Shadow-Enlightened Image Outpainting

Hang Yu[1], Ruilin Li[2], Shaorong Xie[1], Jiayan Qiu[3]*
[1]School of Computer Engineering and Science, Shanghai University, China
[2]Institute of Artificial Intelligence, Shanghai University, China
[3]School of Computing and Mathematical Sciences, Univerisity of Leicester, United Kingdom
{yuhang, ruilinli, srxie}@shu.edu.cn, jiayan.qiu.1991@outlook.com

Abstract

Conventional image outpainting methods usually treat unobserved areas as unknown and extend the scene only in terms of semantic consistency, thus overlooking the hidden information in shadows cast by unobserved areas, such as the invisible shapes and semantics. In this paper, we propose to extract and utilize the hidden information of unobserved areas from their shadows to enhance image outpainting. To this end, we propose an end-to-end deep approach that explicitly looks into the shadows within the image. Specifically, we extract shadows from the input image and identify instance-level shadow regions cast by the unobserved areas. Then, the instance-level shadow representations are concatenated to predict the scene layout of each unobserved instance and outpaint the unobserved areas. Finally, two discriminators are implemented to enhance alignment between the extended semantics and their shadows. In the experiments, we show that our proposed approach provides complementary cues for outpainting and achieves considerable improvement on all datasets by adopting our approach as a plug-in module.

1. Introduction

Given an image of a complex scene, as shown in Fig. 1(a), what will its right-side unobserved area look like? Humans can, without much effort, infer the appropriate categories and rough shapes of invisible semantics (*e.g*. objects) and further imagine an adequately realistic overall appearance. One critical cue humans utilize to reason invisible semantics is the shadows cast in the image. In this regard, we identify the shadows of each visible object, on top of which the shadows cast by the unobserved areas are found. According to their shapes and relative spatial layout, we may reasonably infer that there is a person in the unobserved area, as shown in Fig. 1(b), which aligns with its shadow and the given scene. However, existing outpainting methods [7, 26, 57, 58, 69, 74] overlook shadows in the image, leading to inferior results, as shown in Fig. 1(c), where misalignment exists between the extension and its shadows from the shadowing direction in the scene.

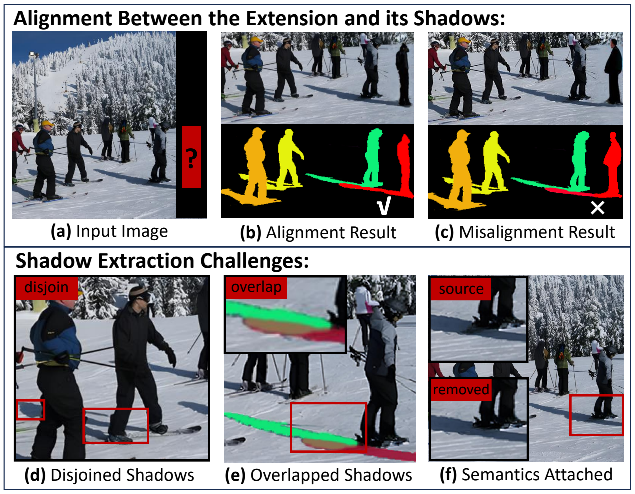


Figure 1. Given the image of a complex scene in (a) for outpainting, the extension needs to align with visible semantics. For example, the result shown in (b) is more aligned with the visible semantics from a global shadowing direction compared to the result shown in (c). For better alignment, we overcome the challenges of shadow extraction in complex scenes, such as disjoined shadows (d), overlapped shadows (e), and shadows attaching unexpected semantic information (f), respectively.

Even though prior efforts on Shape-from-Shadow [22, 32, 36, 60] also aim to extract hidden information, such as the shapes, of the semantics from their shadows in simple scenes with an ideal light source and a given semantic category, they cannot perform complex image outpainting with unknown semantics, due to there are three challenges of shadow extraction in complex scenes. 1) Given the complex spatial layout of the instances, the shadows are usually occluded by other instances, resulting in disjoined shadows of a single object, as shown in Fig. 1(d), which leads to inaccu-

*Corresponding author

rate perception or misinterpretation of invisible semantics. 2) Given the complex illumination of the scene, the shadows cast by different semantics may overlap, thus causing information loss for the unobserved area, as shown in Fig. 1(e). 3) Given the complex spatial layout of the scene, the shadows usually contain semantic information about the planes on which they are cast, which may lead to shadow-irrelevant understanding, as shown in Fig. 1(f).

To perform image outpainting by utilizing the hidden information in the shadows, we propose a novel approach that explicitly looks into the correlation between shadows from the unobserved area and the visible semantics, the shadows with semantic removal, and the consistency between local and global alignment.

Specifically, as depicted in Fig. 2, given an input image, we first conduct shadow extraction to obtain those shadow regions belonging to unobserved instances. Then, we utilize the shadow representations from unobserved instances to predict the scene layout of these instances, and then concatenate them with the shadow representation of instances in unobserved areas for layout-to-image processing. Then, the instance-level shadow representations are concatenated to predict the scene layout of each unobserved instance and outpaint the unobserved areas for the outpainting image. Finally, two discriminators are implemented to enhance alignment between the extended semantics and their shadows.

Our contribution is therefore a novel framework designated to achieve image outpainting by utilizing the shadows, which is the first attempt to the best of our knowledge. This is accomplished by extracting the desired shadow information, followed by scene layout expansion and layout-to-image conversion to produce image outpainting. The whole pipeline is end-to-end trainable. By adopting our proposed framework as a plug-in module, the outpainting methods [69] have achieved considerable quantitative and qualitative performance improvement on all datasets.

2. Related Work

We briefly review the prior works related to ours, including shadow processing and image outpainting.

