhibian Technology Company Ltd., Taizhou 317000, China (e-mail: qianxm@mail.xjtu.edu.cn).

Li Zhu is with the School of Software, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zhuli@mail.xjtu.edu.cn).

Yi Yang is with the School of Computer Science and Technology, Zhejiang University, Hangzhou 310000, China (e-mail: yee.i.yang@gmail.com).

Digital Object Identifier 10.1109/TIP.2024.3357290



(a) Inputs (b) Reference Images (c) w/o references (d) w/ references (e) Groundtruth

Fig. 2. The outpainting examples with (w/) and without (w/o) the reference images. Although the model abandoning the reference images could predict reasonable pixels for the inputs, but its outpainting results suffer from the lack of textural details. The neighboring images share many pixels with the image to be extended, allowing models to borrow valuable pixels from the neighboring images and produce outpainting results with rich textures.

Intuitively, landscape photos typically exhibit similar layouts and appearances to the photos in the same scene. As shown in the top row of Fig. 2, both the input patch and the reference image show the sunset-related scene, and there are many valuable pixels in the reference image aiding in the synthesis of high-quality content for the input patch. Therefore, if we can successfully transfer the knowledge from similar photos to complement the textural details of the predicted content, the authenticity of the generated part can be significantly improved. Straightforwardly, the input image itself is a natural choice for serving as the reference, since it often contains content-consistent pixels to the outpainting part. However, simply adopting the input image for referring often limits the diversity of sketch layout or content pattern of the outpainting part, leading to poor generalization ability, especially for the free-form outpainting.

Motivated by the aforementioned observations and considerations, this work explores a principle for synthesizing highly detailed outpainting results by utilizing pixels from the neighboring images, also called reference images, as guidance. We refer to this approach as Reference-Guided Outpainting (ReGO). However, the reference images inevitably contain some inconsistent content. Therefore, transferring these pixels without adaptive filtering would introduce abrupt pixels and subsequently degrade the quality of the generated content. Consequently, the main challenge of ReGO lies in effectively transferring pixels from neighbors while maintaining the style consistency with the input image.

To this end, an Adaptive Content Selection (ACS) module is first proposed to augment our ReGO. Concretely, an image-guided convolution is first conducted on the reference image to select the compensatory features, and two feature fusion blocks are followed for guiding sketch fusing and boundary stitching, respectively. With the ACS module, our ReGO can effectively filter out the abrupt or profitless contents and only adopt the beneficial features to synthesize texture-rich results. In this context, "beneficial features" refers to the features derived from parts of the reference images that resemble the synthesized content, whereas "non-beneficial features" pertains to the features extracted from sections of the reference images that do not contribute to the synthesized content. Besides, the introduced reference image is only responsible for contributing contents to enrich the final outpainting results, while the style of the synthesized part should keep consistent with the input, instead of being affected by the reference. To achieve the style consistency, we further utilize a hinge-based ranking loss to pull the style of generated part close to the input image and prevent the style of generated image from biasing to the reference image. Particularly, the generated part and the input patch are treated as the positive pair, while the reference image is regarded as the negative sample, then, the triplet loss is imposed to constrain their style representations. The style ranking loss could make our system avoid abrupt pixels and synthesize smoother outpainting results with a cohesive visual style. We perform experiments on three popular benchmarks, NS6K [4], NS8K [10], and SUN Attribute [24]. Extensive quantitative and qualitative comparisons can well demonstrate the superiority of our ReGO over other state-of-the-art approaches.

In summary, the primary contributions of this work include:

- **A new Outpainting framework**. We propose ReGO, a new outpainting method that introduces a reference image as guidance to transfer content-consistent pixels and synthesize texture-rich outpainting results.
- **Adaptive content selection (ACS) module**. ACS module is designed to pick up the beneficial features from the reference image, making ReGO could filter out abrupt pixels and generate semantic-consistent results.
- **Style ranking loss (SRL)**. SRL is proposed to restrain the style of the synthesized part and enables the system to synthesize style-consistent results.
- **Competitive performance** on both random outpainitng and sketch-guided outpainting tasks.

II. Related Work

A. Image Inpainting

The image inpainting has been well explored recent years, whose target is to restore the missing or corrupt regions in images [11], [25], [27], [30], [31], [32], [36], [38], [41], [46], [47], [48]. Benefit from the tremendous success of the generation adversarial networks (GANs) [15] and diffusion models [42], [43], [47], the image inpainting has made great advances these years. In the early exploratory stage of this task, the researchers target on the missing regions with formal shapes [37], [40], the core idea is to collect information from the surrounding context to restore the missing pixels. Liu et al. develop a novel operation named partial convolution, iteratively predicting the missing pixels by collecting information from the surrounding content [35]. With the technique developing, the community pays more attention to the free-form inpainting problems [30], [32], [35], [40], [31], [49]. In [31],

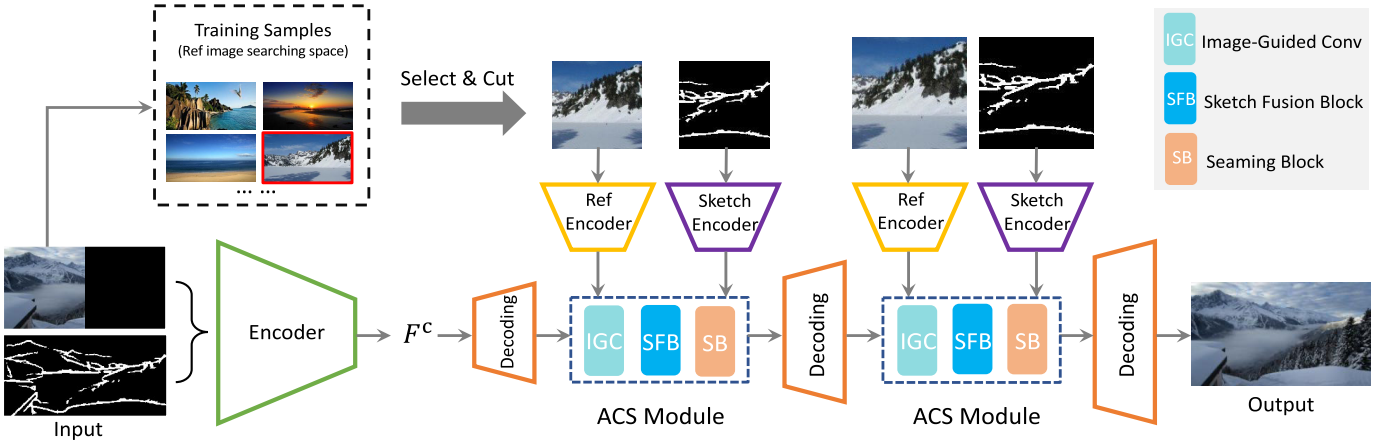


Fig. 3. The overview of the proposed ReGO. The ReGO system takes the left image & the sketch as inputs, and synthesizes the additional right half new content for the input image. The overall architecture follows an encoder-decoder paradigm, where the encoder compresses the inputs and obtain the hidden feature F, and the decoder is responsible to rebuild the complete image from F. Meanwhile, our Adaptive Content Selection (ACS) module could be equipped in each decoder layer, a reference image is first selected from the training samples, whose right half is further cut and fed into the proposed ACS module together with the guiding sketch to replenish the hidden features. As for the guiding sketch, we use the groundtruth from the right half image during training. At the testing stage, the guiding sketch could be manually drawn or borrowed from other image, as shown in Fig.7.

Xie et al. propose a feature re-normalization to adapt to the irregular holes. In [30], Guo et al. propose a full-resolution residual network (FRRN) to restore the missing pixels with irregular shape. Comparing to the inpainting, the missing pixels of outpainting task are far from the valid content, posing more challenges to restore.

B. Image Outpainting

Conventional image outpainting methods follow a search-and-compose pipeline, where the potential patches are first selected from an external library and then stitched with the input image to conduct extrapolation [3], [6], [8], [12], [7], [13]. Inspired by the success of the deep neural network, researchers recently attempt to predict new content beyond the boundaries using the generative adversarial networks (GANs) [2], [4], [15] and diffusion models [42], [45]. For example, Teterwak et al. [2] and Zongxin et al. [4] use encoder-decoder based generator to predict the unseen pixels. In [39], Yao et al. study the potential of transformer for outpainting and develop an transformer-based outpainting framework. Recently, the outpainting has been promoted by the powerful diffusion model [42], [43], [44], [45], Nuwa-Infinity [45] builds a diffusion-based outpainting model that is capable of extending images with very high resolutions over long distances. However, these methods could only predict random contents, to address this weakness, the conditional image outpainting starts to be studied recently. Wang et al. [10] develop a network that allows users to guide the final synthesis by free-form sketches. In [9], the authors propose to utilize the language and the position clues to control the outpainting results. Despite existing methods, the synthesized results still suffer from blurred texture, which motivates us to develop a controllable outpainting framework capable of synthesizing high-quality images with rich texture.

