InOut : Diverse Image Outpainting via GAN Inversion

Yen-Chi Cheng[1], Chieh Hubert Lin[2], Hsin-Ying Lee[3], Jian Ren[3], Sergey Tulyakov[3], Ming-Hsuan Yang[2,4,5]
[1]Carnegie Mellon University [2]UC Merced [3]Snap Inc. [4]Yonsei University [5]Google Research
https://yccyenchicheng.github.io/InOut/

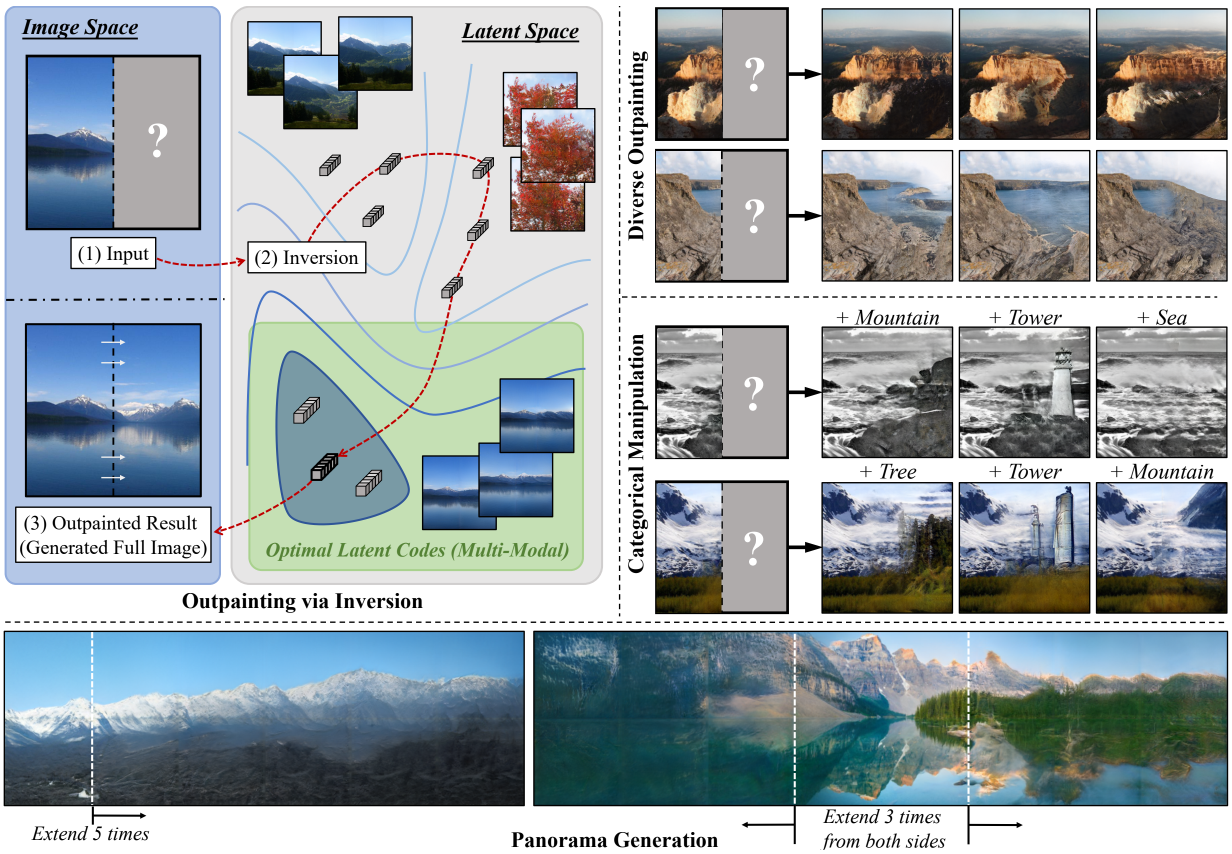

Figure 1. *(top-left)* Given an input image and a trained generator, the proposed algorithm searches for latent codes that can generate images containing the input image. We can naturally achieve *(top-right)* diverse image outpainting, *(middle-right)* categorical manipulation for outpainting area, and *(bottom)* generate panorama with rich and complex structure.

Abstract

Image outpainting seeks for a semantically consistent extension of the input image beyond its available content. Compared to inpainting — filling in missing pixels in a way coherent with the neighboring pixels — outpainting can be achieved in more diverse ways since the problem is less constrained by the surrounding pixels. Existing image outpainting methods pose the problem as a conditional image-to-image translation task, often generating repetitive structures and textures by replicating the content available in the input image. In this work, we formulate the problem from the perspective of inverting generative adversarial networks. Our generator renders micro-patches conditioned on their joint latent code as well as their individual positions in the image. To outpaint an image, we seek for multiple latent codes not only recovering available patches but also synthesizing diverse outpainting by patch-based generation. This leads to richer structure and content in the outpainted regions. Furthermore, our formulation allows for outpainting conditioned on the categorical input, thereby enabling flexible user controls. Extensive experimental results demonstrate the proposed method performs favorably against existing in- and outpainting methods, featuring higher visual quality and diversity.

1. Introduction

Given an input image, we can easily picture how adjacent images might look. For example, given an image of mountains, we can picture the surroundings covered by forests or snow, imagine a lake beneath the hillside, and visualize cliffs near the ocean. This mental skill depends on our prior experience and exposure to diverse scenery. In other words, this is an *image outpainting* task. It enableS various content creation applications such as image editing using extrapolated regions, panorama image generation, and extended immersive experience in virtual reality, to name a few.

Recent advances in image inpainting [20, 23, 33, 34] do not directly address the outpainting problem as the former has more context to deal with — the missing pixels have a larger amount of available surrounding pixels, serving as the boundary conditions and providing crucial guidance for inpainting. In contrast, the outpainting problem can rely only on the context of the available image, with only a scarce number of pixels near the boundary available as the boundary condition. A similar analogy is between video interpolation and video prediction, where the former deals with existing events while the latter tries to model multiple futures.

