Contextual Outpainting with Object-Level Contrastive Learning

Jiacheng Li[1] Chang Chen[2] Zhiwei Xiong[1*]
[1]University of Science and Technology of China
[2]Noah's Ark Lab, Huawei Technologies Co., Ltd.

Figure 1. **Contextual outpainting: given foreground contents, synthesizes coherent and natural background contents.** The proposed method predicts diverse semantic layouts first, and synthesizes realistic background contents with the help of predicted semantic layouts. For each example, we show the semantic layouts (in red dashed boxes) and the outpainted images (in red boxes) generated by our method after the input foreground image and the ground truth image, respectively.

Abstract

We study the problem of contextual outpainting, which aims to hallucinate the missing background contents based on the remaining foreground contents. Existing image outpainting methods focus on completing object shapes or extending existing scenery textures, neglecting the semantically meaningful relationship between the missing and remaining contents. To explore the semantic cues provided by the remaining foreground contents, we propose a novel ConTextual Outpainting GAN (CTO-GAN), leveraging the semantic layout as a bridge to synthesize coherent and diverse background contents. To model the contextual correlation between foreground and background contents, we incorporate an object-level contrastive loss to regularize the learning of cross-modal representations of foreground contents and the corresponding background semantic layout, facilitating accurate semantic reasoning. Furthermore, we improve the realism of the generated background contents via detecting generated context in adversarial training. Extensive experiments demonstrate that the proposed method achieves superior performance compared with existing solutions on the challenging COCO-stuff dataset. Project page: https://ddlee-cn.github.io/cto-gan.

*Corresponding author: zwxiong@ustc.edu.cn.

1. Introduction

Image outpainting, also referred to as image extrapolation or image extension, is a long-lived task in computer vision. Many real-world scenarios have a strong demand for high-quality image extrapolations, like simulating different views of the current visual content in virtual reality. Early image outpainting methods rely on a retrieval and stitching process to extend image patches [18, 49, 69]. Recently, learning-based methods have made impressive progress in synthesizing visually pleasing results [13, 21, 53, 61]. However, existing image outpainting methods mainly focus on completing object shapes or extending existing scenery textures. The contextual relationship between foreground and background contents remains unexplored.

In this work, we study a variant of the outpainting problem, named contextual outpainting, which aims to synthesize coherent and natural background contents from the remaining foreground contents, as shown in Fig. 1. As humans, it is easy for us to hallucinate the empirical context given common objects, since we relate objects with their context unconsciously in everyday life. There are many potential applications of contextual outpainting techniques, such as generating plausible backgrounds for the salient objects in online advertising, film making, and aug-

mented reality. However, for machines, the task of contextual outpainting is significantly harder than the previous image completion tasks (*i.e.* inpainting and outpainting) in two ways. Firstly, the assumption of information redundancy is violated because the foreground and background contents share almost nothing in common in terms of appearance. Secondly, to utilize the constraint provided by the remaining foreground contents, it is necessary to understand the correlations inside the scene at the semantic level.

To address the above obstacles, we utilize the semantic layout as a bridge and exploit the contextual correlation between foreground and background contents in a generative way. Specifically, as shown in Fig. 2, we propose a novel *ConTextual Outpainting GAN (CTO-GAN)*, that infers the possible semantic layout from the foreground contents first and then synthesizes the corresponding background contents under its guidance. We predict diverse semantic layouts from the remaining foreground contents with a Variational Auto-Encoder (VAE).

To better model the contextual correlation between foreground and background contents at the semantic level, we propose an object-level contrastive loss to assist the learning of representations of the foreground contents and the background semantic layout. Specifically, we encode the features of foreground pixels and background semantic layouts into the same cross-modal embedding space and regularize the learning of their representations in a *"relating-by-contrasting" paradigm*, where the network is encouraged to pull the given foreground contents to the coherent semantic layouts and push out-of-context semantic layouts away.

Furthermore, to prevent the discriminator from making lazy decisions merely based on the untouched foreground contents, we incorporate an additional *context-aware discriminator* to detect which region of the generated image is fake, making it harder for the generator to fool the discriminator and thus improving the quality of generated images. We conduct extensive experiments on the challenging COCO-stuff dataset [5] and show that our method is able to generate coherent and diverse background contents, outperforming existing solutions.

2. Related Work

Image outpainting. Early outpainting methods first search similar patches from a candidate pool, and then stitch the retrieved patches with the input image to complete extrapolation [18, 49, 69]. Recently, learning-based methods have been introduced to take advantage of learned representations from large datasets [13, 21, 53, 61]. Following works are conducted based on edge guidance [26, 28, 54, 55], instance mask [4, 19], patch rearrangement [20], and GAN inversion [8]. The aforementioned methods mainly focus on extending a regular portion like the center or a half of the image and producing existing scenery textures or completing image shapes. The closest work to ours is multimodal image outpainting [66], which aims to conquer the mode collapse phenomenon in generating background contents with regularized normalized diversification. Differently, we are interested to model the joint distribution of foreground and background contents and hallucinate coherent context for the remaining foreground content.

Image inpainting. Existing inpainting methods can be divided into two groups: single-solution and multi-solution. Most early single-solution methods are based on diffusion [3, 24] and patch-matching [2, 10]. Recently, learning-based methods model the image inpainting task as a conditional generation problem, leveraging a large dataset to extract powerful priors [17, 25, 31, 33, 38, 52, 60, 62, 63, 64, 72]. Numerous works introduce cues like edges [35, 58, 59], semantic layouts [43], class labels [22], smoothed images [32, 41], and semantic textures [27] to guide the prediction of missing contents. In parallel, multi-solution methods aim to synthesize multiple plausible results given one corrupted image. Methods based on VAE [70, 71], GAN [34], VQ-VAE [39], and transformers [11, 48] are proposed. The extracted priors from large datasets mainly focus on intra-class consistency, while our method infers inter-class priors based on the modeling of the contextual relationship between foreground and background contents.