III. Methodology

The Overview. The primary objective of this study is to enhance the quality of sketch-guided image outpainting by enriching its texture. The pipeline of our outpainting system can be seen in Fig. 3. The architecture operates on an encoder-decoder paradigm where our proposed Adaptive Content Selection (ACS) module is integrated into each decoder layer. The left half image and sketch comprise the inputs to the encoder, and the output is the hidden feature map F, which is transmitted to downstream decoder layers to build the complete image. To synthesize the texture-rich results, the reference image is first chosen from the searching space. *i.e.* , training samples and fed into the ACS module to distill its content for compensating purpose. Besides, to allow the user to harvest the freestyle outpainting, the guiding sketch clue is also integrated to build a flexible system. Fig. 4 shows the details of the ACS module. The image-guided convolution is first employed to distill the beneficial features from the reference image, then the selected features are integrated with the hidden representations of the synthesized part to complement the texture details. In addition, the style ranking loss is designed to encourage the generator to produce style-consistent content.

A. Data Preparation

In our system, an image I from the training set X requires two auxiliary data inputs: the corresponding sketch and reference image. The sketch acts as a conditional clue to guide the synthesis process, as demonstrated in Fig. 1. On the other hand, the reference image is utilized to provide comprehensive detailed features essential for generating texture-rich new content.

To obtain the sketch, we utilize the HED edge detector [16] to extract the edge map, which is then binarized with a predetermined threshold (0.6 in our experiments) to produce the binary sketch $S \in \mathbb{R}^{H\times W\times 1}$. During training, the network is trained to restore the ground-truth parts using original sketches of missing parts drawn from the training data. During testing, users can input manually drawn free-form sketches to synthesize desired results.

To acquire reference images for input I, we begin by extracting feature representations through a pre-trained model,

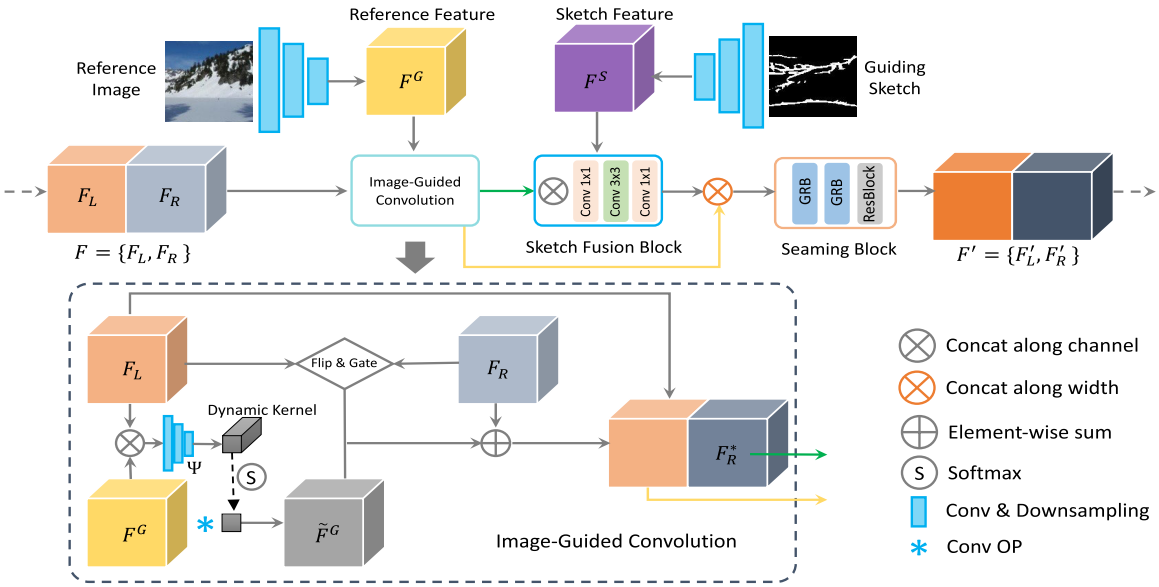


Fig. 4. Our proposed ACS module architecture involves distilling profitable features from the reference image and using them to replenish predicted features. Sketch fusing then follows to combine sketch clues, creating a controlled outpainting system. Lastly, a seaming block ensures a seamless transition between left and right content.

Places365 [17], followed by the utilization of cosine similarity to identify similar samples. Noting that only the left half image is used for similarity calculation. We have observed that visual neighbors typically share similar content with the target image and tend to have more beneficial pixels, thereby making them suitable candidates for reference images. In our practice, we select multiple neighbors for each sample and randomly pick up one in each training iteration as the reference image G. Besides, since we only need to replenish the details of synthesized part, the right half of the reference image is only considered for further processing, as shown in Fig. 4.

B. ACS Module

Given the reference image and guiding sketch, our ACS module has two primary duties: to enhance synthesized new content texture details and combine conditional information from the guiding sketch. To fulfill these duties, we design an image-guided convolution to enrich the details, and a sketch fusing block is utilized to integrate the sketch clue. Furthermore, we implement a seaming block to ensure a uniform boundary between the original left feature and predicted new contents. Subsequently, we elaborate on the details of each block.

Image-Guided Convolution. The proposed Image-Guided Convolution (IGConv) aims to help the network complement texture details for the new content using the distilled the beneficial features from the reference image. Intuitively, the left part of the input could directly serve as the reference image. However, based on our experience, such a strategy often results in an insufficiently diverse training set. Consequently, the trained model becomes overly reliant on the original sketch layout and content pattern, rendering it incapable of well generalizing to freestyle outpainting scenarios. Therefore, in our Image-Guided Convolution,we opt to search for multiple reference image neighbors and select one randomly in each iteration. Such a training fashion allows the model to see diverse input-reference pairs, thus, the generality of the final model could be boosted accordingly.

Formally, let $F_L \in \mathbb{R}^{h\times w\times c}$ be the features encoded from the image to be extended, F_R represents the hidden features for the predicted new content. F_L and F_R form the complete hidden features of the overall image $F \in \mathbb{R}^{h\times 2w\times c}$. And the features of G, which are encoded from a reference image encoder, are denoted as $F^G \in \mathbb{R}^{h\times w\times c}$. The designed image-guided convolution aims to complement F_R by extracting helpful information from reference features F^G. A group of dynamic filters are conditionally produced based on the features of input and reference images, making the network adaptively collect the beneficial content from the reference image. To be specific, a dynamic kernel is produced based on the concatenation of F^G and F_L via a simple feed-forward procedure:

$$k = \Psi(F^G, F_L), \tag{1}$$

where $k \in \mathbb{R}^{3\times 3\times c}$, 3 and c indicates the kernel size and channel number, respectively. The Ψ can be modeled as the neural network, which takes the features of the reference and the input and recurrently use the convolution with stride=2, batch normalization and ReLU[1] to get the features.

The dynamic kernel in Eq. 1 targets on providing guidance to distill the content of the reference image. To adaptively pick up the beneficial pixels and restrain the unhelpful content, we conduct the channel-wise normalization to update the dynamic kernels:

$$k^n_{ijk} = \frac{\exp(k_{ijk})}{\sum_h \exp(k_{ijh})}. \tag{2}$$

Thus, the distilled i-th channel map can be obtained as follow:

$$\widetilde{F}^G_{:,:,i} = F^G * P(k^n_{:,:,i}), i = 1, 2, \ldots, c, \tag{3}$$

[1]ReLU is not applied in the last layer.

where the $*$ denotes the convolution operation, $P(\cdot)$ is an operation to repeat $k^n_{:,:,i}$ c times along the channel dimension. All of the $\widetilde{F}^G_{:,:,i}|^c_{i=1}$ are channel-wise stacked to get the distilled feature map $\widetilde{F}^G \in \mathcal{R}^{h\times w\times c}$.

The dynamic kernel in Eq. 2 aims to selectively emphasize useful features and downplay unprofitable content through the softmax operation, resulting in the network assigning lower weights to unhelpful features and higher weights to helpful ones. The distilled convolution in Eq. 3 then endeavors to succinctly summarize the advantageous semantic regions across feature channels, based on each dynamic kernel's viewpoint. Such a procedure effectively gathers the beneficial features from the reference image and combines them into the map $\widetilde{F}^G$, which are further added to the feature F_R to achieve the compensatory purpose: $F^*_R = F_R + \widetilde{F}^G$.

In addition to the reference image, the input image itself also has the potential to contribute useful pixels. This is due to the synthesized portion highly likely containing the same objects as those present in the input image. Therefore, the features from the input image are also integrated and the F^*_R is updated as follow:

$$F^*_R = F_R + \widetilde{F}^G + F_L \times \sigma(\rho(F_L) \times F_R), \tag{4}$$

where $\rho(\cdot)$ is the horizontally flip operation, $\sigma(\cdot)$ denotes the sigmoid function, which is introduced to learn a dynamic feature selection mechanism.

Sketch Fusing Block. Besides synthesizing the unseen part with thriving and realistic details, our ReGO should also be equipped with a practical mechanism, *i.e.* , allowing users to acquire personal custom outpainting results using their preferred sketches as the guidance. To this end, we introduce a controllable sketch fusion block to achieve the target. To make the final results exactly match the guiding sketch, the sketch fusion block additionally integrates the sketch feature to emphasize the desired shape in the restoring procedure, as shown in Fig. 4.

Concretely, only the right half sketch $S^r \in \mathbb{R}^{H\times W/2\times 3}$ serves as the guiding clues, and its feature maps F^s are first encoded by a sketch encoder E^S. Then, the compressed sketch features are channel-wise concatenated with the complemented feature F^*_R, and fed forward a residual block [20] style structure to get the fused output F^s_R.