Shadow Processing. Conventional shadow processing method reconstructs the surface shape of a single object, under ideal illuminance, from its shadow and visual information [40]. Then, a sequence of methods [1, 2, 13, 14, 31, 62, 66] is developed to improve its performance by involving carefully designed prior knowledge in the model. Then, due to the strong adaptability of deep learning, learning-based methods [22, 32] are proposed to deal with complicated environments, such as objects with non-Lambertian reflectance [32]. Moreover, shadows are utilized as cues for scene geometry information extraction [21, 23, 30, 36, 43–45, 47, 48, 67, 68], segmentation [12], object detection [9, 16, 17], and tracking [41]. Recently, a generative method [36] has been proposed to predict the shape of a single invisible object, with a specific category, from its shadow. However, none of the existing methods can infer the scene layout of the semantics from their shadows in complex scenes.

Image Outpainting. Conventional image outpainting methods extend the image with the background texture and partially observed objects towards specific directions [5, 34, 37, 52, 73, 74]. Then, information on the image margin is utilized to enable the outpainting in any direction [28, 56, 64]. Recently, the extrapolation of segmentation information has been adopted in the outpainting process [26] to add novel semantics in the extending areas. Meanwhile, the scene graph of the image is utilized to introduce new semantics by learning the semantic co-occurrence with Graph Neural Networks (GNN) [19, 33, 49, 50, 59, 69, 72]. Conventional outpainting commonly uses Generative Adversarial Networks (GAN) [15] for image conversion. However, due to the stability of diffusion, diffusion models [3, 18, 42, 51, 55] are being recognized as a promising family of generative models that have proven to be state-of-the-art sample quality for a variety of image generation benchmarks [8, 61, 70], including class-conditional image generation [10, 75], text-to-image generation [27, 51, 53, 71], image-to-image translation [24, 38, 54], layout-to-image generation [6, 76]. However, none of the existing methods takes advantage of the shadows to enhance the outpainting.

3. Method

In this section, we detail the working scheme of the proposed approach which comprises three stages, as shown in Fig. 2. In Stage 1, we utilize the pretrained model of shadow detection and instance shadow detection to grain rough instance-level shadow regions. Meanwhile, a pretrained shadow removal model is to remove unexpected semantic information from the shadows for the purified shadow feature. Then, we calculate the connectivity between shadow regions to merge disjoined shadows and perform overlapped segmentation on the shadows to restore overlapped shadows of instances. After iterative optimization, we obtain refined instance-level shadow regions. In Stage 2, the instance-level shadow representations are concatenated to predict the scene layout of each unobserved instance and outpaint the unobserved areas. In Stage 3, two discriminators are implemented to enhance alignment between the extended semantics and their shadows.

3.1. Stage 1: Shadow Extraction

Given the input image I, we first adopt a pretrained shadow detection model [77] to extract all shadow areas:

$$M_s = \mathrm{Detection}(I), \tag{1}$$

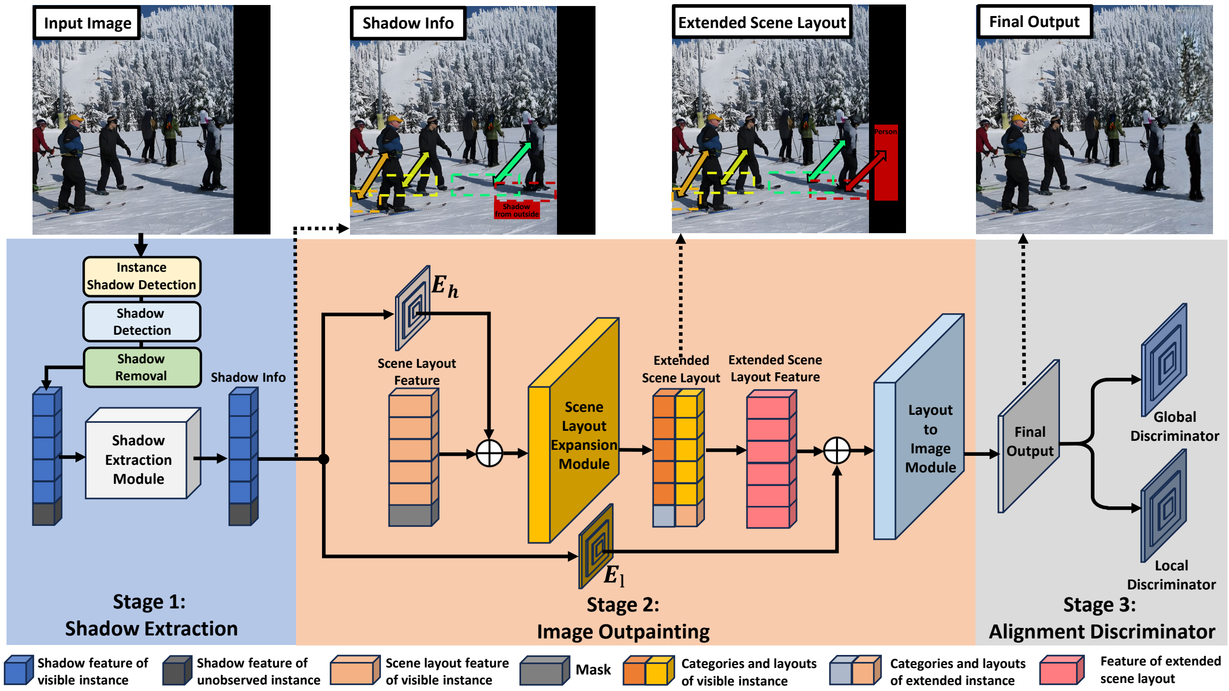

Input Image
Shadow Info
Extended Scene Layout
Final Output
Shadow from outside
Person
Instance Shadow Detection
Shadow Detection
Shadow Removal
Shadow Info
Shadow Extraction Module
E_h
Scene Layout Feature
Scene Layout Expansion Module
Extended Scene Layout
Extended Scene Layout Feature
Layout to Image Module
Final Output
Global Discriminator
Local Discriminator
E_l
Stage 1: Shadow Extraction
Stage 2: Image Outpainting
Stage 3: Alignment Discriminator
Shadow feature of visible instance
Shadow feature of unobserved instance
Scene layout feature of visible instance
Mask
Categories and layouts of visible instance
Categories and layouts of extended instance
Feature of extended scene layout

Figure 2. Illustration of the proposed approach. $\oplus$ denotes concatenation.

where M_s denotes the mask of the shadow areas, respectively. Then, a pretrained instance shadow detection model [63] is introduced to identify the shadow areas for the visible semantics in image I by:

$$M_s^v = \text{Instance\_Detection}(I). \tag{2}$$

where M_s^v denotes the instance-level mask of the shadow areas cast by the visible semantics. Consequently, a binary area-level mask of shadows cast by the unobserved area can be obtained with $M_s^u = M_s - M_s^v$. Thus, we obtain a rough instance-level shadow region.