In the literature, image outpainting is addressed from the image-to-image translation (I2I) perspective [27,31]. These methods aim to learn a deterministic mapping from the domain of partial images to the domain of complete outpainted images. This formulation is limited in several respects. First, the available pixels serve as a strong source of context, thereby facilitating leakage of textures and structures of the input to the output and leading to the repetitive nature of the outpainting (as shown in panorama results in [27]). Second, existing I2I-based methods are deterministic, while in reality there exist numerous ways each image can be outpainted. Applying the available multimodal I2I methods [13, 17] to the outpainting problem is non-trivial.

In this work, we tackle the outpainting problem by inverting generative adversarial networks (GANs) [1,4,7,37]. We first extend a StyleGAN2-based [15] generator to perform generation in a coordinate conditional manner and independently generate spatially consistent micro-patches. Each micro-patch shares the global latent code with the rest of micro-patches in the image, while having a unique coordinate label. Outpainting can then be formulated as finding the optimal latent codes for the available input micro-patches, followed by generating the desired regions by providing the proper coordinate conditioning. To search for the latent code, we propose a GAN inversion process that finds multiple latent codes producing diverse outpainted regions, unlocking diversity in the output. In addition, we propose a categorical generation schema to enable flexible user control. Figure 1 shows examples of multi-modal and categorical outpainting.

We qualitatively and quantitatively evaluate the proposed method on the Place365 [36] dataset, and the Flickr-Scenery dataset which we collected. We use Fréchet Inception Distance (FID) [12] and conduct a user study to evaluate the realism of outpainted images. Since the proposed method can achieve multi-modal generation, we measure the diversity using the Learned Perceptual Image Patch Similarity (LPIPS) metric [35]. Finally, we demonstrate the scenario of categorical generation in the outpainting area and the panorama generation.

2. Related Work

GAN Inversion. Generative models aim to model and sample from a target distribution. Generative adversarial networks (GANs) [11], among various generative models, have demonstrated superior performance in generating high-quality samples. To explore the interpretability of well-trained GANs, GAN inversion has been proposed to find the latent codes that can accurately recover given images for a trained GAN model. There are two main branches of approaches. *Encoder-based methods* [5, 9, 24] adopt an additional encoder to learn the mapping from the image domain to the latent space. *Optimization-based methods* [1,2,7,19,21] use gradient-based optimization methods with reconstruction loss as the objective function to find latent codes that can recover input images. Other variants use encoders to get an initialization for the optimization process [4, 38], or modify the training framework by incorporating invertibility [8, 37].

Image Inpainting. From the aspect of filling missing pixels in images with generative models, the inpainting problem is conceptually related to the outpainting task. Existing image inpainting methods can be categorized into two groups. The first line of work leverages patch similarity and diffusion to obtain essential information from known regions [10, 31]. These methods usually work well on textures but fail to learn semantic structures. The other line of work adopts a learning-based approach to gain better semantic understanding [20, 23, 33, 34]. Most methods apply an encoder-decoder model with the reconstruction loss and adversarial loss to ensure the filled content is smooth and realistic. Image outpainting is more challenging since it entails *creating* new contents rather than *filling in* partial regions, requiring a substantial understanding of scenes.

Image Outpainting. Most image outpainting methods [10, 16, 26, 28] apply patch-based retrieval and matching algorithms to predict possible extrapolation. Recently, several approaches [27, 29, 31] apply GAN models and formulate the problem as an image-to-image translation task. However, the conditional formulation relies heavily on the given available pixels and tends to create repetitive textures and

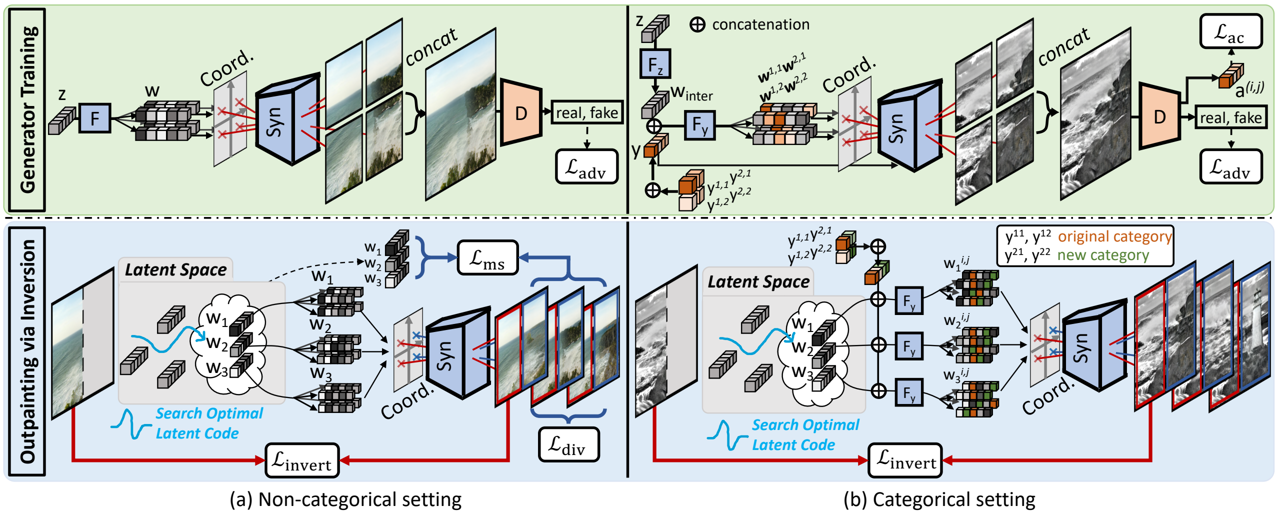

Figure 2. **Overview. Generator training:** The proposed generator adopts StyleGAN2 as the backbone and incorporates coordinate conditioning, where micro-patches are generated conditioned on their positions in the image. **Outpainting via inversion:** We search for latent codes that can recover the given partial images and synthesize diverse samples in the outpainted regions. In addition to the *unconditional setting*, we introduce a *categorical setting* variant, which enables flexible user controls for the categorical manipulation on each outpainting micro-patch.

structures. To the best of our knowledge, the proposed method is the first attempt to tackle the image outpainting task from the GAN inversion perspective.