Seaming Block. Our seaming block is responsible to fuse the raw left half features F_L and the complemented features F^s_R, which in fact attempts to smooth the boundary between the raw features from the input image and the complemented right half features. As shown in Fig. 4, the seaming block consists of two global residual blocks (GRB) [4] and a residual block [20]. We alternately utilize the 1×3 and 7×1 convolution in GRB to strengthen the connection between the original and the predicted regions, especially the boundary between the map from the input image and the complemented map of the predicted new content. Particularly, the F_L and F^s_R are first concatenated along the width dimension, and then sequentially fed through two GRBs and a residual block to get the output F', which is also the final output of our ACS module.

C. Style Ranking Loss

The reference image utilized by ReGO is intended solely to provide texture details, and its style should not be reflected in the synthesized content. To reduce the artifacts, the model should 1) only transfer the texture details from the reference image, 2) keep a consistent style between the given part and the generated part. Given the above considerations, we conclude that hinge-based ranking is well-suited to our needs. We treat the synthesized part and the input image as the positive pair, while treating the reference image as the negative sample. The hinge ranking loss is then applied to their style representations to enforce the style of the input and the new content to be more similar to each other than to that of the reference image.

Following previous practices [21], [22], [23], we utilize the second-order statistics of convolutional feature as style representation. Particularly, the style features of generated part $\hat{I}^r \in \mathbb{R}^{H\times W/2\times 3}$, which is only the right half of the image reconstruction $\hat{I} \in \mathbb{R}^{H\times W\times 3}$, is given by the Gram matrix $R^d \in \mathcal{R}^{N_d\times N_d}$:

$$R^d_{ij} = \sum_k M^d_{ik} M^d_{jk}, \tag{5}$$

where M^d_i is the vectorised i-th feature map in layer d from a convolutional neural network like VGG19 [26], N_d indicates the channel number of layer d.

Analogously, the style representations of the reference image and the input image can be extracted, and our style ranking loss is defined as:

$$\mathcal{L}^d_s = [\alpha - SM(R^d, L^d) + SM(R^d, G^d)]_+, \tag{6}$$

where L^d and G^d represent the Gram matrixs of the left input and the reference image, respectively, $SM(\cdot,\cdot)$ is the cosine similarity, $\alpha \in \mathcal{R}$ is the scalar margin, and $[\cdot]_+ = max(\cdot, 0)$.

By including the feature correlations of multiple layers, the multi-scale style representations are obtained, and the total style loss can be calculated accordingly:

$$\mathcal{L}_s = \sum_{d\in D} w_d \mathcal{L}^d_s, \tag{7}$$

where D is the index collection of selected activation layers, and w_d is the trade-off weight. In our experiments, the activated output of layer relu_Y_1(Y=1,2,3,4,5) of VGG19 network [26] are taken for style representation, *i.e.* , $|D| = 5$. The designed style ranking loss is equipped to the generator loss to train the network.

IV. Experiment

A. Experiment Setup

Dataset. The focus of this research is image outpainting for scenery images, for two main reasons: firstly, the sketches for scenery images are relatively easy to create, and secondly, the pioneering sketch-guided image outpainting work SGIO [10] evaluates its performance on this type of images, which we follow in our experimentation. We conduct extensive experiments on three benchmarks, *i.e.* , NS6K [4], NS8K [10], and SUN Attribute [24], to validate the effectiveness of our

ReGO. The **NS6K** dataset comprises a total of 6,040 scenery images, of which 5,040 are utilized as training data, and the remaining 1,000 are reserved for testing [4]. The **NS8K**, which consists of 8,115 images, contains more diverse scenery images comparing to NS6K. Of these, 6115 images are taken as training data, the rest is used for testing. The **SUN Attribute** dataset has 14,340 diverse enough images from 707 scene categories, we randomly select 80% and 20% for training and testing, respectively.

Implement Details. Our proposed ReGO offers a model-agnostic solution that can be easily incorporated into various off-the-shelf outpainting models. In this work, we apply our ReGO to three state-of-the-art outpainting methods, including NSIO [4], BDIE [2], and SGIO [10], to validate its superiority:

NSIO [4] is originally designed for random content prediction. To achieve the sketch-guided outpainting, we make some modifications for NSIO as follow: the left half sketch is channel-wise concatenated with the input, while the right half sketch is encoded and fed as the initial state of LSTM [14] decoder to predict the hidden feature of the full images. Our ACS module is plugged after each decoding layer except the last one, and the style ranking loss is weighted by 0.5 and added into the generator loss to train the network.

BDIE [2] is a random-outpainting model as well, and the sketch is concatenated with the input to perform the sketch-guided outpainting. Besides, the conditional skip connection and the position channels in SGIO [10] are also equipped to BDIE [2], to build a stronger baseline. The ACS module and the style ranking loss are equipped in an analogous way with NSIO [4].

SGIO [10] is the first attempt for the sketch-guided outpainting. The ACS module is also employed after each decoding layer to enable texture compensation, and the style ranking loss is added to the generator loss with weight 0.5 to ensure the style consistency.

For our study, the style ranking loss is equipped to the generator loss, and the weights of style ranking loss in multiple layers are all set as 0.2, *i.e.* , $w_d = 0.2$. During the training stage, five neighbors are employed in our baseline methods, and the impact of neighbor number will be discussed in the following experiment. At the testing stage, only the most similar neighbor is used to synthesize the outpainting. Besides outpainting models, we also include three state-of-the-art inpainting models for comparison, *i.e.* **DeepFillv2** [32], **CoModGAN** [34], and **LaMa** [33]. For LaMa and CoModGAN, we mask the right half images and introduce the sketch as an additional channel to train the network, while DeepFillV2 is a sketch-guided inpainting model, and it's trained by only restoring the right half image. To make a fair comparison, the loss functions, the hyperparameters and the training details all follow the same settings of their original papers. The sketch augmentation strategy [10] is also employed for all methods to enhance the free-form outpainting.

B. Evaluation Metric

Following Wang et al.'s [10] setting, three metrics, *i.e.* , Fréchet Inception Distance (FID) [28], the Inception Score (IS) [29] and Mean Satisfactory Degree (MSD) [10], are employed for evaluation. To evaluate the free-form outpainting results, we randomly select 555 images from test data and replace the original sketches with manually drawn free-form ones. Finally, a collection of 89 distinct sketches is assembled, and can be broadly categorized into two groups. The first group contains sketches that are similar to the training samples, but are highly simplified. The second group consists of entirely new sketch patterns, such as circles, hearts, checkmarks, *etc*, which are diverse enough to assess the capability of free-form outpainting. We invited 20 volunteers to label the free-form outpainting results as one of three levels: 0-poor, 1-ordinary, and 2-good. The mean satisfaction degree (MSD) was computed as the average of all the assigned labels for the selected images, which is taken for subjective comparison since there is no groundtruth available. Comparing to the FID and the IS, MSD directly reflects the performance in practical situations, therefore, it is a critical metric to evaluate the generalization ability on free-form sketches.

C. Quantitative Comparison

Table I presents the performance of both sketch-guided outpainting and random outpainting on the NS6K and NS8K datasets. The notation $\text{ReGO}_{\text{NSIO}}$ corresponds to the NSIO model equipped with our proposed ReGO module. The results demonstrate that our proposed ReGO module can effectively improve both sketch-guided outpainting and random outpainting performance.

Sketch-Guided Outpainting. Our proposed ReGO module could boost the performance of three state-of-the-art outpainting methods on both NS6K and NS8K. For example, when applied to BDIE [2] on NS6K, our ReGO module results in a reduction of FID from 11.021 to 10.052. In addition to improving image restoration based on original sketches, our ReGO module also enhances free-form outpainting results. The MSD of $\text{ReGO}_{\text{SGIO}}$ can reach 1.201 on NS6K, while the original SGIO's is only 1.01. Our best performance is achieved based on the BDIE [2], which could reach 10.052 FID and 1.357 MSD on NS6K. Table I clearly demonstrates that our proposed ReGO enhances image restoration and free-form outpainting capabilities across all three backbones. These observations provide compelling validation of the effectiveness of our approach.