Meanwhile, to obtain the shadow areas without the semantic information from the cast planes, we introduce a pretrained shadow removal model [39] and obtain the semantically purified shadow image I_s^p by:

$$I_s^p = M_s \odot \big(I - \text{Shadow\_Removal}(I)\big), \tag{3}$$

where $\odot$ denotes the Hadamard product.

After obtaining the shadow image of semantic removal I_s^p, we can use a convolutional net to extract the shadow feature without semantics:

$$F_s^p = \text{ConvNet}(I_s^p), F_s^p \in \mathbb{R}^{W\times H\times d}, \tag{4}$$

where the ConvNet consists of 8 Residual Blocks and preserves the resolution of the output representation to be the same as the input image. W, H denotes the width and height of the input image. d denotes the dimension of the shadow feature map.

Considering that the formation of shadows is influenced by the illumination conditions in the given scene, which can be a cue in addressing the issues of disjoined shadows and overlapped shadows, we concatenate the image I and the shadow mask M_s and fed it into a Convolutional Neural Network (CNN) to obtain the illumination feature $F^{\text{illum}} \in \mathbb{R}^{(W/16)\times(H/16)\times d}$ in the scene as a cue for the shadow regions optimization.

3.1.1 Shadow Merging Operation

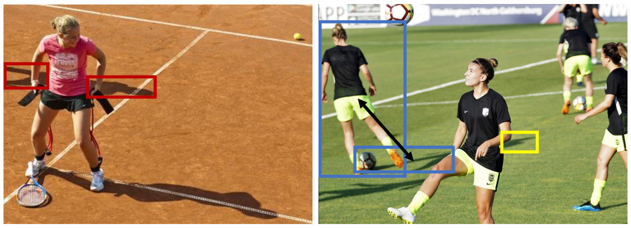
In complex scenes, due to the presence of instances that can occlude shadows, it can result in an instance having multiple disjointed shadows, as shown in Fig. 3.

To classify a single instance with disjoined shadows, we first treat each connected component in M_s^u as a shadow region weight mask. Let $M^u = \{m_x^u\} \in \mathbb{R}^{W\times H\times 1}, x \in [1, N_u]$ be the shadow weight masks that don't belong to any visible instances, where N_u denotes the number of connected components in M_s^u. For the rough instance-level mask of visible instances in M_s^v, let $M^v = \{M_y^v\} \in \mathbb{R}^{W\times H\times 1}, y \in [1, N_v]$ be each shadow weight mask of visible instance, where N_v denotes the number of visible instances. Consequently, we grain the initial instance-level shadow region M:

$$N = N_u + N_v \tag{5}$$

$$M = \{m_z\} = \{M^u, M^v\}, z \in [1, N], \tag{6}$$

For each shadow region, we extract the corresponding region-level shadow feature F by the feature map by



(a) the disjoined shadow of the unobserved instance

(b) the disjoined shadow of the visible instance

Figure 3. Samples of disjoined shadows caused by occlusion in complex scenes. (a) The shadow of an unobserved object is obstructed, creating two disjointed shadow regions. The red boxes indicate the disjoined shadows. (b) Due to the defect of the instance shadow detector, a portion of the shadow of the instance was not classified as a single instance. The blue boxes indicate the detected corresponding pair, and the yellow box indicates the undetected shadow region.

element-wise multiplying the shadow feature without semantics F_s^p and the shadow weight mask M. Then, we use the max pooling operation to resize these features to match the size of the illumination feature. Thus, we obtain the cropped feature $\bar{F}$ by:

$$F = F_s^p \odot M, F \in \mathbb{R}^{W\times H\times d}, \tag{7}$$

$$\bar{F} = \text{maxpool}(F), \bar{F} \in \mathbb{R}^{(W/16)\times(H/16)\times d}, \tag{8}$$

To improve the computation of connectivity between shadow regions, we combine the cropped features $\bar{F}$ of two regions along with the illumination feature by concatenating them together. This concatenated input is fed into an encoder $E_{connect}$, and then passed through a sigmoid activation function for connectivity between two shadow regions:

$$\bar{F} = \{\bar{f}_m\}, m \in [1, N], \tag{9}$$

$$\bar{f}_{(i,j)} = \mathrm{E}_{\text{connect}}(\text{concat}(\bar{f}_i, \bar{f}_j, F^{\text{illum}})), i \neq j \tag{10}$$

$$Connectivity_{(i,j)} = \text{Sigmoid}(\bar{f}_{(i,j)}), i \neq j \tag{11}$$

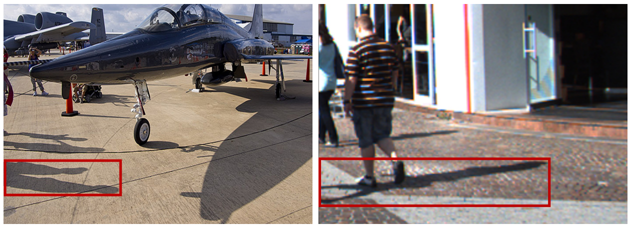
When the connectivity between two shadow regions is higher than a hand-setting threshold α_1, these two shadow regions are considered to be the shadows of the same object, accordingly merging their weight masks. Conversely, they stay unchanged. Finally, we obtain the updated shadow features F^{update} after merging operation:

$$F^{update} = \{f_m\} \in \mathbb{R}^{W\times H\times d}, m \in [1, N^m], \tag{12}$$

where N^m denotes the number of instances after merging.