3. Diverse Outpainting via Inversion

Overview. The goal of image outpainting is to outward-synthesize unknown regions with respect to the given input image. Our pipeline consists of two stages, ***generator training*** and ***outpainting via inversion***. In Section 3.1, we first introduce a generator based on the StyleGAN [14, 15] and COCO-GAN [18]. It is trained to output micro-patches conditioned on the joint latent and on the coordinate of the patch in the output image. We do not specifically optimize the generator to perform outpainting. During the outpainting stage (Section 3.2) we find the optimal latent code for the available input patches in the latent space of the trained patch-based generator. Outpainting is then performed by combining the desired coordinates and the found optimal latent code and generating a new patch. We further propose a categorical-conditioning scheme to enable controllable outpainting. Finally, a simple blending algorithm to further mitigate the artifacts is introduced in Section 3.3.

3.1. Coordinate-conditioned Generator

In this work, we handle two different settings: (a) *non-categorical generation* that synthesizes images from latent codes, and (b) *categorical generation* that uses categorical labels as additional conditional context, enabling more user-control in the following inversion stage.

Non-categorical generation. We use the StyleGAN2 [15] as our backbone architecture. Given a latent $\mathbf{z}$ from the input latent space $\mathcal{Z}$, we obtain an intermediate code $\mathbf{w} \in \mathcal{W}$ by a non-linear mapping network F. Similar to in [30], we map $\mathbf{w}$ to a Gaussianized space $\mathcal{V}$. The mapping is achieved via a Leaky ReLU (LRU) with a negative slope of 5, that is, $\mathbf{v} = \mathrm{LRU}_{5.0}(\mathbf{w})$. The outpainting quality in the later GAN inversion stage can be substantially improved with the additional Gaussianized space. The necessity of adopting the Gaussianized space is discussed in Section 3.2.

We formulate the image outpainting problem as finding the latent codes that synthesizes images overlapping with the input image. In the inversion process, we seek for a latent code for the whole image while having only a part of the image available. Therefore, instead of generating a full image, the generator G synthesizes several *micro-patches* $\{I^{i,j}_{\text{micro}}\}_{i,j=1,\dots,n}$, which will be concatenated to form a full image I_f. Each patch depends on the joint latent code and its coordinates. For an $n \times n$ micro-patches generation setting, the corresponding coordinates to $\{I^{i,j}_{\text{micro}}\}_{i,j=1,\dots,n}$ are $\{c^{i,j}\}_{i,j=1,\dots,n}$. We set $c^{1,1} = (-1,-1)$, $c^{n,n} = (1,1)$, and the rest, if any, are obtained by linear interpolation. The output image I_f is generated as follows:

$$\begin{aligned} \mathbf{w} &= F(\mathbf{z})\,, & \mathbf{v} &= \mathrm{LRU}_{5.0}(\mathbf{w}) \\ I^{i,j}_{\text{micro}} &= G(\mathbf{v}, c^{i,j})\,, & I_{\text{f}} &= \underset{i,j=1,\dots,n}{\text{concat}}\,(I^{i,j}_{\text{micro}})\,. \end{aligned} \tag{1}$$

We train the generator using the Wasserstein-GAN loss [3] with real full-images I_{r} and generated full-images I_{f}:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{I_{\text{r}}}\left[D(I_{\text{r}})\right] - \mathbb{E}_{z}\left[D(I_{\text{f}})\right]. \tag{2}$$

Categorical generation. To enable fine-grained user control in the inversion stage, we propose a categorical generation schema. Given a real image I_r, we divide it into micro-patches $\{I^{i,j}_{\text{micro}}\}$. We then obtain the categorical labels $\{y^{i,j}\}$ for each micro-patch using the off-the-shelf DeepLabV3 [6] model, and set the k-th element in the multi-class binary label vector $y^{i,j}$ to 1 if any of the pixels within $I^{i,j}_{\text{micro}}$ is recognized as the k-th class.

To use categorical information as a conditional input, we split the nonlinear mapping network F into $\{F_z, F_y\}$. Here, F_z operates the same way as F in the non-categorical setting, while F_y takes $\{y^{i,j}\}$ as an additional input and fuses the information with the output of F_z. The new $w_{i,j}$ under the categorical setting is computed by

$$\mathbf{w}_{\text{inter}} = F_z(\mathbf{z})\,, \quad \mathbf{w}^{i,j} = F_y(\mathbf{w}_{\text{inter}}, y^{i,j})\,. \tag{3}$$

Next, similar to the non-categorical setting, we Gaussianize the code with $\mathbf{v}^{i,j} = \text{LRU}_{5.0}(\mathbf{w}^{i,j})$, generate micro-patches $I^{i,j}_{\text{micro}} = G(\mathbf{v}, c_{i,j})$, then concatenate $\{I^{i,j}_{\text{micro}}\}$ into a full image I_f. We apply an auxiliary classifier that uses the last intermediate features of D to perform multi-class classification $a^{i,j}$ for all $I^{i,j}_{\text{micro}}$, aiming at learning a proper conditional distribution of $I^{i,j}_{\text{micro}}$ regarding the $y^{i,j}$ input to G.

$$\mathcal{L}_{\text{cls}} = \text{BCE}(a^{i,j}, y^{i,j})\,, \tag{4}$$

where BCE is the binary cross entropy loss function. The full training objective is:

$$\min_D \max_G \mathcal{L}_{\text{adv}} + \min_{G,D} \mathcal{L}_{\text{cls}}\,.$$