Contrastive learning. With the progress made in self-supervised representation learning [7, 14, 45], contrastive learning attracts increasing interest from the community. Most image-level contrastive learning methods rely on an elaborately designed augmentation process to generate suitable positive samples. Recent research interests shift from image-level to pixel-level or object-level, where the positive and negative samples can be naturally defined. Progress is made in fields like object detection [56], semantic segmentation [51], and object-level representation learning [57]. Along the other line, contrastive learning methods for cross-modal data such as depth map, semantic layout, audio, and text have been introduced [1, 44, 65]. Our method distinguishes itself from the above works by operating at the object level (foreground vs background) and across modality (image vs semantic layout).

Context modeling. Contextual information has been exploited in many computer vision tasks, such as visual recognition [9, 12, 15], representation learning [36, 38], and dynamics prediction [47]. Recently, learning-based methods which synthesize and insert objects based on their context [23, 67] or predict context from objects [40] have been proposed. These methods operate on either images or segmentation maps. In comparison, we adopt a "relating by contrasting" paradigm to model the cross-modal contextual relationship between the foreground contents and the background semantic layout.

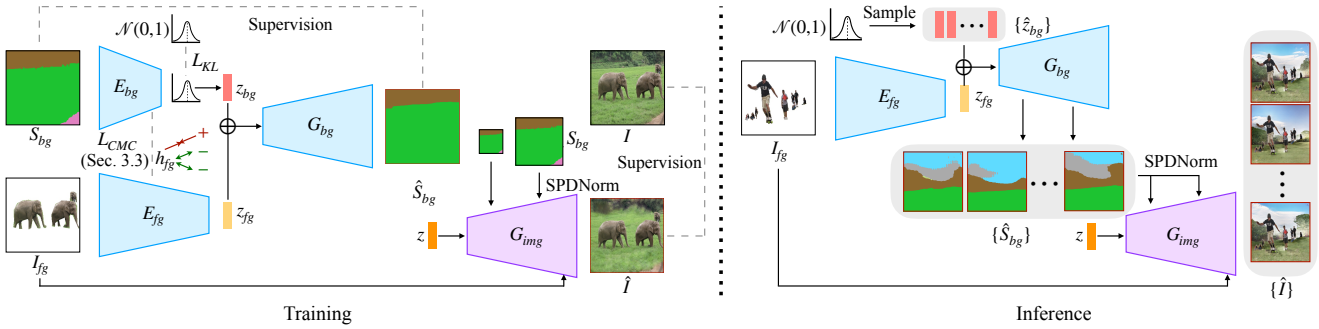
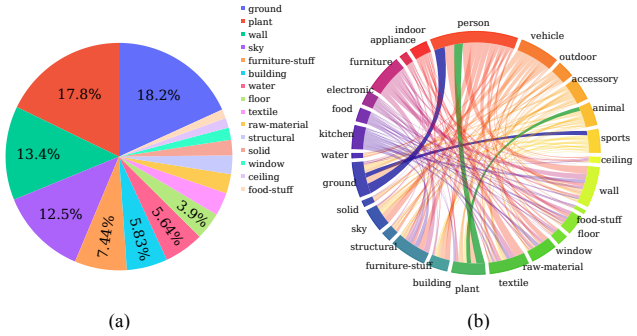


Figure 2. The overview of the proposed CTO-GAN. Left: We train the proposed method in two stages independently: semantic reasoning (blue) and content generation (purple). Firstly, the input foreground image I_{fg} and the background semantic layout S_{bg} are encoded by the foreground encoder E_{fg} and the background encoder E_{bg} into latent codes z_{fg} and z_{bg}, which are then decoded into semantic layout by the layout generator G_{bg}. Secondly, with S_{bg} as conditional input, the image generator G_{img} learns to outpaint I_{fg} to obtain the final output image $\hat{I}$. Note we omit the discriminators in the illustration for similicity. Right: At the inference time, $\{\hat{z}_{bg}\}$ are sampled from a known distribution, say $\mathcal{N}(0, 1)$. Then, G_{bg} and G_{img} synthesize diverse background semantic layouts and contents from $\{\hat{z}_{bg}\}$ and I_{fg}. $\bigoplus$ denotes the concatenation operation. Detailed architectures of these components are provided in the supplementary material.



(a) (b)

Figure 3. (a) Statistics of pixel area grouped by background classes in the COCO-Stuff dataset [5]. (b) The contextual relationship between foreground and background classes revealed by counting co-occurrences. We highlight the contextual correlation of `person` + `sports` → `ground` (blue) and `person` + `animal` → `plant` (green).

3. Contextual Outpainting

3.1. Revealing Contextual Relationships

Recent learning-based completion methods benefit from extracting appearance priors from a large dataset. For example, the COCO-stuff dataset [5, 30] contains versatile image patches with both thing (foreground) and stuff (background) categories. The dataset-level distribution of background classes can be described with the pixel area statistics grouped by class labels, as shown in Fig. 3(a). With the extracted patch appearance priors of different classes, learning-based completion methods are capable of filling the missing region based on the intra-class similarity. But these intra-class appearance priors are not enough for the contextual outpainting task, whose crucial point is understanding the contextual relationship *between* the remaining foreground and missing background classes. Intuitively, we reveal the inter-class contextual correlations inside the dataset by counting co-occurrences among supercategories. As illustrated in Fig. 3(b), a group of objects with certain foreground classes is more likely to appear in a specific context. For example, `person` and `sports` classes are related to context `ground`, while `person` and `animal` are related to `plant`. From this point of view, we assume that the images which share similar foreground classes would share similar background context and reorganize the dataset into image groups. Inside each group, the foreground images should be associated with shared contextual semantics. We build these associations in a generative way, by introducing the semantic layout as bridging information and setting these shared similar semantic layouts as training targets for each group of foreground images.