To well explore the effectiveness of our reference image, we further study three alternatives of reference types, *i.e.* , non-reference, self-reference (employ the input image itself as reference image), and the searched neighbors as reference. The $\text{ReGO}_{\text{BDIE}}$ without reference will degrade to BDIE*. To validate the pure benefits of reference image, we further remove the style ranking loss from $\text{ReGO}_{\text{BDIE}}$, name the resulted method as $\text{ReGO}_{\text{BDIE}}$-R. To investigate the performance differences resulting from utilizing three different types of references, we conducted experiments on the NS6K dataset using the BDIE backbone. Let $\text{ReGO}_{\text{BDIE}}$-SR (**S**elf-**R**eference) denotes the $\text{ReGO}_{\text{BDIE}}$ with the input as reference. From the comparison of Table I, we can find the methods with reference all outperform the non-guidance method "BDIE*". For example, the FID of BDIE* on NS6K is 11.02, while

TABLE I

PERFORMANCE COMPARISONS ON THREE DATASETS UNDER CRITERIA IS, FID AND MDS, FOR SKETCH-GUIDED AND RANDOM OUTPAINTING TASKS. ∗ MEANS THE METHOD IS MODIFIED TO PERFORM SKETCH-GUIDED OUTPAINTING AS DESCRIPTED IN SUBSECTION IV-A. WE ADOPTED THE SAME BACKGROUND FOR METHODS WITH THE SAME BACKBONE TO MAKE THE COMPARISON CLEARER

	Sketch-Guided Outpainting								
	NS6K			NS8K			SUN Attribute		
	IS↑	FID↓	MSD↑	IS↑	FID↓	MSD↑	IS↑	FID↓	MSD↑
DeepFillv2* [32]	2.316	16.712	0.511	3.012	15.132	0.592	9.132	27.993	0.562
CoModGAN* [34]	2.758	15.145	0.583	3.221	13.774	0.685	9.884	26.076	0.641
LaMa* [33]	3.016	13.639	0.567	3.347	12.312	0.629	10.147	24.135	0.619
NSIO* [4]	2.891	12.870	0.649	3.250	10.813	0.837	10.391	23.021	0.801
SGIO [10]	2.920	10.998	1.010	3.321	10.390	1.170	10.126	22.536	0.792
BDIE* [2]	3.002	11.021	0.963	3.323	9.639	0.892	10.857	21.412	0.886
$\text{ReGO}_{\text{NSIO}}$	2.923	12.030	0.839	3.329	10.232	0.966	10.683	21.452	0.869
$\text{ReGO}_{\text{SGIO}}$	2.924	10.104	1.201	3.387	9.787	1.293	10.542	21.012	0.894
$\text{ReGO}_{\text{BDIE}}$-SR	3.042	10.561	1.012	3.392	9.047	1.015	10.719	20.779	0.889
$\text{ReGO}_{\text{BDIE}}$-R	3.038	10.269	1.221	3.417	8.813	1.263	10.906	20.332	0.978
$\text{ReGO}_{\text{BDIE}}$	**3.126**	**10.052**	**1.357**	**3.444**	**8.738**	**1.396**	**10.992**	**19.433**	**1.014**
Random Outpainting									
DeepFillv2 [32]	2.273	18.693	-	2.816	18.109	-	8.987	29.016	-
CoModGAN [34]	2.612	18.014	-	2.983	17.223	-	9.763	27.973	-
LaMa [33]	2.773	15.061	-	3.019	15.014	-	9.974	25.863	-
NSIO [4]	2.883	13.612	-	3.123	12.871	-	10.229	25.271	-
SGIO [10]	2.951	15.857	-	3.178	12.316	-	9.889	24.978	-
BDIE [2]	2.880	13.252	-	3.155	11.373	-	10.647	24.465	-
$\text{ReGO}_{\text{NSIO}}$	2.792	14.224	-	2.948	13.230	-	10.553	26.196	-
$\text{ReGO}_{\text{SGIO}}$	2.848	15.396	-	3.204	13.748	-	10.438	25.441	-
$\text{ReGO}_{\text{BDIE}}$-SR	3.119	13.174	-	3.388	11.023	-	10.623	23.788	-
$\text{ReGO}_{\text{BDIE}}$-R	3.178	12.829	-	3.514	10.806	-	10.696	23.509	-
$\text{ReGO}_{\text{BDIE}}$	**3.243**	**12.606**	-	**3.573**	**10.586**	-	**10.796**	**23.209**	-

$\text{ReGO}_{\text{BDIE}}$-SR's, $\text{ReGO}_{\text{BDIE}}$-R's, and $\text{ReGO}_{\text{BDIE}}$ can attain 10.561 10.269, 10.052 respectively. This reveals that the reference image is a reliable and effective clue to boost the performance. $\text{ReGO}_{\text{BDIE}}$-SR can also achieve acceptable performance on image rebuilding according to the original sketches, however, it shows poor generality when encountering the free-style sketches. For example, the FID of $\text{ReGO}_{\text{BDIE}}$-SR could reach 10.561 on NS6K and surpasses the method BDIE, however, its MDS for free-form outpainting is only 1.012, which is much worse than $\text{ReGO}_{\text{BDIE}}$. We think the barren sketch layout and content pattern cause somewhat overfitting, consequently, the model trained with self-reference could not well generalize to the free-style outpainting. In contrast, When the neighbors serves as the reference, the model could see diverse training pairs, as a result, the trained model could perform well on both the image rebuilding and free-form outpainting.

Pair-wise Study for Free-form Outpainting. To conduct a comprehensive study of focused free-form outpainting, we further conduct a pair-wise comparison to evaluate the effectiveness of our ReGO. For two methods, A and B, their corresponding result pair (I_A, I_B) is first displayed. Method A is assigned a score of 1 if I_A is better than I_B, -1 if I_A is worse, and 0 if they are comparable. The final score is calculated by averaging across all 555 test samples. The comparison results are presented in Table II, where the column denotes method A. It can be observed that the methods incorporating ReGO outperform their corresponding baseline methods. For instance, $\text{ReGO}_{\text{BDIE}}$ surpasses BDIE* by 0.247,

TABLE II

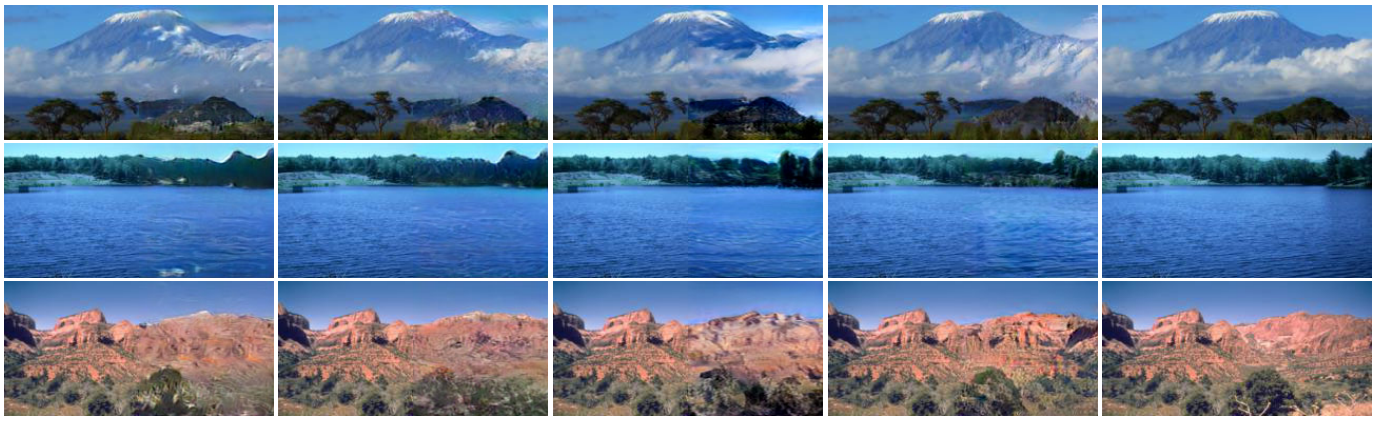
PAIR-WISE COMPARISON FOR FREE-FORM OUTPAINTING

Method	NSIO	SGIO	BDIE*
$\text{ReGO}_{\text{NSIO}}$	0.276	0.215	0.111
$\text{ReGO}_{\text{SGIO}}$	0.309	0.267	0.158
$\text{ReGO}_{\text{BDIE}}$	0.465	0.382	0.247

and all methods with ReGO demonstrate better performance, well validating the effectiveness of our proposed ReGO.

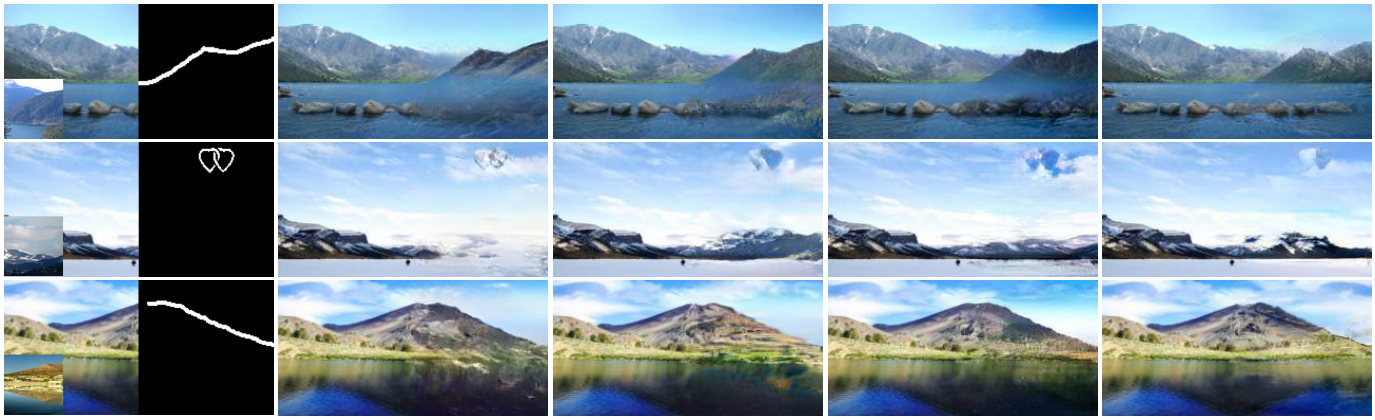
Random Outpainting. Besides providing the sketches to harvest the desired outpainting, another possible scenario is that the users may not wish to drawn any guiding sketches and only attempt to obtain the random results. How would our system perform if no guiding sketches are fed? To validate the effectiveness of our method under such a scenario, we conduct experiments to predict random results and report the performance on both datasets. Since the NSIO [4] and BDIE [2] are originally designed for random outpainting, we follow the same pipelines as their original papers [2], [4] to train the networks. As for the inpainting methods, CoModGAN [34], and LaMa [33], we directly mask the right half of the image for training. For sketch-guided systems, we simply set the right half sketch as zeros to conduct random outpainting without retraining, *i.e.* , *zero-shot*.