3.2. GAN Inversion with Diversity Loss

Given a trained coordinate-conditioned generator G as discussed in the previous section and an input image R as a reference, we generate a set of possible outpainted images by composing R with generated micro-patches $\{O^m\}$. For brevity and notation clarity, here we assume G is trained with a grid of 2×2 micro-patches, $\{R^{1,1}_{\text{micro}}, R^{1,2}_{\text{micro}}, R^{2,1}_{\text{micro}}, R^{2,2}_{\text{micro}}\}$. Furthermore, for presentation simplicity, we assume R to be on the left and consists of two left-side micro-patches (i.e., $R^{1,1}_{\text{micro}}$ and $R^{1,2}_{\text{micro}}$), while the outpainted area $\{O^m\}$ on the right, as shown in the lower-half of Figure 2. Note that, in practice, G is not restricted to 2×2, R can be of any resolution, and outpainting can be performed using an arbitrary direction.

Similar to existing inversion methods, we seek for the optimal latent code w that recovers the input image. With $R_f = \text{concat}(G(\mathbf{v}, c^{1,1}), G(\mathbf{v}, c^{1,2}))$, we get:

$$\mathcal{L}_{\text{mse}} = \|R - R_f\|_2\,, \quad \mathcal{L}_{\text{percept}} = \text{Percept}(R, R_f)\,, \tag{5}$$

where $\mathbf{v} = \text{LRU}_{5.0}(\mathbf{w})$ and Percept is the perceptual distance proposed in [35].

The outpainting process requires not only the reconstructed parts to be correct R (i.e., $I^{1,1}_{\text{micro}}$ and $I^{1,2}_{\text{micro}}$), but the outpainted parts (i.e., $I^{2,1}_{\text{micro}}$ and $I^{2,2}_{\text{micro}}$) to be realistic and consistent. Note that the continuity and consistency between micro-patches are enforced by joint latent and the coordinate conditioning schema. During the generator training, the latent is sampled from a Gaussian distribution. Therefore, it is crucial to encourage the sought latent code $\mathbf{w}$ to belong to the domain of the training data, and be interpretable by G, instead of overfitting to the given image with the out-of-domain latent code. As the first step, in Section 3.1, we add an additional *Gaussianized* space [30] $\mathcal{V}$ after $\mathcal{W}$, simplifying the complex and arbitrarily shaped $\mathcal{W}$ with $\text{LRU}_{5.0}$. Next, with the Gaussianized $\mathcal{V}$, we can easily derive the mean μ and covariance matrix $\mathbf{\Sigma}$ of the distribution $p(\mathbf{v})$, where $\mathbf{v} \in \mathcal{V}$. We encourage recovered $\mathbf{v}$ to be in the training distribution by regularizing its prior:

$$\mathcal{L}_{\text{prior}} = (\mathbf{v} - \mu)^\top \mathbf{\Sigma}^{-1} (\mathbf{v} - \mu)\,. \tag{6}$$

To enable diverse outpainting, we apply two different objective functions. Assuming we target at generating m different outpainted results, we first explicitly penalize the inverted latent codes with their pairwise distance:

$$\mathcal{L}_{\text{div}} = -\sum_{i=1}^{m} \sum_{j=i+1}^{m} \|\mathbf{w}^i - \mathbf{w}^j\|_1\,. \tag{7}$$

Then, to further encourage the model to seek for different final latent codes within the latent space, we apply a mode-seeking regularization [22]:

$$\mathcal{L}_{\text{ms}} = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \left(\frac{\|G(\mathbf{w}^i) - G(\mathbf{w}^j)\|_1}{\|\mathbf{w}^i - \mathbf{w}^j\|_1}\right)\,. \tag{8}$$

The full objective of our inversion is:

$$\begin{aligned} \mathop{\arg\min}_{\{\mathbf{w}_\mathrm{i}\} \in \mathcal{W}} \quad & \lambda_{\text{mse}} \mathcal{L}_{\text{mse}} + \lambda_{\text{percept}} \mathcal{L}_{\text{percept}} + \\ & \lambda_{\text{prior}} \mathcal{L}_{\text{prior}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}} + \lambda_{\text{ms}} \mathcal{L}_{\text{ms}}\,, \end{aligned} \tag{9}$$

where the hyper-parameters λ's control the importance of each term.The proposed pipeline can be applied to either optimization-based, encoder-based, or hybrid methods. Note that such an inversion paradigm is the same for both non-categorical and categorical settings, except the categorical setting seeks for $\mathbf{w}_{\text{inter}}$ instead of $\mathbf{w}$.

3.3. Patch Blending

As described in Section 3.2, the inversion process requires the reconstruction of the given parts and the prediction of the outpainting parts, where the continuity and consistency are enforced in the training stage. However, even with the help of the prior loss $\mathcal{L}_{\text{prior}}$, the outpainting occasionally leads to tiny seams between patches after concatenating patches. Due to simple merging of patches, the outpainted images are likely to contain artifacts. As such, we introduce an image blending method to address this issue. In addition to the reference image R and outpainted area O, we generate the patches located halfway between R and O. Take $R = I^{1,1}_{\text{micro}}, I^{1,2}_{\text{micro}}$ and $O = I^{2,1}_{\text{micro}}, I^{2,2}_{\text{micro}}$ as an example, the additional region A is generated with coordinate

Table 1. **Quantitative comparisons.** The proposed method outperforms related state-of-the-art baselines on both Places365 and Flickr-Scenery datasets in FID and IS metrics, measuring both the visual quality and diversity.