3.2. ConTextual Outpainting GAN (CTO-GAN)

From the above observation, we design CTO-GAN that leverages the semantic layout as a bridge to model the contextual correlation between foreground and background contents. The benefit of bridging the foreground and background contents with the semantic layout is two-fold. Firstly, the semantic layouts lie in a more compact domain, which is easier to be abstracted by neural networks. Secondly, it explicitly describes the intermediate semantic reasoning result from the remaining foreground contents, making our method more explainable. As illustrated in Fig. 2, we infer the possible semantic layout $\hat{S}_{bg}$ from the foreground image I_{fg} first, and then obtain the outpainted image $\hat{I}$ with the predicted $\hat{S}_{bg}$ as the conditional signal.

During training, the proposed CTO-GAN contains two independent stages: semantic reasoning and content generation. In the semantic reasoning stage, a conditional VAE

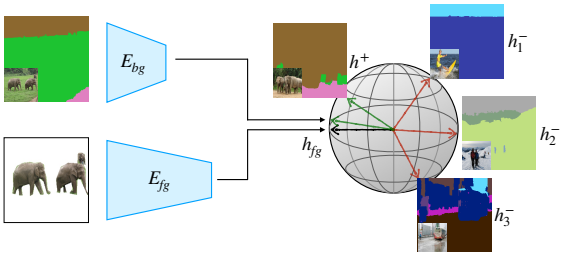


Figure 4. "Relating-by-contrasting" paradigm. The proposed CMC loss regularizes the learning procedure of encoders by pulling h_{fg} to context-coherent semantics (h^+) and pushing out-of-context semantics (h^-) away. We show the original images on the bottom-left corner for reference.

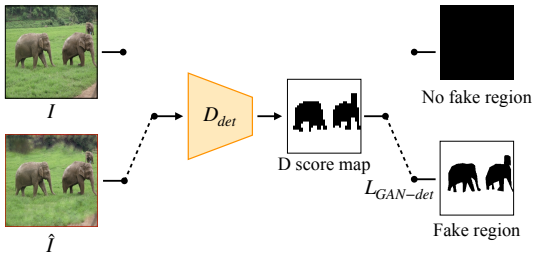


Figure 5. Context-aware discriminator. The proposed context-aware discriminator learns to detect the region of synthetic context with the supervision of the ground truth mask.

is trained to model the joint distribution of foreground and background contents. It is composed of a background encoder E_{bg}, a foreground encoder E_{fg}, and a background layout generator G_{bg}. Given a foreground image I_{fg} and a background semantic layout S_{bg}, E_{fg} encodes I_{fg} into a latent code z_{fg}, and E_{bg} encodes S_{bg} into a distribution, from which the background latent code z_{bg} is resampled. Then, G_{bg} decodes z_{fg} and z_{bg} into $\hat{S}_{bg}$. In the content generation stage, the image generator G_{img} learns to outpaint I_{fg} to obtain $\hat{I}$. We use the SPDNorm [37] condition scheme and upsample from a random noise z.

During inference, the background semantic layout plays a bridging role between foreground and background contents. Sampled from the normal distribution, latent codes $\{\hat{z}_{bg}\}$ are decoded into multiple semantic layouts $\{\hat{S}_{bg}\}$ under the condition provided the foreground latent code z_{fg}. Then G_{img} outpaints I_{fg} to obtain diverse results $\{\hat{I}\}$ with the help of $\{\hat{S}_{bg}\}$.

3.3. Relating-by-Contrasting Paradigm

To better exploit the contextual relationship between foreground and background contents, we embrace the idea of object-level contrastive learning to encode contents in the semantic reasoning stage. Specifically, with the foreground encoder E_{fg} and the background encoder E_{bg}, we map the foreground image I_{fg} and the background semantic layout S_{bg} into the same cross-modal embedding space. As illustrated in Fig. 4, in this shared space, the foreground representation h_{fg} is viewed as an anchor, the background representations from the same image group act as positive samples h^+, and the background representations from the other image groups serve as negative samples h^-. From this point of view, we formulate the following cross-modal contrastive (CMC) loss for training as

$$L_{CMC}(h_{fg}, h^+, h^-) = \\ -\log\left[\frac{\exp\left(h_{fg}\cdot h^+/\tau\right)}{\exp\left(h_{fg}\cdot h^+/\tau\right)+\sum_{n=1}^{N}\exp\left(h_{fg}\cdot h_n^-/\tau\right)}\right],$$

where τ denotes the temperature value. This regularization term helps the foreground encoder to derive a better-structured embedding space by pulling closer to the context-coherent semantics and pushing away the out-of-context semantics. It enables E_{fg} to encode the foreground images according to their most related background semantics, facilitating the semantic reasoning process. In practice, we utilize another momentum encoder for E_{bg} to provide a large set of negative samples via the MoCo scheme [14]. The learned representations are then further abstracted by convolutions to obtain latent codes z_{fg} and z_{bg}.