3.1.2 Shadow Splitting Operation

In complex scenes, due to the complex spatial layout in the scene, it is common to have overlapped shadows between instances, as shown in Fig. 4.



(a) overlapped with shadows cast by the unobserved instances

(b) overlapped with shadow cast by the visible instances

Figure 4. Samples of overlapping shadows in complex scenes. The red boxes indicate the overlapped shadows. (a) The shadows of multiple unobserved instances overlap. (b) The shadows of unobserved instances overlap with the shadows of visible instances.

In this work, each overlapped shadow is decoupled into two intersecting graph spaces. After concatenating the cropped feature $\bar{F}$ and illumination feature F^{illum}, we input them into the Overlapping BiLayers [25]. Then, we grain two weight masks M'_0, and M'_1, each weight mask represents the shadow region of an individual instance. Then, we can obtain the "de-overlapped" shadow feature f of an instance by element-wise multiplying the weight mask M' with the purified shadow feature F_s^p:

$$M'_0, M'_1, \in \mathbb{R}^{W\times H\times 1} \tag{13}$$

$$f_0 = M'_0 \odot F_s^p, f_1 = M'_1 \odot F_s^p, \tag{14}$$

when the ratio of the combined area of two weight masks to the source area is higher than a hand-setting threshold α_2, it is considered an overlapped shadow and needs to be split. Conversely, it stays unchanged. Finally, we obtain the updated shadow features F^{update} after splitting operation:

$$F^{update} = \{f_m\} \in \mathbb{R}^{W\times H\times d}, m \in [1, N^s], \tag{15}$$

where N^s denotes the number of instances after splitting.

3.1.3 Iterative Optimization

In complex scenes, it is common to encounter situations where disjoined shadows and overlapped shadows co-occur. Therefore, we utilize an iterative combination of merging and splitting operations to optimize the categorization of all shadow regions in the image. Thus, the issues of disjoined shadows and overlapped shadows, shown in Fig. 3 and Fig. 4, are addressed. Specifically, the iterative shadow optimization is summarized as Algorithm 1.

3.2. Stage 2: Outpaiting with Shadows

After obtaining the refined instance-level shadow regions M from Stage 1, we grain the cropped shadow representations $\bar{f^u}$ of unobserved instances by:

$$\bar{f^u} = \text{maxpool}(M^u \odot F_s^p), M^u \in M, \tag{16}$$

Algorithm 1: Iterative Shadow Optimization

```
Input  : rough instance-level shadow weight mask
Output: refined instance-level shadow weight mask
$M^{(0)} \leftarrow M$
for $t = 1$ to $T$ do
    $M^{(t)} \leftarrow$ ShadowMerging($M^{(t-1)}$)
    $M^{(t)} \leftarrow$ ShadowSplitting($M^{(t)}$)
    if $\|M^{(t)} - M^{(t-1)}\| < \epsilon$ then
        break
    end
end
return $M^{(t)}$
```

which are concatenated to predict the categories and layouts of new instances and outpaint the unobserved areas.

For scene layout expansion, given a partial input image with the partial scene graph and the corresponding layout, we use GNN [11] to reveal the scene layout L_{exp} of new instances. Specifically, we pass the shadow representations $\bar{f^u}$ from the unobserved instances through an encoder E_h, and then concatenate the output with the node features f^h of the masked instances from the encoder of GNN.

$$\hat{f^h} = \mathrm{concat}(f^h, \mathrm{E_h}(\bar{f^u})) \tag{17}$$

For layout to image, given an input image I with an expanded scene layout L_{exp}, we use layout-to-image model [76] to perform layout to image conversion. Specifically, we pass the shadow representations f^u from the unobserved instances through an encoder E_l, and then concatenate the output with scene layout features f^l of the expanded instance from the encoder of the layout-to-image module.

$$\hat{f^l} = \mathrm{concat}(f^l, \mathrm{E_l}(\bar{f^u})) \tag{18}$$

Note that we focus on the design of extracting shadow representation, so producing high-accuracy predictions for scene layout and high-quality image outputs is not within the main scope of this work. Thus, the graph-based and diffusion-based designs can be replaced with any scene layout expansion and layout-to-image modules if desirable.

3.3. Stage 3: Alignment Discriminator

In stage 2, we outpaint the partial input image I into the extended image I_E with introduced instances. To enhance alignment between unobserved area and visible semantics, we adopt the pretrained instance shadow detector [63] to first extract the mask M_E^u of the unobserved area and its corresponding shadow from I_E, and then feed it into a local shadow-instance alignment discriminator to ensure the generated alignment is visually real. This is calculated as follows:

$$\begin{aligned}\min_{LD} V(LD) = {} & \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(LD(x)-b)^2] \\ & + \frac{1}{2}\mathbb{E}_{z\sim pz(z)}[(LD(M_E^u \odot G(z)) - a)^2],\end{aligned} \tag{19}$$

$$\min_{G} V(G) = \frac{1}{2}\mathbb{E}_{z\sim pz(z)}[(LD(M_E^u \odot G(z)) - c)^2], \tag{20}$$

where LD denotes the local alignment discriminator. G denotes the generator for outpainting. a and b denote the ground truth real and fake labels, respectively. c denotes the value that G wants LD to believe for the fake data.

Then, the mask M_E that consists of all semantics and their shadows in I_E is extracted, and we feed it into a global alignment discriminator to ensure the newly generated alignment is consistent with the existing ones. This is calculated as follows:

$$\begin{aligned}\min_{GD} V(GD) = {} & \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(GD(x)-b)^2] \\ & + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(GD(M_E \odot G(z)) - a)^2],\end{aligned} \tag{21}$$

$$\min_{G} V(G) = \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(GD(M_E \odot G(z)) - c)^2], \tag{22}$$

where GD denotes the global alignment consistency discriminator.