	Place365		Flickr-Scenery	
Method	FID ↓	IS ↑	FID ↓	IS ↑
Boundless	35.02	6.15	61.98	6.98
NS-outpaint	50.68	4.70	61.16	4.76
DeepFillv2	56.14	5.69	62.47	5.38
Image2StyleGAN++	<u>25.36</u>	6.71	40.39	7.10
InOut (Ours)	**23.57**	<u>7.18</u>	**30.34**	**7.16**
InOut-C (Ours)	29.24	**7.69**	<u>33.17</u>	<u>7.15</u>

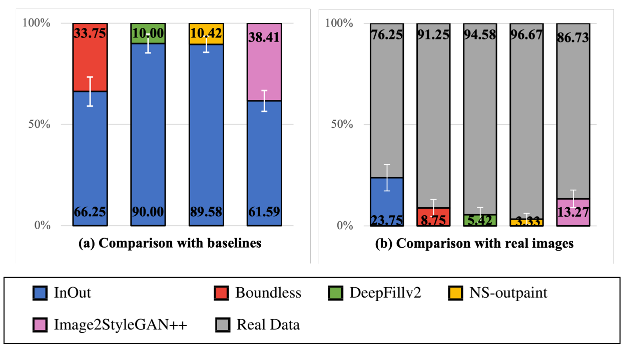

Figure 3. **User Studies.** We conduct user studies to quantify the visual quality in two settings: (a) comparison with baselines, and (b) comparison with real images. We mark the 95% confidence interval with white bars.

Table 2. **Ablation studies**. We show the necessity of each component using FID and LPIPS for quality and diversity.

# output	m=2		m=3	
Method	FID ↓	Diversity ↑	FID ↓	Diversity ↑
InOut w/o $\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{ms}}$	30.12	0.183	30.28	0.176
InOut w/o $\mathcal{L}_{\text{ms}}$	29.85	0.201	29.80	0.206
InOut w/o $\mathcal{L}_{\text{div}}$	29.75	0.204	33.97	0.201
InOut w/o $\mathcal{L}_{\text{prior}}$	36.56	0.216	36.53	0.220
InOut	30.26	0.211	30.18	0.223

$(0, -1)$ and $(0, 1)$. We then linearly blend the overlapped area between R and A and the area between O and A.

Despite its simplicity, this post-processing step provides sufficient quality for our purpose. In practice, we observe that the generator can accurately interpolate the positions of the extended silhouette of landscapes with respect to the coordinate interpolation.

4. Experimental Results

Dataset. We evaluate our method on scenery datasets since they are the most representative and natural use-cases of outpainting. We perform experiments on the Places365 [36] dataset and a collected Flickr-Scenery dataset. More results on images with structured samples (e.g., buildings) could be found in the supplementary material. Similar to [27], we evaluate our method on a subset of the Places365 dataset. We select 25 scenery classes from the Places365 dataset with a subset of 62,500 samples. To further analyze the generalization of our method, we construct a Flickr-Scenery dataset by collecting a large-scale scenery image database of 54,710 images from Flickr. All images are center-cropped and resized to 256×256 pixels. For both datasets, we split the data into 80%, 10%, 10% for training, validation, and testing. All quantitative and qualitative experiments are evaluated on the testing set only. The source code, trained models, and Flickr-Scenery dataset will be made publicly available.

Evaluated Methods. We carry out quantitative and qualitative experiments with the state-of-the-art image outpainting methods (Boundless [27] and NS-outpaint [31]) and also image-inpainting methods (DeepFillv2 [33, 34] and Image2stylegan++ [2]). In order to have a fair comparison to Image2stylegan++, we adopt the optimization-based method for all comparisons. We compare and discuss optimization- and encoder-based methods in the supplementary materials.

4.1. Quantitative Evaluation

We evaluate the results in terms of realism and diversity using the Fréchet Inception Distance (FID) [12] and Inception Score (IS) [25]. Note that we *do not* apply the blending scheme introduced in Section 3.3 across the quantitative evaluations, as we aim to demonstrate the strength of the proposed pipeline without additional postprocessing.

As shown in Table 1, all of the proposed inversion-based methods perform favorably against I2I-based methods. The FID and IS results demonstrate that the generated-image distributions from our InOut variants to the real distribution are significantly more similar than the generation distributions from I2I-based baseline methods. Furthermore, compared to Image2stylegan++ [2], the results show that coordinate conditioning not only enables the categorical manipulation feature of InOut-C, but also naturally improves the generation diversity and quality.

User studies. We conduct user studies to make explicit pairwise qualitative comparisons in two settings: (a) our method against each of the baseline methods, and (b) all methods against real samples. For each round of comparison, we present a pair of two outpainting results generated from the same real sample to the users. The images are either sampled from our method, baselines, or real images. Then, the subjects are asked to select a more realistic and preferred sample out of the image pair. We collect the results from 80 volunteers. Each of them makes 21 rounds of