3.4. Context-Aware Discriminator

Since the foreground image is preserved in the contextual outpainting task, the vanilla discriminator that distinguishes the realism of the generated image may be easily fooled by the almost untouched foreground pixels. To address this issue, we propose a context-aware discriminator that detects the synthesized region of the generated images and apply it to the content generation stage for context-aware adversarial training. As shown in Fig. 5, the context-aware discriminator D_{det} predicts a score map, indicating the probability to be real or fake for every spatial location. We use the binary cross-entropy (BCE) criterion and the input mask as the target to supervise the learning of D_{det}. During optimization, the following losses are alternatingly updated as

$$L_{GAN-det}(D_{det}) = \mathcal{E}(D_{det}(\hat{I}), \mathbf{m}) + \mathcal{E}(D_{det}(I), \mathbf{m}^0),$$
$$L_{GAN-det}(G_{img}) = \mathcal{E}(D_{det}(\hat{I}), \mathbf{m}^0),$$

where $\mathcal{E}$ denotes the BCE criterion, $\mathbf{m}$ denotes the ground truth mask, and $\mathbf{m}^0$ denotes a mask tensor with all zero values, indicating no fake region in the ground truth image I.

3.5. Loss Functions

In addition to the CMC loss for semantic reasoning, we utilize the Kullback-Leibler divergence term to regularize the sampling of z_{bg} into a the normal distribution as

$$L_{KL}(E_{bg}) = \mathcal{D}_{KL}(E_{bg}(z_{bg}|S_{bg})||\mathcal{N}(0,1)),$$

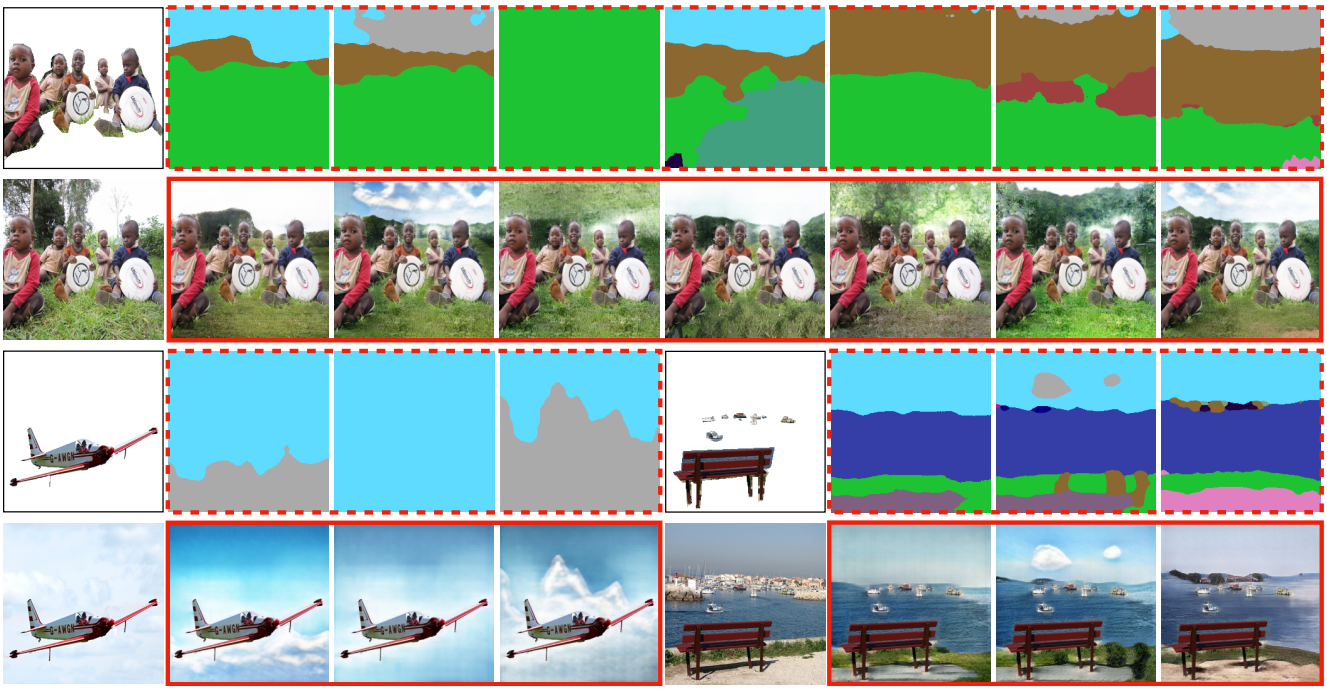


Figure 6. Visual results generated by our method. The proposed CTO-GAN predicts coherent and diverse semantic layouts with different classes and shapes, and then synthesize realistic background contents.

where $\mathcal{D}_{KL}$ denotes the KL divergence distance. Besides, cross-entropy loss and focal loss [29] are applied at multiple scales to supervise the generation of the background semantic layout. Following previous work [50], a multi-scale patch discriminator is incorporated for adversarial training.

In the content generation stage, we utilize the ℓ_1 distance and the feature matching distance from a pre-trained VGG network [42] to supervise the reconstruction of outpainted images. Besides the proposed context-aware discriminator, we apply a multi-scale patch discriminator with the semantic layout as conditional input to judge the realism and the alignment with the desired layout of the generated images. The multi-scale features extracted by the discriminator are used for feature-level reconstruction. The two stages of CTO-GAN are trained in parallel, and their respective loss terms are balanced by hyperparameters during optimization.

4. Experiments

4.1. Settings

Dataset. We conduct experiments on the COCO-Stuff dataset [5, 30], which contains 80 thing (foreground) categories and 91 stuff (background) categories. Collected by searching for common objects in their common context, the COCO-Stuff dataset encapsulates rich and challenging contextual correlations among classes. It includes over 118K training images and 5K validation images. We focus on the outdoor scene and omit the images with too small foreground area, resulting in 53,865 training images and 2,252 test images. During training, for each image, we construct the mask to indicate the missing background regions, where the pixels are annotated as stuff classes. We rescale all the images to 256×256 pixels. The detailed process to obtain pseudo stuff annotations for foreground regions and background semantic layouts with all-stuff annotations are provided in the supplementary material.