In summary, shadow extraction, scene layout expansion, and layout-to-image module are end-to-end trained together, thus the visual authenticity loss may facilitate shadow representation learning. Therefore, the final objective function of our approach is:

$$L = \lambda_1 L_t + \lambda_2 L^G_{L\text{-}GAN} + \lambda_3 L^G_{G\text{-}GAN}, \tag{23}$$

where λ_1, λ_2 and λ_3 denote the balancing weights. L_t denotes the loss adopted from the base model. $L^G_{L\text{-}GAN}$ and $L^G_{G\text{-}GAN}$ denote the alignment loss from the local alignment and global consistency.

4. Implementation Details

This section provides details of the training settings and module implementations.

Training Settings. Our approach is implemented using PyTorch [46] on four NVIDIA-RTX A6000 GPUs. During the training process, we adopt the Adam optimizer, 0.001 for the local discriminator and the global discriminator, and 0.01 for all modules in shadow representation extractors. The batch size is set to 32 and the size of the input images is 256 × 256. The training is manually stopped.

Shadow Detection Models. We adopt the popular shadow segmentation method [77] as the shadow detector, which is able to segment the shadow areas in the image. Furthermore, we adopt the state-of-the-art instance shadow detection method [63] as our instance shadow detector, which is able to detect the shadow instances and their corresponding object in the image. However, it is unable to detect shadows cast by unobserved objects.

Alignment Discriminator. We adopt diffusion timestep-dependent discriminators from Diffusion-GAN [65]. It not only can minimize the divergence between real and diffused data at the end of the process, but also minimizes the divergence between the diffused real data distribution and the diffused generator distribution over several timesteps. In training, the balancing weight λ_1 is set to be 0.5, λ_2 and λ_3 are set to 0.25. Note that the architecture of the discriminator can be adjusted based on the specific type of layout-to-image module being implemented.

Scene Layout Expansion. We adopt GTwE [11] to predict the categories and layout of unobserved instances simultaneously. GTwE has the same aforementioned hyper-parameters as [69], which are the attention hidden size d_{atten} = 512, feed forward size d_{ff} = 2048, multi-head number n_{head} = 4, and dropout = 0.1.

Layout to Image. We adopt LayoutDiffusion [76] for better outpainting the unknown areas with the condition of layouts predicted by GTwE. Note that LayoutDiffusion can not take the visible semantics as guidance to outpaint the unknown areas. Thus, we follow RePaint [38] to modify the standard denoising process for conditioning on the given image content. In each step, we sample the known region from the input and the outpaint painted part from the DDPM [18] output.

5. Experiments

In this section, we provide our experimental setups and both the objective and subjective results. Since we are not aware of any existing work that performs the same task as we do, we mainly focus on showing the promise of the proposed approach. Specifically, we evaluate the effectiveness of our approach by conducting experiments with an image outpainting pipeline proposed by [69]. Our goal is, again, to show the possibility of learning effective shadow representations thus enhancing complex scene extension, rather than trying to beat the state-of-the-art shadow detection and outpainting method. Other modules with the same functionality, provided they are end-to-end trainable, can be adopted in our approach to achieve potentially better performance.

5.1. Datasets

We evaluate our proposed approach on two common datasets, Visual Genome (VG) [29] and COCO-stuff [4], which gives an adequate amount of scene layouts.

Visual Genome [29] dataset collects 108,077 images with dense annotations of objects, attributes, and relationships. Following the setting of SG2Im [20], we divide the data into 80%, 10%, and 10% for the train, val, and test set, respectively. We select the object and relationship categories occurring at least 2000 and 500 times in the train set, respectively, leaving 178 objects and 45 relationship types, and select the images with 3 to 30 bounding boxes and ignoring all small objects. Finally, the training/validation/ test set will have 62565 / 5062 / 5096 images, respectively.

COCO-stuff [4] dataset augments a subset of the COCO dataset [35] with additional semantic categories. Thus, a total of 80 object categories (car, dog, etc.) and 91 semantic categories (sky, snow, etc.) are available, with 118K / 5K annotated images for training/validation. Following the setting of SG2Im [20], the ground truth object coordinates in the images are utilized for constructing the synthetic relationships, a total of six relationships are considered: *left of, right of, above, below, inside, and surrounding*. In addition, we use images in the train and val set with 3 to 8 objects that cover more than 2% of the image. Finally, there are 25,210 train and 3,097 val images.

5.2. Quantitative and Qualitative Result

In this section, we evaluate the quantitative and qualitative effectiveness of our proposed approach to the image outpainting task by adopting it as a plug-in module of the SOTA method SGT [69]. The outpainting performance can be divided into two steps, in which the first one focuses on predicting the category and layout of the unobserved semantic and then the second one aims to generate the outpainting images. Therefore, we evaluate our approach on both the two steps. As can be seen from Tab. 1, by adopting our approach as a plug-in module, SGT achieves significant improvement in unobserved semantic prediction and its relevant relationship estimation. Although our target is to enhance only the semantic prediction, the accurately predicted semantics consequently strengthen the relationship estimation, thus leading to overall improvement on all metrics. Then, as can be seen from Tab. 2, we achieve considerable improvement in unobserved layout prediction, which also explicitly impacts the alignment with visible semantics in addition to the categories of instances. Moreover, in order to evaluate the quality of outpainting images, we introduce the FID metric and the results shown in Tab. 3.

Then, the visual results are shown in Fig. 5 and Fig. 6, it can be found that our proposed approach is able to enhance the outpainting in the unobserved area, which is achieved by constraining both the shape and the semantic of the outpainting to be aligned with its visible shadows.