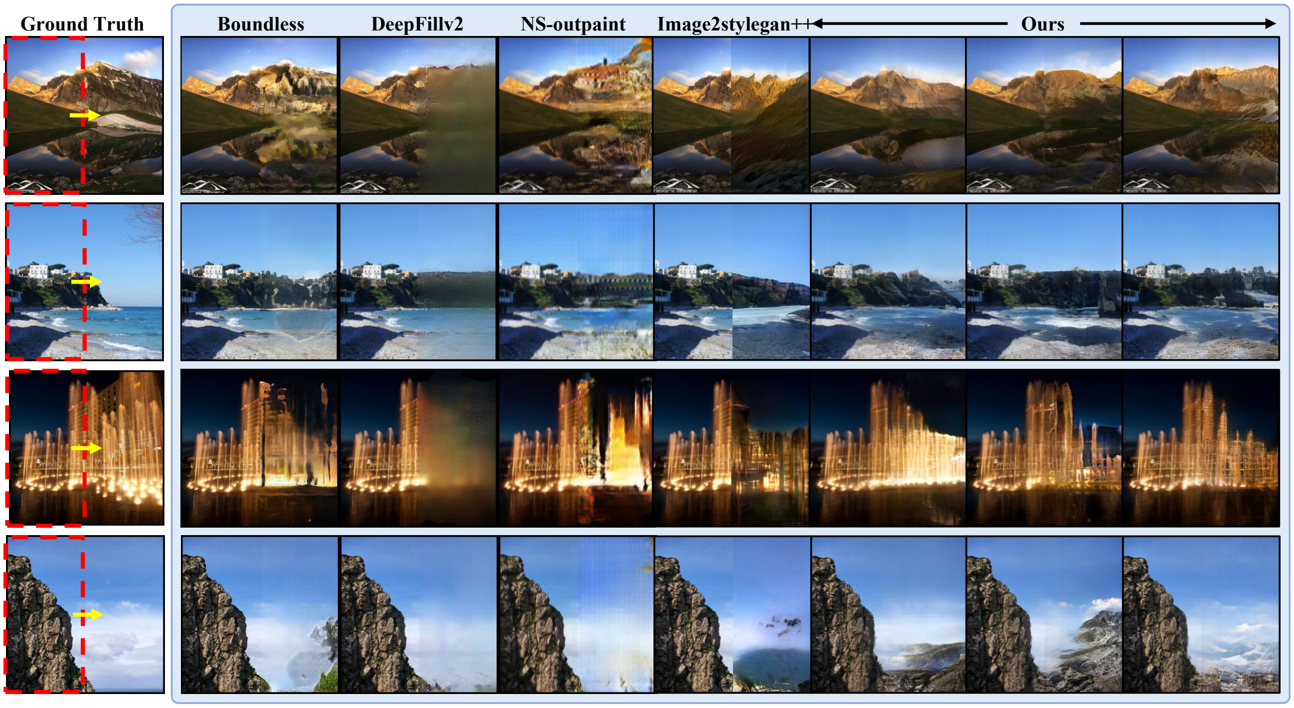

Figure 4. **Comparison to related work.** The qualitative comparison against other related methods show that the proposed approach is more stable, synthesizes richer context with more complex structures, and is able to handle some of the difficult complex scenes. (The input regions to all methods are marked with red dashes.)

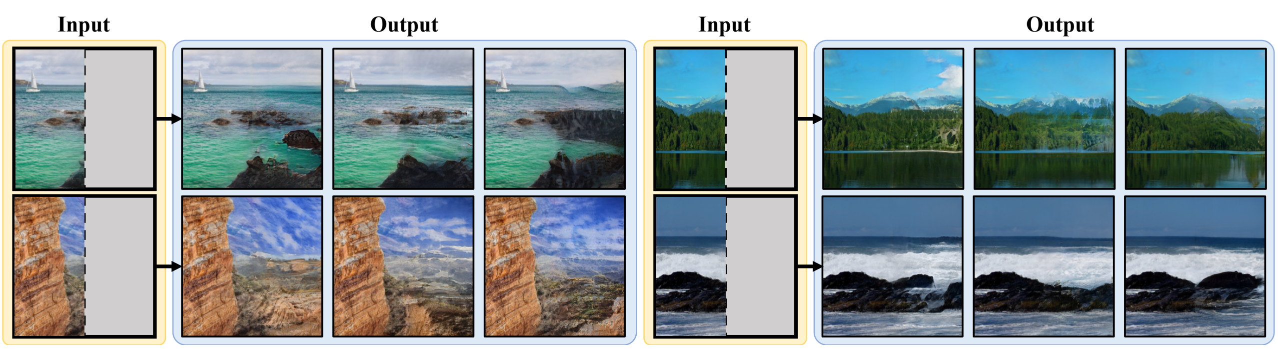

Figure 5. **Diverse outpainting.** We show that the proposed method can seek various solutions for a given input, achieving a high-variety of outpainting results without sacrificing the generation quality.

selections, resulting in 1,680 data points.

Figure 3 shows that subjects prefer the outpainting results by our model than those by other evaluated methods. Especially in comparison with real images, we observe a noticeable gap between the proposed InOut and Boundless. This may be attributed to that our method can frequently synthesize complex structures and novel objects (as shown in Figure 4 and Figure 6) that match the varieties and details of real images, while Boundless tends to create overly-smoothed results with raindrop-shaped artifacts.

Ablation studies. In Table 2, we analyze and quantify the effectiveness of each component with ablation studies. We introduce a diversity score that measures the perceptual distance [35] among m outpainting results with respect to a real image. Each diversity score is averaged over 2,048 samples in the testing set. The diversity loss functions $\mathcal{L}_{\mathrm{ms}}$ and $\mathcal{L}_{\mathrm{div}}$ improve the diversity without compromising the quality. Note that the worst-case diversity quantity without any diversity loss is unlikely to be zero. This is due to the stochastic nature of gradient descent and randomization techniques introduced in [15] for encouraging the exploration during optimization. Furthermore, we show that

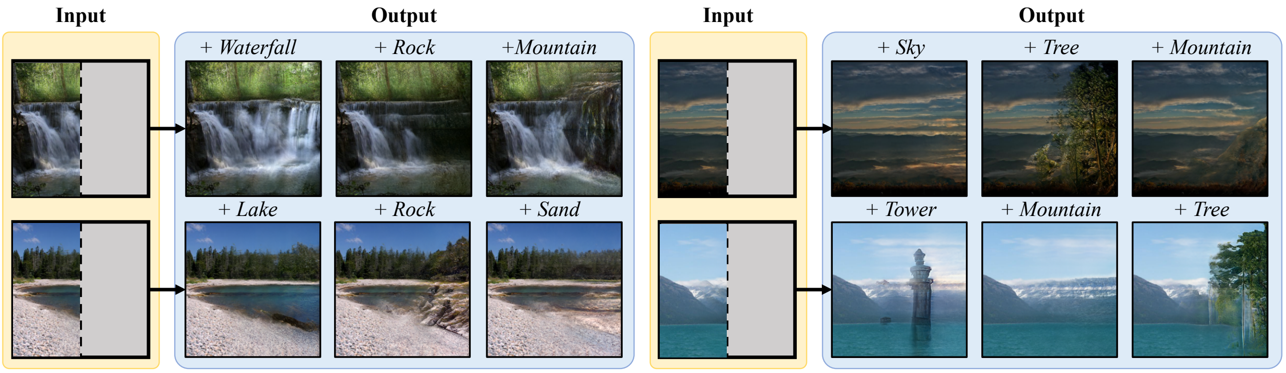

Figure 6. **Categorical generation.** We show the effectiveness of categorical manipulation by assigning different categorical labels to the outpainting area of the same real-image input. The results demonstrate that the proposed method can smoothly impose novel objects and calibrate the landscape to accommodate different categorical controls from users.