The results are also reported in Table I, we can observe that abandoning the original guiding sketches significantly damnify the performance of the sketch-guided outpainting systems. For example, the $\text{ReGO}_{\text{SGIO}}$ with guiding sketches could reach 10.104 FID on NS6K, while its FID w/o the



(a) BDIE [2] (b) +RI (c) +SRL (d) +ACS (e) Groundtruth

I: Visual ablation on image rebuilding.



(a) Inputs (b) BDIE [2] (c) +RI (d) +SRL (e) +ACS

II: Visual ablation on free-form outpainting.

Fig. 5. The visual ablation of each component in our method on image rebuilding and free-form outpainting, where RI, SRL, and ACS indicate the reference image, style ranking loss, and adaptive content selection module, respectively.

guiding sketches deteriorates to 15.396, this is because the systems are trained with the original sketches. The SGIO heavily relys on the guiding sketch, the $\text{ReGO}_{\text{SGIO}}$ inherits this weakness and even more severe, we think this is the reason that $\text{ReGO}_{\text{SGIO}}$ performs worse than the SGIO. For the random prediction methods, $\text{ReGO}_{\text{NSIO}}$ performs slightly worse than the NSIO [4], since the NSIO is retrained with random prediction setting. While $\text{ReGO}_{\text{BDIE}}$ is more outstanding comparing to BDIE [2] and inpainting methods (LaMa [33]). From Table I, we can see that even though no guiding sketches are provided, the methods with our ReGO module could also produce comparable results with the original methods. With the designed ACS module, we can develop an unified framework that could simultaneously deal with the random prediction and the sketch-guided outpainting. what's more, the BDIE with ReGO module could achieve the SOTA performance on both tasks.

D. Ablation Study

Validate the Components of ReGO. To assess the individual contributions of each component in our proposed ReGO model, we utilize BDIE as the backbone and perform ablations on the NS6K dataset. Quantitative results are reported in Table III (a). As shown in Table III (a), when the reference image is introduced, the FID of BDIE could be improved from 11.02 to 10.99 and the MSD is boosted as well, which reveals compensating the texture details from the neighbors is a promising idea. However, only the reference image does not make the performance outstanding enough. The incorporation of the proposed ACS module allows for the network to effectively filter out unnecessary content and emphasize beneficial pixels. As a result, a significant improvement in performance is observed, with the MSD reaching 1.221 and FID reaching 10.269. Incorporating the style ranking loss to ensure style consistency further improves the performance. The model, which utilizes all three parts simultaneously, achieves the best performance. The addition of a new mechanism results in further improvement, which confirms the contributions of each component. Figure 5 displays the visual results of the ablation comparison on image rebuilding and free-form outpainting. The contribution of each component can be clearly observed from the figure.

Validate the Style Ranking Loss. To produce the style-consistent results, a style ranking loss is proposed to prevent the reference image from affecting its style. Alternatively, a regression procedure can also be used to directly achieve a close style between the synthesized content and input.

TABLE III
ABLATION STUDY AND DISCUSSION ON NS6K DATASET WITH $\text{ReGO}_{\text{BDIE}}$ AS THE BASELINE METHOD

	RI	ACS	SRL	IS↑	FID↓	MSD↑
BDIE* [2]				3.002	11.021	0.963
BDIE* [2]	✓			2.918	10.991	1.034
BDIE* [2]	✓	✓		3.038	10.269	1.221
BDIE* [2]	✓	✓	✓	**3.126**	**10.052**	**1.357**

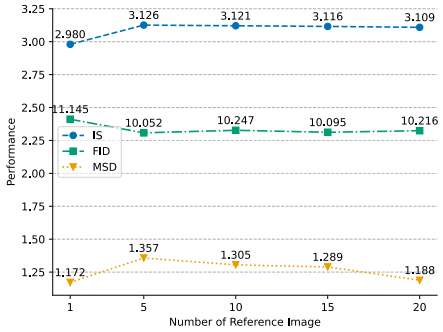
(a): The contributions of each part in our method. RI, ACS, and SRL indicate reference image, ACS module, and style ranking loss, respectively.

Method	IS↑	FID↓	MSD↑
$\text{ReGO}_{\text{BDIE}}$-Reg	3.089	10.956	1.163
$\text{ReGO}_{\text{BDIE}}$	**3.126**	**10.052**	**1.357**

(b): The performance of the style ranking loss and the style regression loss.

	$\text{ReGO}_{\text{NSIO}}$		$\text{ReGO}_{\text{SGIO}}$		$\text{ReGO}_{\text{BDIE}}$	
	Train with Searched References					
Inference	Search	Random	Search	Random	Search	Random
FID↓	12.03	12.635	10.104	10.451	10.052	10.289
IS ↑	2.923	2.871	2.924	2.825	3.126	2.885
MSD ↑	0.839	0.799	1.201	1.031	1.357	1.196
	Train with Random References					
Inference	Search	Random	Search	Random	Search	Random
FID↓	12.791	12.816	10.864	10.871	10.889	10.902
IS ↑	2.892	2.879	2.941	2.935	3.102	2.992
MSD ↑	0.661	0.653	0.997	0.984	0.971	0.969

(c): Performance with the searched and the randomly selected reference images on NS6K.



(d): Performance *w.r.t* reference numbers, we scale the FID by logarithmic function to make the value of three criteria close and exhibit clearer tendency.

In this subsection, we analyze the effects of both solutions. Table III (b) shows the performance under IS, FID, and MSD, where $\text{ReGO}_{\text{BDIE}}$-Reg indicates $\text{ReGO}_{\text{BDIE}}$ using the l_2 style reconstruction loss instead of the style ranking loss. From Table III (b), the $\text{ReGO}_{\text{BDIE}}$ with the proposed style ranking loss performs much better on both image restoring and free-form outpainting. The observed results may be attributed to overfitting. Despite being from the same image, the style representations captured by the Gram matrices in the left half and right half parts are different. Thus, performing a rigid regression procedure can easily result in overfitting. Furthermore, the primary objective of the style ranking loss is to prevent the reference image's style from being reflected in the extended content. It is not necessary to enhance the style consistency between the input and the synthesized part, as it can be achieved through pixel-wise reconstruction and adversarial training [2], [4], [10].

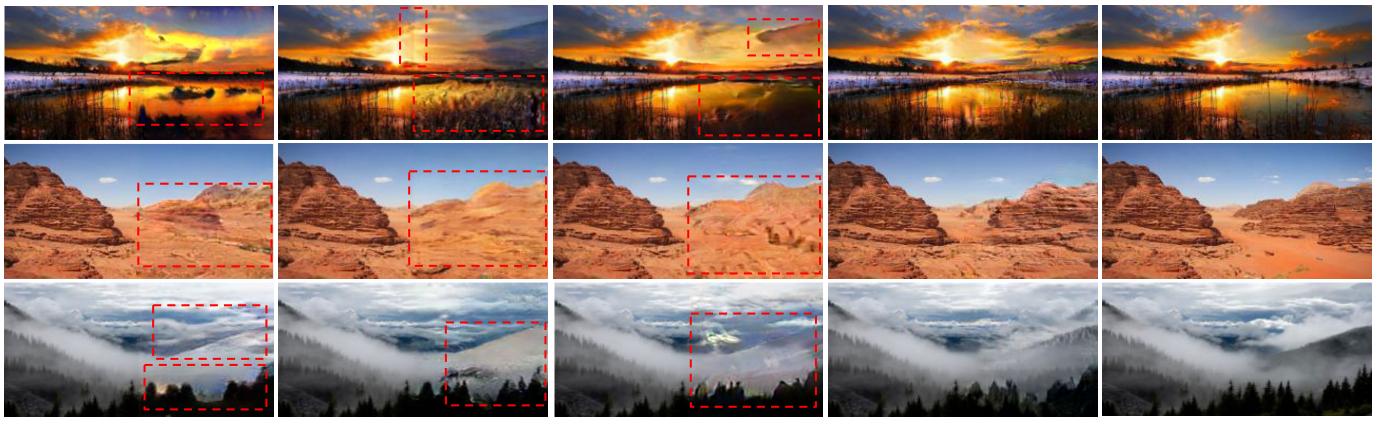
Robustness about the Reference Image. The reference image serves as a basis for enriching the outpainting details and plays a crucial role in our system. However, the ideal reference image is not always available for the current test image. Hence, an intuitive question arises: What happens if the model utilizes a "bad" reference image? To thoroughly study this question, we train and evaluate our models with searched and random references, respectively. Table III (c) compares the results, where 'Search' indicates the model using the picked similar neighbor as reference, while 'Random' means the reference image is randomly selected. If the models are trained with similar references, the performance of random reference during inference is worse than those of picked references. For examples, the FID of $\text{ReGO}_{\text{BDIE}}$ is dropped from 10.052 to 10.289. In contrast, the models trained with randomly references harvest incremental performance gains comparing with the original methods. For example, the FID of BDIE is 11.021, while the FID of $\text{ReGO}_{\text{BDIE}}$ with the searched reference during inference is 10.889, while if the model is trained with similar samples, its FID can be boosted to 10.052. Given the above, we can conclude the following: 1) Our methods can work well even without a similar reference image being provided; and 2) using similar reference images for both training and testing can lead to better performance, which validates our motivation.