Comparison methods. We compare the proposed method with both single-solution and multi-solution image completion methods, including GatedConv [63] (inpainting, single-solution), Boundless [21] (outpainting, single-solution), Multimodal Image Outpainting (MIO) [66] (outpainting, multi-solution), Pluralistic Image Completion (PIC) [71] (inpainting, multi-solution) and Diverse Structures for Inpainting (DSI) [39] (inpainting, multi-solution).

Implementation details. The encoders and the generator in the semantic reasoning stage of CTO-GAN follow PIC [71]. The content generation stage of CTO-GAN is inspired by SPADE [37]. We add a UNet-like generator to aggregate the input foreground image and background features to obtain the final result. The context-aware discriminator follows the architecture of DeepLabV2 [6]. The architectures of other discriminators are similar to pix2pixHD [50], but with the projection of the semantic layout as conditional input for the one in the content generation stage. Our method is implemented in PyTorch and trained on 2 NVIDIA GTX 3090 GPUs.

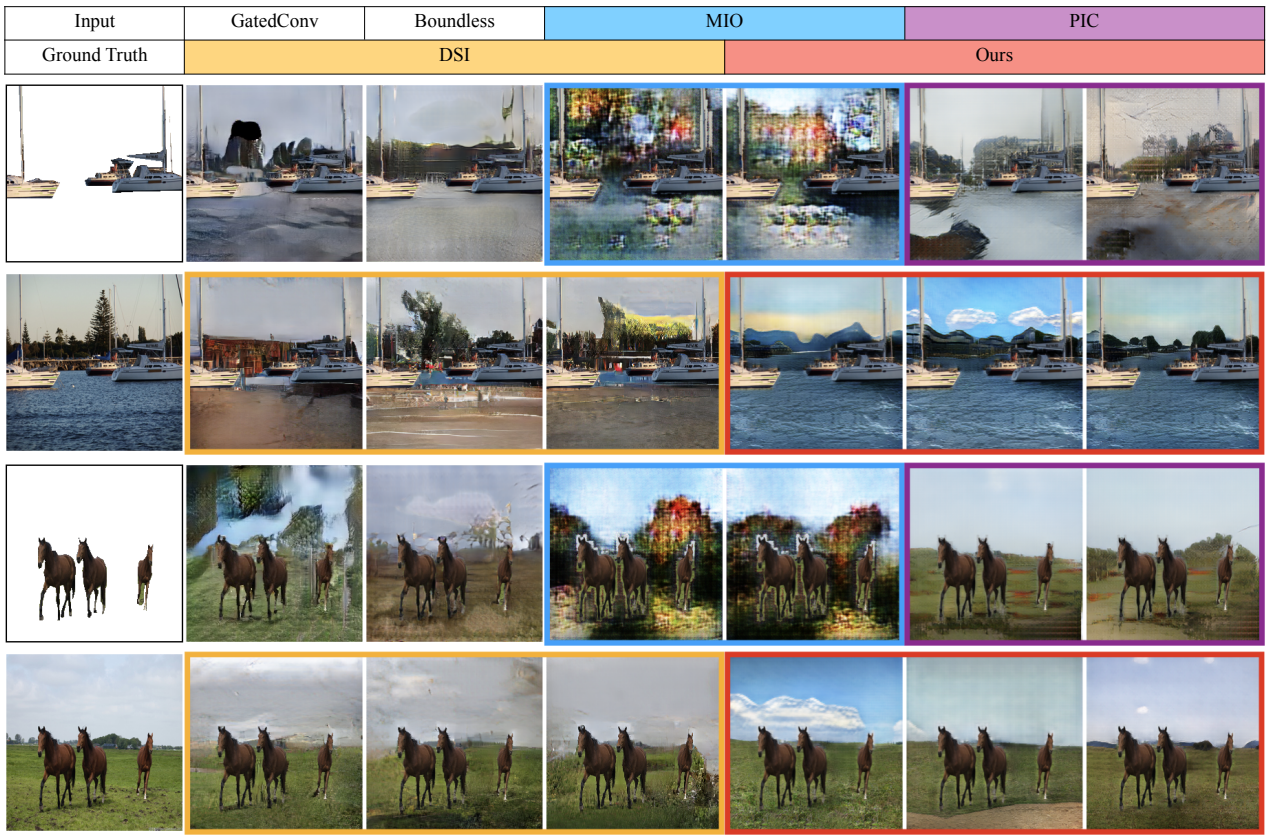


Figure 7. Qualitative comparison with existing methods. For each example, from top to bottom, from left to right, the pictures are: the input foreground image, results of GatedConv [63], Boundless [21], results of MIO [66] (in blue box), results of PIC [71] (in purple box), the ground truth image, results of DSI [39] (in yellow box) and results of our method (in red box).

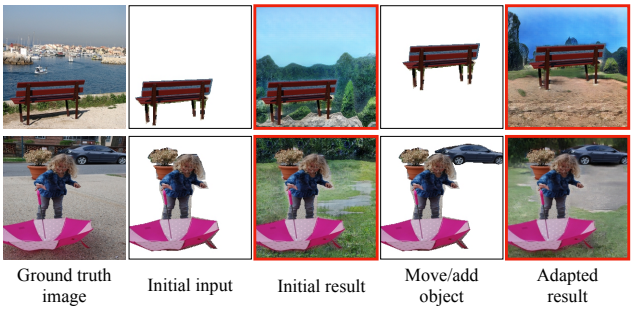


Figure 8. Scene-adaptive results generated by our method for manipulated input foreground images.