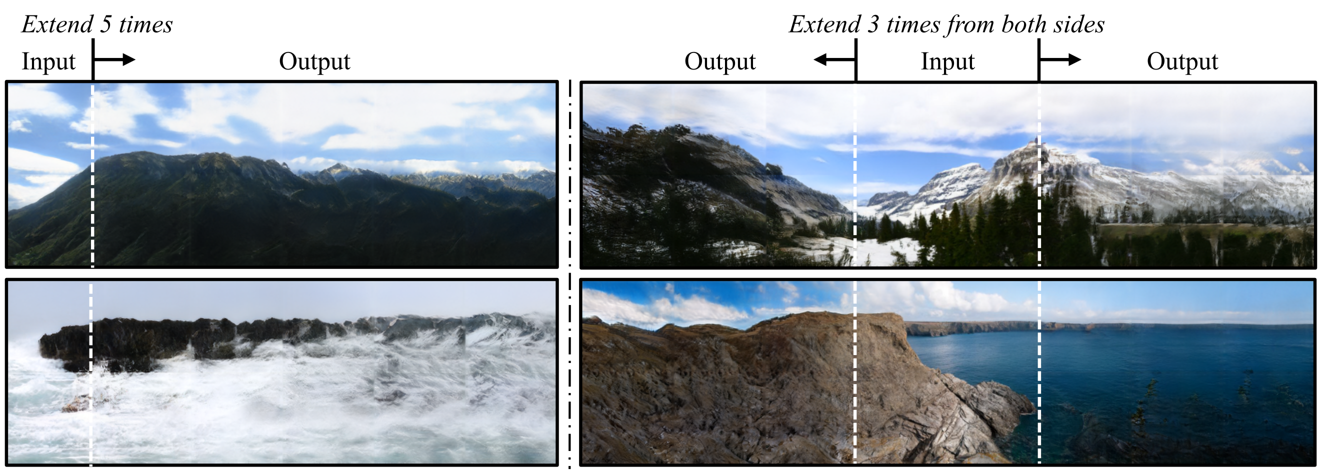

Figure 7. **Panorama generation.** We synthesize panoramic images by performing recursive outpainting. The results are of high quality and high structural complexity without repeating patterns.

the prior loss $\mathcal{L}_{\text{prior}}$ is essential for securing the visual quality. As we have discussed in Section 3.2, the $\mathcal{L}_{\text{prior}}$ regularizes the final state of the inverted latent codes to be located within the dense area of the Gaussian prior. Hence, the generated contents within the outpainted area remain realistic, instead of creating artifacts with the unconstrained latent codes that drift far away from the training distribution. As shown in Figure 8, inversion without $\mathcal{L}_{\text{prior}}$ resulting in either obvious seams between input regions and outpainted areas or replicating input regions to the outpainted areas.

We evaluate the effect of different m values of $\mathcal{L}_{\text{ms}}$ and $\mathcal{L}_{\text{div}}$. We demonstrate that the diversity losses provide significant improvement in diversity score without compromising visual quality by seeking distinctive latent codes.

4.2. Qualitative Evaluation

In this section, we demonstrate the visual quality and diversity of the proposed method, and present applications, including categorical generation, panorama generation, and outpainting from different shapes and directions. Please refer to the supplementary materials for more visual results.

Visual quality. In Figure 4, we compare the visual quality of outpainting results from the proposed InOut against baselines. The results show that InOut is generally more realistic, coherent, diverse, exhibit more novel structures/objects, yet introduces fewer noticeable artifacts. In contrast, Boundless [27] tends to introduce raindrop-shaped artifacts, DeepFillv2 [33, 34] creates blurry extensions, while NS-outpaint [31] and Image2stylegan++ [2] frequently generates strong artifacts and obvious color differences.

Diverse outpainting. In Figure 5, we show the diverse outpainting results when $m = 3$. The results show that the proposed diversity loss enables the inversion pipeline to seek for different outpainting solutions. Notice that despite the variety of outpainting solutions, all inverted results remain visually compelling and matches the real-image input.

Categorical Generation. Figure 6 shows the results of categorical manipulation enabled with the InOut-C variant. Users can insert class-specific objects or manipulate the outpainted landscape structure with the categorical conditions

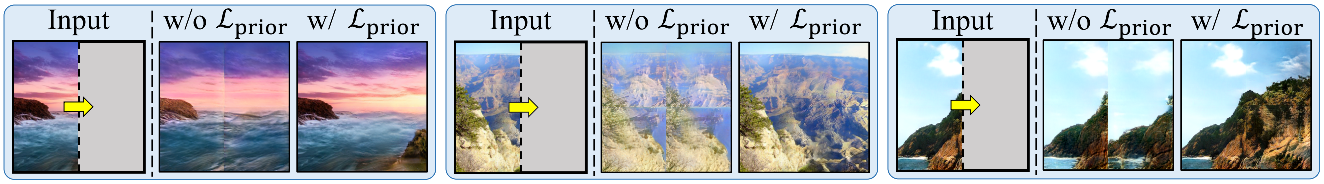

Figure 8. **Qualitative on $\mathcal{L}_{\text{prior}}$.** Without $\mathcal{L}_{\text{prior}}$, the inversion will overfit to the reconstruction loss, resulting in latent codes extremely far away from the training distribution. The outpainted area result in obvious seams (left) or replication of the input image (middle and right).