TABLE IV
THE PERFORMANCE COMPARISON OF $\text{ReGO}_{\text{BDIE}}$ WITH TOP 1-5 SIMILAR REFERENCES

Method	IS↑	FID↓	MSD↑
$\text{ReGO}_{\text{BDIE}}$	3.126	10.052	1.357
$\text{ReGO}_{\text{BDIE}}$-TOP2	3.137	10.064	1.354
$\text{ReGO}_{\text{BDIE}}$-TOP3	3.129	10.059	1.351
$\text{ReGO}_{\text{BDIE}}$-TOP4	3.132	10.061	1.353
$\text{ReGO}_{\text{BDIE}}$-TOP5	3.134	10.056	1.351

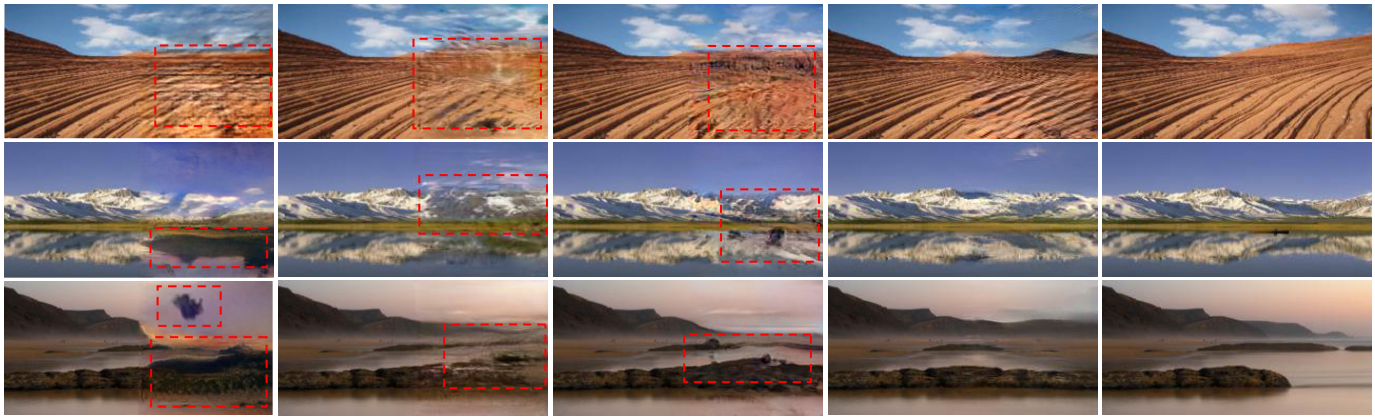
Table IV provides additional insights into the robustness of the reference image during inference. In our default setting, we utilize the most similar image identified through search as the reference. To investigate the impact of various reference images, this section examines the performance of the Top 1-5 similar images. The performance comparison based on $\text{ReGO}_{\text{BDIE}}$ on the NS6K dataset is presented in Table IV, the results reveal that using the Top 1-5 similar images yields comparable performance. Additionally, using the most similar reference image achieves the best results.

Discuss the Number of Reference Image. In our baseline methods, five neighbors for each training sample are



(a) LaMa [35] (b) SGIO [10] (c) BDIE [2] (d) ReGO$_{\text{BDIE}}$ (e) Groundtruth

I: Image rebuilding according to the original sketches.



(a) LaMa [35] (b) SGIO [10] (c) BDIE [2] (d) ReGO$_{\text{BDIE}}$ (e) Groundtruth

II: Random outpainting

Fig. 6. The results for the image rebuilding and the random outpainting, where the dotted red line indicates the imperfect region. Part I shows the results on image rebuilding according to the original sketches. The comparison methods could predict reasonable pixels for the input, however, they all suffer from the blurry synthesized content. While our methods could synthesize texture-rich content. The part II exhibits the random outpainting. Even though no sketch is provided for guidance, ReGO$_{\text{BDIE}}$ could also synthesise texture-rich results and performs much better than the methods originally designed for random outpainting, *i.e.* , NSIO [4] and BDIE [2].

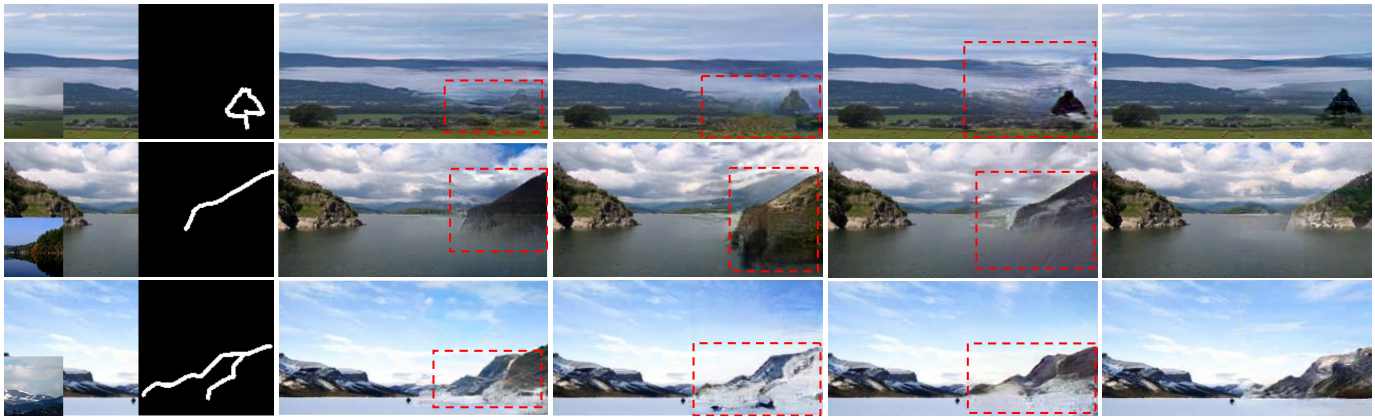
selected, and we randomly pick up one in each iteration to serve as the reference image. This subsection investigates the impacts of the number of the reference image. The performance tendencies with five different reference numbers are shown in Table. III (d), where the FID is scaled by the logarithmic function. Two important observations can be made from Table. III (d). First, Comparing to employing only one reference image, using multiple references could enhance the model generality and train a more robust generation model. The FID of ReGO$_{\text{BDIE}}$ with only one reference image is 10.728, when the reference number increases to 5, the FID could be improved to 10.052. Secondly, it is observed that increasing the number of reference images beyond five does not further improve the performance, which is evident from the performance tendencies when the number of references ranges from 5 to 20. The situation with five reference images achieves the best performance on average.

E. Qualitative Results

Image Rebuilding. Fig. 6 provides the visualizations of the rebuilding results according to the original sketches and random outpainting for SOTA inpainting and outpainting models. To ease the visual exhibition as well as saving some space, we only exhibit the outpainting results of our best model ReGO$_{\text{BDIE}}$. It can be observed that the results of ReGO$_{\text{BDIE}}$ are more authentic and natural due to the richer textural details.

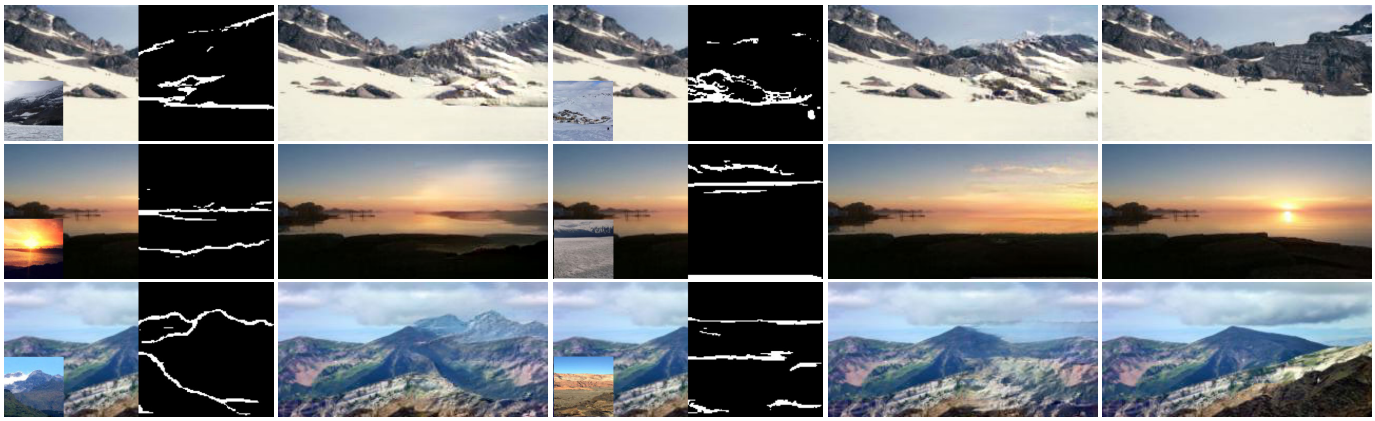
From Fig. 6 I, the comparison methods, LaMa [33], SGIO [10] and BDIE [2], could extend reasonable pixels for the input image, but the predicted content is blurry and lacks textural details, which makes the overall image not authentic enough. While ReGO$_{\text{BDIE}}$ could produce texture-rich outpainting results. The results of random prediction are exhibited in Fig. 6 II, comparing to the competing methods, ReGO$_{\text{BDIE}}$ could also successfully synthesize the results with more textural details when no sketches are fed, and the synthetic images are even more satisfactory than the method original designed for the random prediction, *i.e.* , BDIE [2]. From Fig. 6, we could find that the guiding sketch is not one of requisite inputs for our system, when the users do not provide the guiding sketch, ours system could also produce satisfactory random outpainitng results.