4.2. Main Results

Qualitative results. We visualize both the generated semantic layouts and outpainted results produced by our method in Fig. 6. As can be seen, our method generates both coherent and diverse semantic layouts, as well as realistic background contents. We compare the visual quality of our proposed method with existing methods, as shown in Fig. 7. Although comparison methods can generate plausible colors, they tend to predict blurry textures or irrelevant contents. In comparison, the proposed CTO-GAN is able to synthesize semantically coherent contents as well as vivid textures. For example, in the first example in Fig. 7, our method hallucinates sky views in both morning and sunset. Furthermore, we analyze the ability of scene understanding of our method by manipulating the semantic classes and spatial relations of foreground objects of the input image. As shown in Fig. 8, our method predicts the adapted results according to the semantic changes in terms of presented foreground objects and their spatial locations.

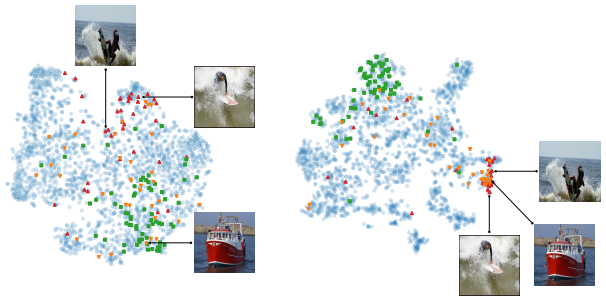
Quantitative comparison. We evaluate and compare our method with existing methods across multiple metrics. We adopt deep features based metrics FID [16] and LPIPS [68] to evaluate the perceptual quality of the outpainted images since deep features show superior consistency with the human visual system than traditional metrics [68]. To evaluate the semantic coherence of generated images, we perform semantic segmentation with a pre-trained DeepLabV2 model [6] and calculate the weighted mean intersection-over-union (mIoU) and pixel accuracy (Accu). These metrics are calculated across 10 random samples for multi-

Metric	Perceptual		Semantic		Subjective	Distortion	
	FID ↓	LPIPS ↓	mIoU ↑	Accu ↑	Avg. Rank. ↓	PSNR ↑	SSIM ↑
GatedConv	40.10	0.436	26.6	38.2	4.25	14.29	0.436
Boundless	31.11	0.411	26.8	38.8	3.40	**15.54**	0.514
MIO	60.33	0.487(0.455)	26.6	31.6	5.39	11.36(12.86)	0.433(0.462)
PIC	33.14	0.417(0.378)	25.4	39.0	3.92	14.37(15.88)	0.467(0.510)
DSI	30.74	0.395(0.351)	26.6	39.1	2.42	14.94(16.22)	0.494(0.542)
Ours	**27.34**	**0.371(0.341)**	**31.5**	**47.0**	**1.61**	14.79(16.01)	**0.529(0.560)**

Table 1. Quantitative comparison with existing methods. For multi-solution methods, we report the performances of LPIPS, PSNR and SSIM in the format of `Average(Best)`. Our method outperforms the existing solutions in almost all metrics, especially in terms of perceptual quality and semantic coherence.

	person sports	animal	person vehicle	vehicle	animal person
GatedConv	0.714	0.594	0.639	0.651	0.732
Boundless	0.704	0.602	0.639	0.644	0.732
MIO	0.594	0.476	0.574	0.595	0.632
PIC	0.699	0.569	0.614	0.630	0.716
DSI	**0.721**	0.573	0.616	0.620	0.732
Ours	0.702	**0.644**	**0.645**	**0.661**	**0.754**

Table 2. Semantic precision of top-5 foreground super-category combinations.



(a) (b)

Figure 9. The effectiveness of contrastive regularization illustrated by the t-SNE visualization of learned representations (h_{fg}) of foreground images. (a) Without contrastive regularization, the learned representations inside the same class are quite separated, leading to inaccurate semantic reasoning. (b) With the proposed contrastive regularization, foreground images with similar semantic classes are well grouped. ▲ denotes foreground images with `person` + `surfboard` classes, ▼ denotes `person` + `boat`, and ■ denotes `person` + `car` + `bus`.

solution methods. As listed in Table 1, our method outperforms the existing solutions in almost all metrics, especially in terms of perceptual quality and semantic coherence. We also report distortion metrics peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for reference, although all methods perform poorly in terms of pixel-level restoration. Additionally, we report the semantic

	FID ↓	LPIPS ↓	mIoU ↑	Accu ↑
Ours w/o contra. reg.	33.19	0.407	25.6	38.0
Ours w/o context dis.	28.31	0.387	31.1	46.4
Ours	**27.34**	**0.371**	**31.5**	**47.0**

Table 3. Ablation studies on contrastive regularization (contra. reg.) and context-aware discriminator (context dis.).

precision of the top-5 super-category combinations in Table 2. It measures the accuracy of the generated semantic classes of each method. Our method outperforms existing solutions in most combinations. For the diversity metric, we use the average pairwise LPIPS distance (LPIPS-D) between 5 samples from 1K images from the test set. As discussed in previous works [39, 48], meaningless but diverse completion results may lead to high LPIPS-D, so we show the FID metric accompany with LPIPS-D in Fig. 11(a). Our method achieves similar performance with PIC and DSI. Finally, We conduct a subjective evaluation, where 20 participants are asked to rank the results produced by comparison methods and our method for 20 random input images. We report the average ranking in Table 1. As can be seen, our method obtains the most favorable results.