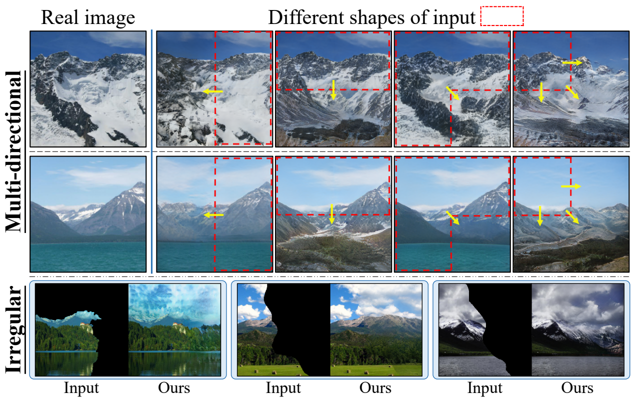

Figure 9. **Multi-directional and irregular boundary outpainting.** Our pipeline can inherently tackle *(top)* different outpainting directions and *(bottom)* irregular input shapes.

of the two micro-patches on the right. The generator is able to automatically complete the background as well as blending the presented objects into the scene.

Panorama generation. Our framework naturally supports panorama generation by recursively taking previous outpainted micro-patches as the new inversion target. Figure 7 shows the panoramas generated from our method by outpainting to the left and right. The results show that the recursively outpainted area contains highly diverse structures without repeating patterns.

Multi-directional and irregular-boundary outpainting. For brevity, most results presented are generated given two micro-patches on the left-hand side as inputs. Nevertheless, the proposed method can perform outpainting from various directions with different input shapes and even arbitrary input shape. In the top of Figure 9, from left to right, we demonstrate outpainting results generated from the right, from the top, given three micro-patches, and given one micro-patch. In the bottom of Figure 9, we present outpainted results given inputs of irregular boundaries.

Generalizability. To demonstrate the generalizability of the proposed method, we present the results on the LSUN Church [32] dataset. As shown in Figure 10, the proposed method can be applied to images with structural and artificial contents in addition to landscape images.

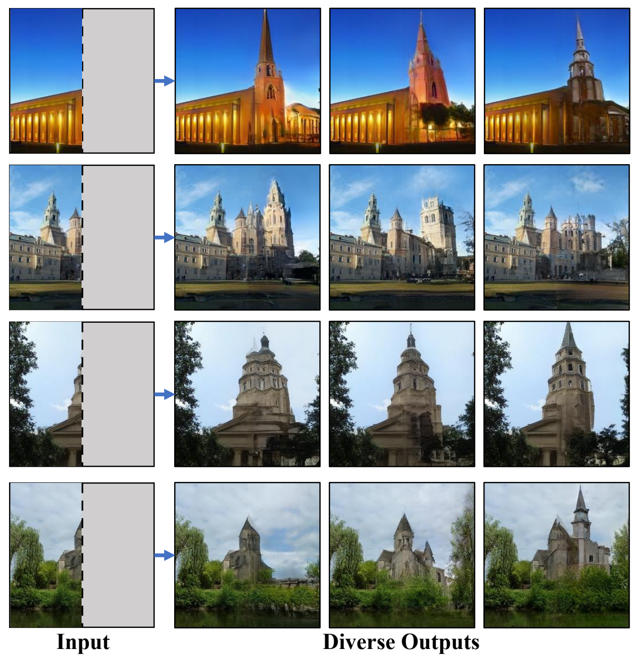

Figure 10. **Qualitative results on LSUN church.** The proposed method could handle datasets with structured objects as well.

5. Conclusion

In this work, we tackle the image outpainting task from the GAN inversion perspective. We first train a generator to synthesize micro-patches conditioned on their positions. Based on the trained generator, we propose an inversion process that seeks for multiple latent codes recovering available regions as well as predicting outpainting regions. The proposed framework can generate diverse samples and support categorical specific outpainting, enabling more flexible user controls. Qualitative and quantitative experiments demonstrate the effectiveness of the proposed framework in terms of visual quality and diversity.

6. Acknowledgements

This work is supported in part by the NSF CAREER Grant #1149783 and a gift from Snap Inc.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 2, 5, 7

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *ICML*, 2017. 3

[4] David Bau, Hendrik Strobelt, William Peebles, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. In *SIGGRAPH*, 2019. 2

[5] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In *ICLR*, 2017. 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[7] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *TNNLS*, 2018. 2

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *2017*, 2017. 2

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017. 2

[10] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 2, 5

[13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2, 3, 6

[16] Johannes Kopf, Wolf Kienzle, Steven Drucker, and Sing Bing Kang. Quality prediction for image completion. *ACM TOG (Proc. SIGGRAPH)*, 2012. 2

[17] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[18] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: generation by parts via conditional coordinating. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4512–4521, 2019. 3

[19] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. In *ICLR Workshop*, 2017. 2

[20] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 2

[21] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In *NeurIPS*, 2018. 2

[22] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019. 4

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[24] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. In *NIPS Workshop*, 2016. 2

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 5

[26] Josef Sivic, Biliana Kaneva, Antonio Torralba, Shai Avidan, and William T Freeman. Creating and exploring a large photorealistic virtual space. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 2

[27] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *ICCV*, 2019. 2, 5, 7

[28] Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R Martin, and Shi-Min Hu. Biggerpicture: data-driven image extrapolation using graph matching. *ACM TOG (Proc. SIGGRAPH)*, 2014. 2

[29] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *CVPR*, 2019. 2

[30] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 3, 4

[31] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *ICCV*, 2019. 2, 5, 7

[32] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 8

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 2, 5, 7

[34] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 2, 5, 7

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 4, 6

[36] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 2, 5

[37] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 2

[38] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 2