Free-form Outpainting. The comparison for free-form outpainting is exhibited in Fig. 7 I, ReGO$_{\text{BDIE}}$ could not only synthesize the expected content matching the guiding sketch but achieve authentic and natural enough results. Especially



(a) Inputs (b) LaMa [35] (c) SGIO [10] (d) BDIE [2] (e) ReGO$_{\text{BDIE}}$

I: Outpainting using manually drawn free-form sketches.



(a) Inputs (b) Outputs (c) Inputs (d) Outputs (e) Original images

II: Outpainting using sketches from the other images as guidance.

Fig. 7. Outpainting results according to the manually drawn free-form sketches and the sketch from other images, where the dotted red line indicates the imperfect region. Part (I) shows the results for free-form outpainting. While the part (II) exhibits the outpainting using sketches from the other images as guidance. The inputs in (a) and (c) use the same reference images (lower left in (a)) but different sketches. The sketches in (a) are directly from the reference images, while the sketches in (c) are extracted from a randomly selected image (shown in lower left). The corresponding predictions are shown in (b) and (d), respectively. The results are produced by ReGO$_{\text{BDIE}}$.

the boundaries of different semantic regions are much clearer than the competing methods. Additionally, we surprisingly find that the reference image could also help fill reasonable pixels for the free-form outpainting, as shown in the bottom row in Fig. 7 I. Besides the manually drawn sketches, we could also use the sketch from another image to guide the outpainting, as shown in Fig. 7 II. The inputs in Fig. 7 II(a) directly use the sketches of the reference images to control the outpainting, while the ones in Fig. 7 II(c) use the sketches from randomly selected images, two types serve as the simple and the difficult cases, respectively. From Fig. 7 II, our method could not only predict new content matching the guiding sketches but achieve satisfactory style-consistency for both simple and difficult cases. It's worth noting that this paper mainly focuses on synthesizing new content along left to right, however, the prediction of other directions could also be performed based on BDIE backbone, just as shown in Fig. 9, we leave everything unchanged except for using the mask to indicate the missing regions. In this task, our ReGO$_{\text{BDIE}}$ could also achieve more outstanding performance comparing to the BDIE model, 10.817 (FID) 3.004 (IS)-ReGO$_{\text{BDIE}}$ VS 12.132 (FID), 2.893 (IS)-BDIE.

Visualization of Feature Maps. To make it clear that where the beneficial contents come from, we visualize the feature map surrounding the image-guided convolution. The results are shown in Figure 8, where the first row represents the image for outpainting, and the second to fourth rows depict the features before the IGConv, the features extracted from the reference image, and the feature maps after the IGConv, respectively. As shown in the second row, the features corresponding to the predicted part are sparse. In the third row, some contents are extracted from the reference image, which are used to compensate for the features in the second row, resulting in denser and more activated features.

Results on High-Resolution Images. Besides the low resolution dataset, we also collect 558 high-resolution scenery images from Internet using the key word "scenery images" to further evaluate our model. We resize the images as 512×768 and directly evaluate the performance on this dataset, the performance of our method could also outperform the most competitive method BDIE, 34.012(FID) 4.078(IS)-ReGO$_{\text{BDIE}}$ VS 37.841(FID) 3.917(IS)-BDIE. Fig. 10 shows the high resolution results of three state-of-the-art methods. From the above, the proposed framework allows users to harvest three

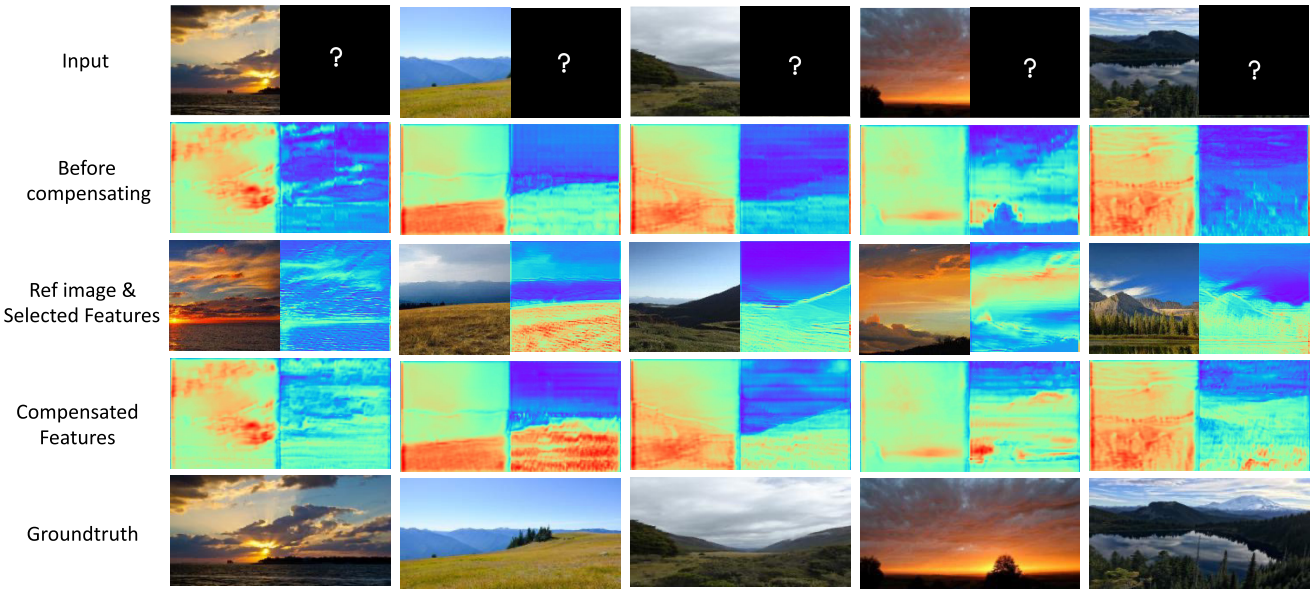
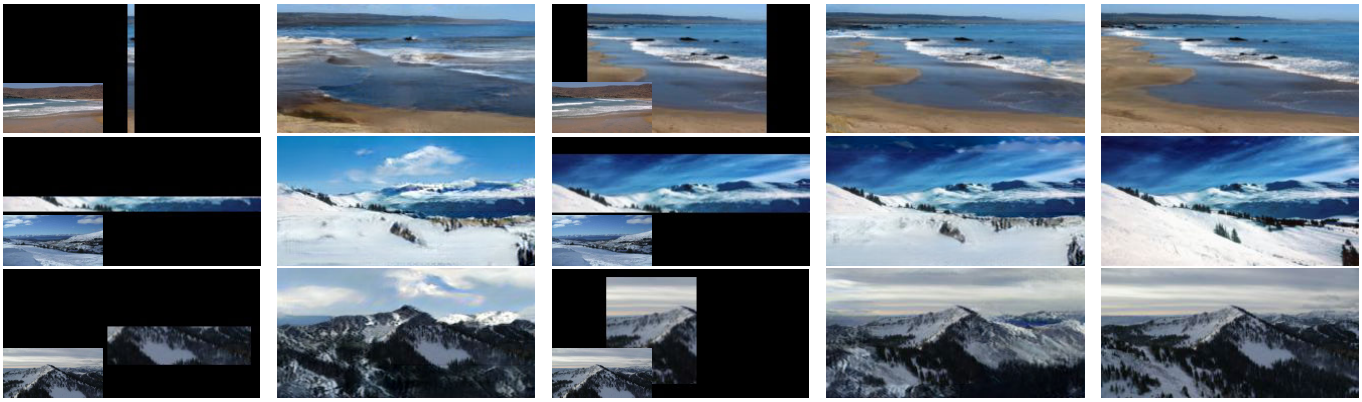
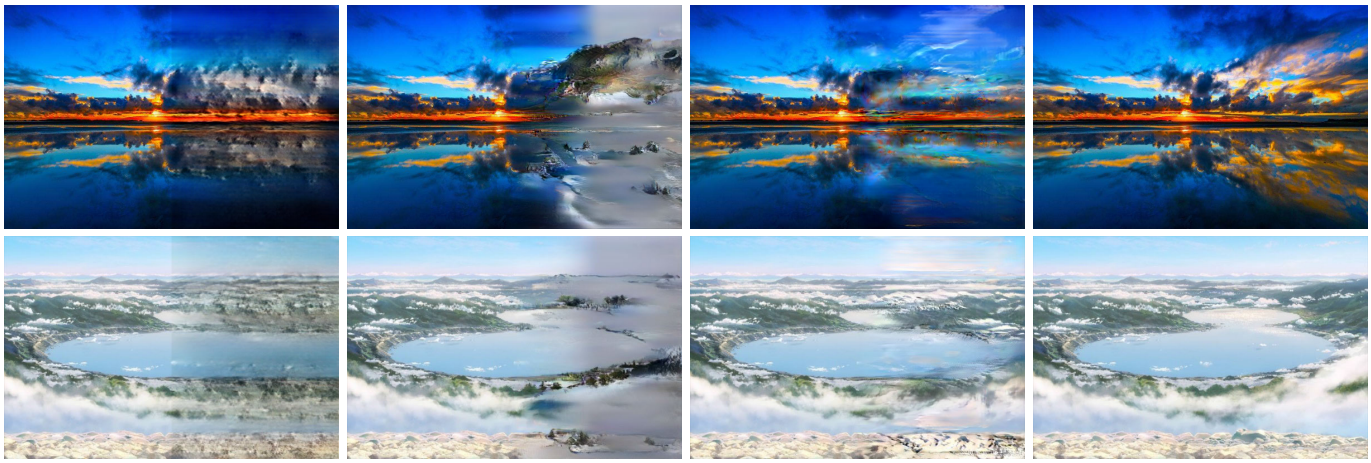


Fig. 8. Visualization of feature maps around the last IGConv of $\text{ReGO}_{\text{BDIE}}$. The images for outpainting, maps before IGConv, maps picked from the reference image, and the groundturth image are subsequently shown from top to bottom rows.