4.3. Ablation Studies

The effectiveness of contrastive regularization. We validate the regularization effect of utilizing the CMC loss by comparing the learned representations from the foreground encoder. Specifically, we obtain the learned representations (h_{fg}) of the test images from the foreground encoder and run the t-SNE algorithm [46] to visualize these representations in a 2D plane. As illustrated in Fig. 9(a), without contrastive regularization, the learned representations across different foreground images are heavily intersected with each other, leading to inaccurate semantic reasoning. In comparison, as shown in Fig. 9(b), with our contrastive regularization, foreground images with the same semantic classes are well grouped. Note that even the images from

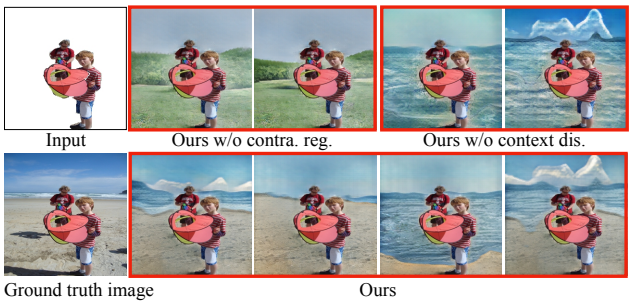


Figure 10. Results of ablation studies on contrastive regularization (contra. reg.) and context-aware discriminator (context dis.).

other classes are grouped together due to the consistency they share in context, although they are very different in terms of appearance. Besides, as shown in Fig. 10, without our contrastive regularization, intersected representations may lead to predicting common but inaccurate semantics. As listed in Table 3, our method with object-level contrastive regularization achieves better quantitative performance, since it helps reason the missing background semantics and synthesize coherent semantic layouts.

The effect of the context-aware discriminator. We retrain the content generation stage of CTO-GAN without incorporating the proposed context-aware discriminator. As listed in Table 3, the method trained with our context-aware discriminator achieves better performance in terms of perceptual quality and semantic coherence. It also can be observed in Fig. 10. Given the same semantic layouts, our method trained with the proposed context-aware discriminator generates more realistic and vivid background contents, while the one trained without the context-aware discriminator mixes different classes (sea and sand) together.

The effect of sharing semantics inside the same image group. As described in Sec. 3.1, we reorganize the images into groups and share semantics inside each group. During training, we control the percentage of involving shared semantic layouts for a training image via a hyperparameter γ. We adopt the widely used precision and recall curve to evaluate the accuracy and coverage of generated semantic classes, respectively. As shown in Fig. 11(b), without sharing semantics ($\gamma = 0$), our method achieves high precision, but with low recall. Involving shared semantic layouts during training helps increase the semantic diversity of our method. We choose $\gamma = 0.25$ to obtain a balance between semantic accuracy and coverage. More experimental results and an interactive demo are in the supplementary material.

4.4. Limitation

Our method fails in certain cases. As shown in the first example of Fig. 12, our method fails to generate semantic layouts with too many classes and fine boundaries. As shown in the second example, when the photo is

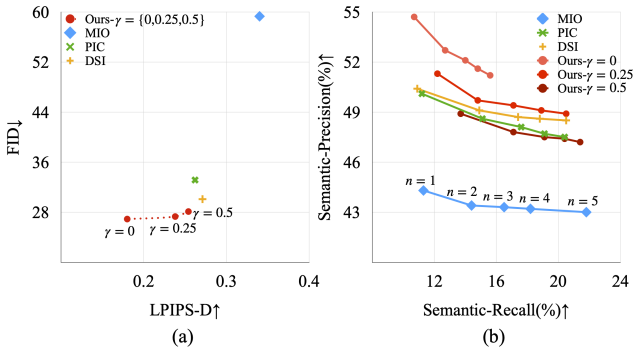


Figure 11. The quantitative comparison of diversity with existing multi-solution methods. (a) The FID vs LPIPS-D plane. (b) The semantic precision vs recall curve. n denotes the number of generated samples.

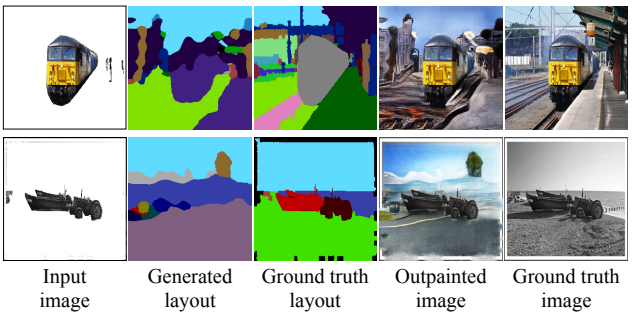


Figure 12. Failure cases of our method.

monochrome, our assumption that the semantic layout provides necessary cues for background generation is violated, leading to an inconsistent style in the outpainted image. Our future work would include introducing more explicit constraints like bounding boxes and scene graphs to facilitate the reasoning and generation of background semantics and contents.

5. Conclusion

In this work, we propose CTO-GAN, which generates coherent and diverse background contents according to the remaining foreground contents. Our method leverages the semantic layout as a bridge and utilizes contrastive regularization to model the contextual correlation between foreground and background contents. We conduct extensive experiments on the challenging COCO-stuff dataset and demonstrate the semantic reasoning ability of our method and its superiority over existing solutions.