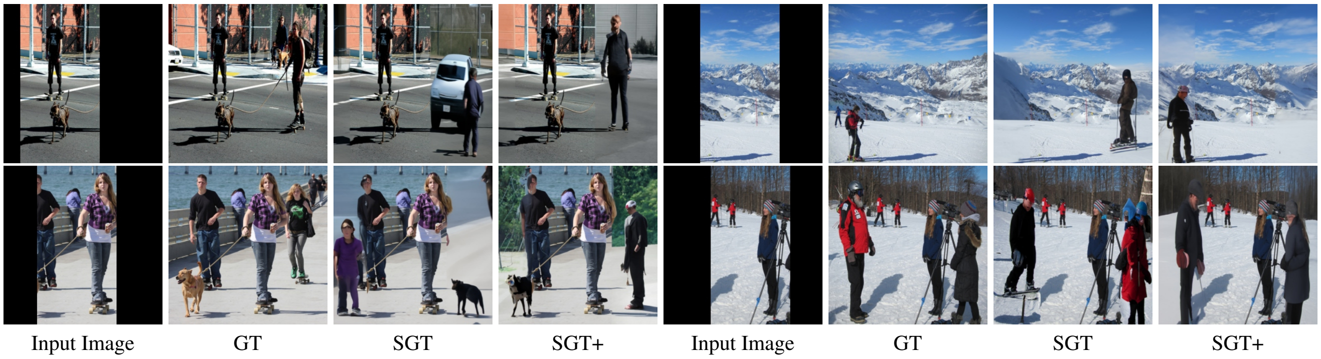


Figure 5. Visual results from SGT [69] on VG [29]. SGT+ denotes the result after adopting our approach as a plug-in module.

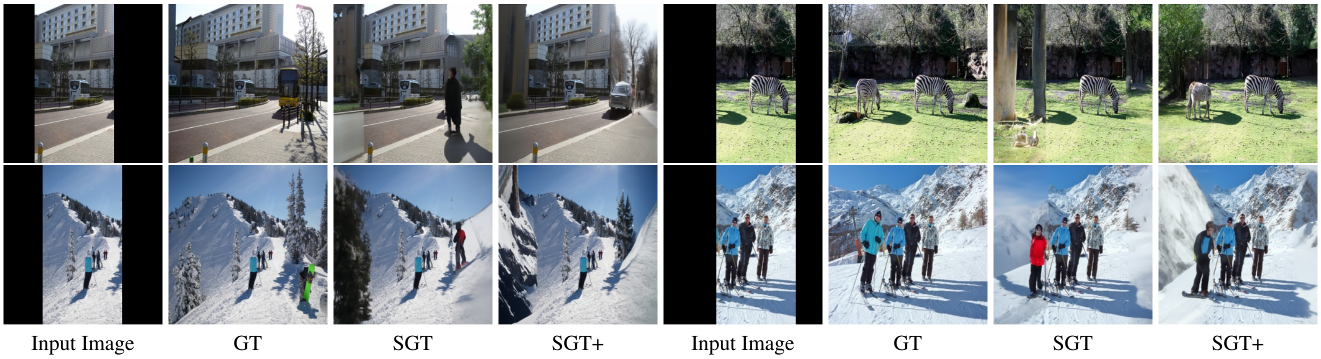


Figure 6. Visual results from SGT [69] on COCO-stuff [4]. SGT+ denotes the result after adopting our approach as a plug-in module.

5.3. Multiple Instances Extension Comparison

To verify whether our approach can enhance alignment with visible semantics in complex scenes, even with fewer shadows. We increase the number of instances requiring inference from unobserved areas to validate alignment between the extension and visible semantics shown in Fig 7.

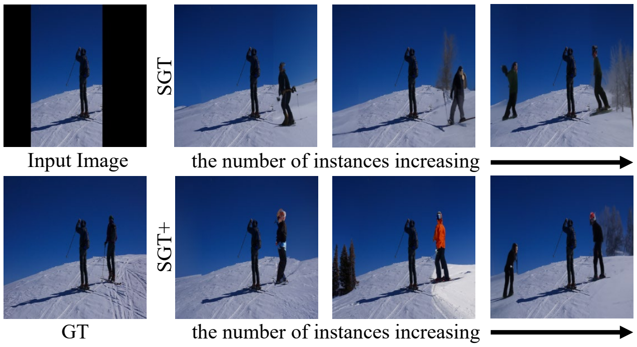

SGT
Input Image
the number of instances increasing
SGT+
GT
the number of instances increasing

Figure 7. Qualitative comparison of multiple instances extension. Each image adds one more instance from left to right.

As we can see, our method can effectively leverage shadow information in complex scenes to infer a scene layout that aligns well with the visible semantics. Specifically, instances that its shadow is not in the visible area can still align with visible semantics compared to the method without incorporating our approach, which ensures a harmonious generation of content with visible semantics.

5.4. Shadow Extension Comparison

As seen in Fig. 5 and Fig. 6, our visual results demonstrate that outpainting with our module achieves excellent alignment with visible semantics. However, there is an additional issue we need to consider: the shadows of visible instances can also affect the overall shadow consistency of the image. Therefore, we have also conducted a comparison specifically addressing this concern, as shown in Fig. 8.

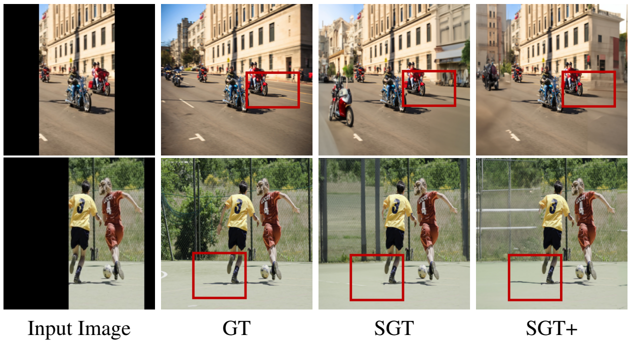


Figure 8. Qualitative comparison of shadow extension on the visible instance. Red boxes show the differences in the extended shadow of the visible instance.

As we can see, our shadow extension harmonizes the generated with the visible semantics. However, the method without incorporating our module randomly elongates or abruptly truncates shadows, resulting in a lack of global alignment among the semantic-shadow pairs in the scene.