(a) Input (b) Output (c) Input (d) Output (e) Groundtruth

Fig. 9. The results for multi-direction prediction. Based on the BDIE backbone, our method $\text{ReGO}_{\text{BDIE}}$ could predict the content for multiple directions.



(a) LaMa [35] (b) BDIE [2] (c) $\text{ReGO}_{\text{BDIE}}$ (d) GroundTruth

Fig. 10. The high resolution results of all methods, we input the images with 512×384 to rebuild 512×768 images.

types of results: random outpainting, free-form outpainting from manually drawn sketches and controllable outpainting using sketch from another image. Therefore, our proposed method is with higher practical value.

Weaknesses and Limitations. Our ReGO promotes the outpainting perforamnce at the cost of following aspects. **(1) More complex pipeline.** The reference is helpful to enrich the texture but also introduce an additional step to search.

(2) More parameters and lower efficiency. To select the beneficial contents from the reference, we introduce more parameters to process the reference, which will inevitably complex the model and increase the inference time. Besides, this paper mainly focuses on the scenery images and didn't investigate the performance under more complex scenes like indoor, street-view, portrait, etc, which will be studied in our future works.

V. Conclusion and Future Work

In summary, this work introduces a novel ReGO module that enhances outpainting quality by incorporating neighboring pixels. The proposed method effectively improves the results of sketch-guided image outpainting by enriching textural details. An ACS module is developed to filter out non-beneficial pixels and emphasize useful ones. This helps the generator use helpful pixels to enhance its output. A style ranking loss is used to prevent the synthesized content from being affected by the reference image's style. Experiments conducted on three benchmarks using three backbones demonstrate the effectiveness of the proposed method. The idea of enhancing details from neighbors may also be applicable to other generation tasks. In the future, we plan to explore the effectiveness of the proposed method under more complex scenes.

References

[1] Y. Wang, X. Tao, X. Shen, and J. Jia, "Wide-context semantic image extrapolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1399–1408.

[2] D. Krishnan et al., "Boundless: Generative adversarial networks for image extension," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10520–10529.

[3] Y.-C. Cheng, C. H. Lin, H.-Y. Lee, J. Ren, S. Tulyakov, and M.-H. Yang, "InOut: Diverse image outpainting via GAN inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11421–11430.

[4] Z. Yang, J. Dong, P. Liu, Y. Yang, and S. Yan, "Very long natural scenery image prediction by outpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10560–10569.

[5] J. Kopf, W. Kienzle, S. Drucker, and S. B. Kang, "Quality prediction for image completion," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–8, Nov. 2012.

[6] J. Sivic, B. Kaneva, A. Torralba, S. Avidan, and W. T. Freeman, "Creating and exploring a large photorealistic virtual space," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8.

[7] Y. Zhang, J. Xiao, J. Hays, and P. Tan, "FrameBreak: Dramatic image extrapolation by guided shift-maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1171–1178.

[8] M. Wang, Y.-K. Lai, Y. Liang, R. R. Martin, and S.-M. Hu, "BiggerPicture: Data-driven image extrapolation using graph matching," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–13, Nov. 2014.

[9] Y. Li, L. Jiang, and M.-H. Yang, "Controllable and progressive image extrapolation," 2019, *arXiv:1912.11711*.

[10] Y. Wang, Y. Wei, X. Qian, L. Zhu, and Y. Yang, "Sketch-guided scenery image outpainting," *IEEE Trans. Image Process.*, vol. 30, pp. 2643–2655, 2021.

[11] Y. Zhou, C. Barnes, E. Shechtman, and S. Amirghodsi, "TransFill: Reference-guided image inpainting by merging multiple color and spatial transformations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2266–2267.

[12] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 341–346.

[13] Q. Shan, B. Curless, Y. Furukawa, C. Hernández, and S. M. Seitz, "Photo uncrop," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 16–31.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[15] Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[16] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5767–5777.

[19] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[22] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6997–7005.

[23] Y. Yao, J. Ren, X. Xie, W. Liu, Y.-J. Liu, and J. Wang, "Attention-aware multi-stroke style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1467–1475.

[24] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, nos. 1–2, pp. 59–81, May 2014.

[25] W. Quan, R. Zhang, Y. Zhang, Z. Li, J. Wang, and D.-M. Yan, "Image inpainting with local and global refinement," *IEEE Trans. Image Process.*, vol. 31, pp. 2405–2420, 2022.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[27] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4672–4681.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6626–6637.

[29] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2226–2234.

[30] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, "Progressive image inpainting with full-resolution residual network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2496–2504.

[31] C. Xie et al., "Image inpainting with learnable bidirectional attention maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8857–8866.

[32] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4470–4479.

[33] R. Suvorov et al., "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3172–3182.

[34] S. Zhao et al., "Large scale image completion via co-modulated generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[35] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 89–105.

[36] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[37] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.

[38] N. Wang, Y. Zhang, and L. Zhang, "Dynamic selection network for image inpainting," *IEEE Trans. Image Process.*, vol. 30, pp. 1784–1798, 2021.

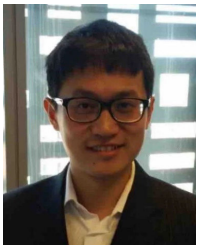
[39] K. Yao, P. Gao, X. Yang, J. Sun, R. Zhang, and K. Huang, "Outpainting by queries," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 153–169.
[40] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.
[41] M. Zhu et al., "Image inpainting by end-to-end cascaded refinement with mask awareness," *IEEE Trans. Image Process.*, vol. 30, pp. 4855–4866, 2021.
[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020.
[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
[44] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023, *arXiv:2302.05543*.
[45] J. Liang et al., "NUWA-infinity: Autoregressive over autoregressive generation for infinite visual synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2022.
[46] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "MISF: Multi-level interactive Siamese filtering for high-fidelity image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1859–1868.
[47] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11451–11461.
[48] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12127–12136.
[49] D. Ding, S. Ram, and J. J. Rodríguez, "Image inpainting using nonlocal texture matching and nonlinear filtering," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1705–1719, Apr. 2019.
[50] W. Chen and J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9416–9425.

Yaxiong Wang received the B.S. degree from Lanzhou University in 2015 and the Ph.D. degree from Xi'an Jiaotong University in 2021. He is currently an Associate Professor with the Hefei University of Technology, Hefei, China. His research interests include image generation, image-text retrieval, and image segmentation.

Yunchao Wei received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2016. He was a Postdoctoral Researcher with the Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, IL, USA, from 2017 to 2019. He is currently a Full Professor with Beijing Jiaotong University. Before that, he was a Senior Lecturer with the University of Technology Sydney, Sydney, NSW, Australia. His current research interests include computer vision and machine learning. He is the ARC Discovery Early Career Researcher Award Fellow from 2019 to 2021.

Xueming Qian (Member, IEEE) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, in 2008. He was a Visiting Scholar with Microsoft Research Asia from 2010 to 2011. He was an Assistant Professor with Xi'an Jiaotong University, where he was an Associate Professor, from 2011 to 2014. He is currently a Full Professor. He is also the Director of the Smiles Laboratory, Xi'an Jiaotong University. His research interests include social media big data mining and search. His research is supported by the National Natural Science Foundation of China, Microsoft Research, and Ministry of Science and Technology. He received the Microsoft Fellowship in 2006. He received the Outstanding Doctoral Dissertations of Xi'an Jiaotong University and Shaanxi Province in 2010 and 2011, respectively.

Li Zhu received the B.S. degree from Northwestern Polytechnical University in 1989 and the M.S. and Ph.D. degrees from Xi'an Jiaotong University in 1995 and 2000, respectively. He is currently a Professor with the School of Software, Xi'an Jiaotong University. His main research interests include multimedia processing and communication, parallel computing, and networking.



Yi Yang (Senior Member, IEEE) received the Ph.D. degree in computer science from Zhejiang University in 2010. He was a Postdoctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with Zhejiang University, Hangzhou, China. Before that, he was a Full Professor with the University of Technology Sydney, Sydney, NSW, Australia. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimodal generation, video analysis, and AI for science.