Acknowledgement

We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, and National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2

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

[3] Marcelo Bertalmío, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. Image inpainting. In *ACM SIGGRAPH*, 2000. 2

[4] Richard Strong Bowen, Huiwen Chang, Charles Herrmann, Piotr Teterwak, Ce Liu, and Ramin Zabih. Oconet: Image extrapolation by object completion. In *CVPR*, 2021. 2

[5] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 2, 3, 5

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 5, 6

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[8] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. In&out : Diverse image outpainting via GAN inversion. *arXiv:2104.00675*, 2021. 2

[9] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. *Pattern Recognit. Lett.*, 33(7):853–862, 2012. 2

[10] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B. Goldman, and Pradeep Sen. Image melding: combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.*, 31(4):82:1–82:10, 2012. 2

[11] Ye Deng, Siqi Hui, Sanping Zhou, Deyu Meng, and Jinjun Wang. Learning contextual transformer network for image inpainting. In *ACM MM*, 2021. 2

[12] Santosh Kumar Divvala, Derek Hoiem, James Hays, Alexei A. Efros, and Martial Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 2

[13] Dongsheng Guo, Hongzhi Liu, Haoru Zhao, Yunhao Cheng, Qingwei Song, Zhaorui Gu, Haiyong Zheng, and Bing Zheng. Spiral generative network for image extrapolation. In *ECCV*, 2020. 1, 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 4

[15] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 2

[16] 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 *NeurIPS*, 2017. 6

[17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, 2017. 2

[18] Biliana Kaneva, Josef Sivic, Antonio Torralba, Shai Avidan, and William T. Freeman. Infinite images: Creating and exploring a large photorealistic virtual space. *Proc. IEEE*, 98(8):1391–1407, 2010. 1, 2

[19] Bholeshwar Khurana, Soumya Ranjan Dash, Abhishek Bhatia, Aniruddha Mahapatra, Hrituraj Singh, and Kuldeep Kulkarni. Semie: Semantically-aware image extrapolation. In *ICCV*, 2021. 2

[20] Kyunghun Kim, Yeohun Yun, Keon-Woo Kang, Kyeongbo Kong, Siyeong Lee, and Suk-Ju Kang. Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning. In *WACV*, 2021. 2

[21] Dilip Krishnan, Piotr Teterwak, Aaron Sarna, Aaron Maschinot, Ce Liu, David Belanger, and William T. Freeman. Boundless: Generative adversarial networks for image extension. In *ICCV*, 2019. 1, 2, 5, 6

[22] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided GAN based semantic inpainting. In *CVPR*, 2020. 2

[23] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. In *NeurIPS*, 2018. 2

[24] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, 2003. 2

[25] Jiacheng Li, Zhiwei Xiong, Dong Liu, Xuejin Chen, and Zheng-Jun Zha. Semantic image analogy with a conditional single-image GAN. In *ACM MM*, 2020. 2

[26] Yijun Li, Lu Jiang, and Ming-Hsuan Yang. Controllable and progressive image extrapolation. In *WACV*, 2021. 2

[27] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Image inpainting guided by coherence priors of semantics and textures. In *CVPR*, 2021. 2

[28] Han Lin, Maurice Pagnucco, and Yang Song. Edge guided progressively generative image outpainting. In *CVPR Workshops*, 2021. 2

[29] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 3, 5

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

[32] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *ECCV*, 2020. 2

[33] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *ICCV*, 2019. 2

[34] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. PD-GAN: probabilistic diverse GAN for image inpainting. In *CVPR*, 2021. 2

[35] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *ICCV Workshops*, 2019. 2

[36] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*,

2016. 2

[37] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 4, 5

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

[39] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In *CVPR*, 2021. 2, 5, 6, 7

[40] Xiaotian Qiao, Quanlong Zheng, Ying Cao, and Rynson W. H. Lau. Tell me where I am: Object-level scene context prediction. In *CVPR*, 2019. 2

[41] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *ICCV*, 2019. 2

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[43] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C.-C. Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. In *BMVC*, 2018. 2

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2

[45] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2

[46] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008. 7

[47] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, 2016. 2

[48] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, 2021. 2, 7

[49] Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R. Martin, and Shi-Min Hu. Biggerpicture: data-driven image extrapolation using graph matching. *ACM Trans. Graph.*, 33(6):173:1–173:13, 2014. 1, 2

[50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 5

[51] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 2

[52] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *NeurIPS*, 2018. 2

[53] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *CVPR*, 2019. 1, 2

[54] Yaxiong Wang, Yunchao Wei, Xueming Qian, Li Zhu, and Yi Yang. Rego: Reference-guided outpainting for scenery image. *arXiv:2106.10601*, 2021. 2

[55] Yaxiong Wang, Yunchao Wei, Xueming Qian, Li Zhu, and Yi Yang. Sketch-guided scenery image outpainting. *IEEE Trans. Image Process.*, 30:2643–2655, 2021. 2

[56] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *arXiv:2106.02637*, 2021. 2

[57] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 2

[58] Shunxin Xu, Dong Liu, and Zhiwei Xiong. Edge-guided generative adversarial network for image inpainting. In *VCIP*, 2017. 2

[59] Shunxin Xu, Dong Liu, and Zhiwei Xiong. E2I: generative inpainting from edge to image. *IEEE Trans. Circuits Syst. Video Technol.*, 31(4):1308–1322, 2021. 2

[60] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, 2018. 2

[61] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *ICCV*, 2019. 1, 2

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

[63] 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, 6

[64] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *ECCV*, 2020. 2

[65] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 2021. 2

[66] Lingzhi Zhang, Jiancong Wang, and Jianbo Shi. Multimodal image outpainting with regularized normalized diversification. In *WACV*, 2020. 2, 5, 6

[67] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *ECCV*, 2020. 2

[68] 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. 6

[69] Yinda Zhang, Jianxiong Xiao, James Hays, and Ping Tan. Framebreak: Dramatic image extrapolation by guided shift-maps. In *CVPR*, 2013. 1, 2

[70] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. UCTGAN: diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, 2020. 2

[71] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 2, 5, 6

[72] Heliang Zheng, Jianlong Fu, Yanhong Zeng, Jiebo Luo, and Zheng-Jun Zha. Learning semantic-aware normalization for generative adversarial networks. In *NeurIPS*, 2020. 2