	VG [29]				**COCO-stuff** [4]			
	Object		Relationship		Object		Relationship	
	rAVG↓	Hit@ 1/3↑	rAVG↓	Hit@ 1/3↑	rAVG↓	Hit@ 1/3↑	rAVG↓	Hit@ 1/3↑
SGT	10.51	22.9 / 45.7	5.81	36.9 / 69.4	10.54	26.4 / 52.3	2.92	25.7 / 66.4
$+F_s^p$	9.61	24.6 / 47.0	5.23	37.7 / 72.4	10.01	27.0 / 53.4	2.89	25.9 / 67.1
$+F^{illum}$	9.25	26.4 / 49.3	4.82	38.2 / 73.1	9.54	27.5 / 54.4	2.82	27.0 / 67.2
$+L^G_{L\text{-}GAN}$	10.1	23.1 / 46.5	5.24	37.4 / 71.3	10.41	26.3 / 53.1	2.98	25.9 / 66.9
$+L^G_{G\text{-}GAN}$	9.97	23.4 / 46.2	5.42	37.2 / 71.7	10.75	26.7 / 52.9	2.93	26.3 / 66.5
SGT+	**8.93**	**27.9 / 50.4**	**4.59**	**39.4 / 75.2**	**8.83**	**28.4 / 55.3**	**2.73**	**28.3 / 68.4**

Table 1. Quantitative evaluation and ablation study of scene graph expansion on VG [29] and COCO-stuff [4] datasets. ↑ denotes the higher is better, ↓ the lower is better. SGT+ denotes the pipeline that adopts our approach as a plug-in module.

mIoU	**VG**	**COCO-stuff**
SGT	22.7	32.6
$+F_s^p$	25.4	34.8
$+F^{illum}$	26.4	36.2
$+L^G_{L\text{-}GAN}$	23.6	33.4
$+L^G_{G\text{-}GAN}$	24.9	33.1
SGT+	**28.4**	**38.1**

Table 2. Quantitative evaluation and ablation study on VG [29] and COCO-stuff [4] datasets, with mIoU metric.

FID	**VG**	**COCO-stuff**
SGT	20.12	19.12
$+F_s^p$	19.89	19.02
$+F^{illum}$	19.92	18.93
$+L^G_{L\text{-}GAN}$	19.48	18.60
$+L^G_{G\text{-}GAN}$	19.52	18.59
SGT+	**19.47**	**18.38**

Table 3. Quantitative evaluation and ablation study on VG [29] and COCO-stuff [4] datasets, with FID metric.

5.5. User Studies & Ablation Study

User Studies. For image outpainting, any appropriate generation should be treated to be correct, thus without a specific target output. Therefore, in order to further evaluate the qualitative effectiveness of our approach, we conduct two user studies, where 91 users are involved to evaluate the outpainting correctness of our approach. In the first experiment, we send each user 20 randomized selected pairs of contents, in which the first one is the observed image with filling black in the unobserved area and the second one is the scene layout we predict to appear in the unobserved area, we then ask the users whether the predicted semantic could appear in the unobserved area. Finally, 89% of the selected pairs are scored to be true, which denotes that our proposed approach is able to predict the appropriate scene layout for the unobserved semantics.

In the second experiment, we send each user 20 randomized selected pairs of images, where one of them comes from SGT and the other one comes from adopting our approach as a plug-in module, we then ask the users which one is better in visual authenticity. Finally, images from ours achieve 87% better chosen, which shows the significant quality enhancement produced by our approach.

Ablation Study. We evaluate the individual effectiveness of the semantic removal, the illumination concatenation, the local alignment discriminator, and the global consistency discriminator in this experiment. The results show in Tab. 1, Tab. 2 and Tab. 3. It can be seen that the most significant enhancement in scene layout prediction comes from illumination concatenation, which shows the importance of incorporating a comprehensive illumination understanding to tackle the challenges of shadow extraction for obtaining accurate predictions. Meanwhile, the semantic removal enhances the accuracy notably, which shows that the shadow-irrelevant semantic information in the shadow areas leads to unexpected and noisy shadow representations. And the two discriminators achieve considerable enhancement in FID, which shows the importance of ensuring the local alignment between the unobserved semantics and its visible shadows and the consistency among all semantic-shadow pairs.

6. Conclusion

In this paper, we proposed a novel approach that aims to perform image outpainting by utilizing shadow information, which has never been explored. This is accomplished by extracting instance-level shadows cast by the unobserved areas from the input image. Then, the instance-level shadow representations are concatenated to infer the scene layout of each instance and outpaint the unobserved areas. Finally, two discriminators are implemented to enhance alignment between the extended semantics and their shadows. Subsequently, extensive objective and subjective experiments are conducted which strongly proves the proposed approach successfully captures semantic and geometry information in the shadows. In future work, we will endeavor to learn shadow representations from dynamic environments and incorporate more tasks like novel view synthesis and pedestrian trajectory prediction into the unified framework.

Acknowledgement The research was supported by the Shanghai Committee of Science and Technology, China (Grant No.23ZR1423500).

References

[1] Abdelrehim Ahmed and Aly Farag. Shape from shading for hybrid surfaces. In *2007 IEEE International Conference on Image Processing*, pages II–525. IEEE, 2007. 2

[2] Abdelrehim H Ahmed and Aly A Farag. Shape from shading under various imaging conditions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[3] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022. 2

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6, 7, 8

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 2

[6] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023. 2

[7] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11431–11440, 2022. 1

[8] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9): 10850–10869, 2023. 2

[9] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[11] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021. 5, 6

[12] Aleksandrs Ecins, Cornelia Fermüller, and Yiannis Aloimonos. Shadow free segmentation in still images using local density measure. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2014. 2

[13] Jiacheng Fan, Min Chen, Jinqiu Mo, Shigang Wang, and Qinghua Liang. Variational formulation of a hybrid perspective shape from shading model. *The Visual Computer*, pages 1–14, 2022. 2

[14] Silvano Galliani, Yong Chul Ju, Michael Breuß, and Andrés Bruhn. Generalised perspective shape from shading in spherical coordinates. In *Scale Space and Variational Methods in Computer Vision: 4th International Conference, SSVM 2013, Schloss Seggau, Leibnitz, Austria, June 2-6, 2013. Proceedings 4*, pages 222–233. Springer, 2013. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[16] Yinsai Guo, Hang Yu, Liyan Ma, Liang Zeng, and Xiangfeng Luo. Thfe: A triple-hierarchy feature enhancement method for tiny boat detection. *Engineering Applications of Artificial Intelligence*, 123:106271, 2023. 2

[17] Yinsai Guo, Hang Yu, Shaorong Xie, Liyan Ma, Xinzhi Cao, and Xiangfeng Luo. Dsca: A dual semantic correlation alignment method for domain adaptation object detection. *Pattern Recognition*, page 110329, 2024. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 6

[19] Weiqiang Jin, Biao Zhao, Hang Yu, Xi Tao, Ruiping Yin, and Guizhong Liu. Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning. *Data Mining and Knowledge Discovery*, 37(1):255–288, 2023. 2

[20] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 6

[21] Imran N. Junejo and Hassan Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. In *Computer Vision – ECCV 2008*, pages 318–331, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. 2

[22] Asaf Karnieli, Ohad Fried, and Yacov Hel-Or. Deepshadow: Neural shape from shadow. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 415–430. Springer, 2022. 1, 2

[23] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011. 2

[24] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 2

[25] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4019–4028, 2021. 4

[26] Bholeshwar Khurana, Soumya Ranjan Dash, Abhishek Bhatia, Aniruddha Mahapatra, Hrituraj Singh, and Kuldeep Kulkarni. Semie: Semantically-aware image extrapolation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14880–14889, 2021. 1, 2

[27] Jihyun Kim, Seong-Hun Jeong, Kyeongbo Kong, and Suk-Ju Kang. An unified framework for language guided image completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2568–2578, 2023. 2

[28] Daehyeon Kong, Kyeongbo Kong, Kyunghun Kim, Sung-Jun Min, and Suk-Ju Kang. Image-adaptive hint generation

via vision transformer for outpainting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3572–3581, 2022. 2

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6, 7, 8

[30] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating natural illumination from a single outdoor image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 183–190, 2009. 2

[31] Kyoung Mu Lee and C-C Jay Kuo. Shape from shading with a generalized reflectance map model. *Computer vision and image understanding*, 67(2):143–160, 1997. 2

[32] Junxuan Li and Hongdong Li. Neural reflectance for shape recovery with shadow handling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16221–16230, 2022. 1, 2

[33] Pengbo Li, Hang Yu, Xiangfeng Luo, and Jia Wu. Lgm-gnn: A local and global aware memory-based graph neural network for fraud detection. *IEEE Transactions on Big Data*, 2023. 2

[34] Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *Advances in Neural Information Processing Systems*, 35:15420–15432, 2022. 2

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[36] Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, and Carl Vondrick. What you can reconstruct from a shadow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17059–17068, 2023. 1, 2

[37] Chia-Ni Lu, Ya-Chu Chang, and Wei-Chen Chiu. Bridging the visual gap: Wide-range image blending. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 843–851, 2021. 2

[38] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2, 6

[39] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1680–1691, 2023. 3

[40] Pascal Mamassian. *Shape from Shadows*, pages 724–725. Springer, 2014. 2

[41] S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1079–1087, 2004. 2

[42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[43] Takahiro Okabe, Imari Sato, and Yoichi Sato. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1693–1700. IEEE, 2009. 2

[44] Alexandros Panagopoulos, Dimitris Samaras, and Nikos Paragios. Robust shadow and illumination estimation using a mixture model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 651–658, 2009.

[45] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. Simultaneous cast shadows, illumination and geometry inference using hypergraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):437–449, 2013. 2

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[47] Jiayan Qiu, Xinchao Wang, Pascal Fua, and Dacheng Tao. Matching seqlets: An unsupervised approach for locality preserving sequence matching. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):745–752, 2019. 2

[48] Jiayan Qiu, Xinchao Wang, Stephen J Maybank, and Dacheng Tao. World from blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8493–8504, 2019. 2

[49] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. Hallucinating visual instances in total absentia. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 264–282. Springer, 2020. 2

[50] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. Scene essence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8322–8333, 2021. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[52] Mark Sabini and Gili Rusak. Painting outside the box: Image outpainting with gans. *arXiv preprint arXiv:1808.08483*, 2018. 2

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep

language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[54] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[56] Cheng-Yo Tan, Chiao-An Yang, Shang-Fu Chen, Meng-Lin Wu, and Yu-Chiang Frank Wang. Robust image outpainting with learnable image margins. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1159–1163, 2021. 2

[57] Cheng-Yo Tan, Chiao-An Yang, Shang-Fu Chen, Meng-Lin Wu, and Yu-Chiang Frank Wang. Robust image outpainting with learnable image margins. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1159–1163. IEEE, 2021. 1

[58] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019. 1

[59] Pinzhuo Tian and Hang Yu. Can we improve meta-learning model in few-shot learning by aligning data distributions? *Knowledge-Based Systems*, 277:110800, 2023. 2

[60] Kushagra Tiwary, Tzofi Klinghoffer, and Ramesh Raskar. Towards learning neural representations from shadows. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 300–316. Springer, 2022. 1

[61] Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022. 2

[62] Guohui Wang and Jin Cheng. Three-dimensional reconstruction of hybrid surfaces using perspective shape from shading. *Optik*, 127(19):7740–7751, 2016. 2

[63] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2022. 3, 5, 6

[64] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019. 2

[65] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs with diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[66] Chenyu Wu, Srinivasa G Narasimhan, and Branislav Jaramaz. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *International Journal of Computer Vision*, 86:211–228, 2010. 2

[67] Lin Wu and Xiaochun Cao. Geo-location estimation from two shadow trajectories. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 585–590, 2010. 2

[68] Xiaoyu Xu, Jiayan Qiu, Xinchao Wang, and Zhou Wang. Relationship spatialization for depth estimation. In *European Conference on Computer Vision*, pages 615–637. Springer, 2022. 2

[69] Chiao-An Yang, Cheng-Yo Tan, Wan-Cyuan Fan, Cheng-Fu Yang, Meng-Lin Wu, and Yu-Chiang Frank Wang. Scene graph expansion for semantics-guided image outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15626, 2022. 1, 2, 6, 7

[70] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 2

[71] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 2023. Just Accepted. 2

[72] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7074–7083, 2020. 2

[73] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10560–10569, 2019. 2

[74] Kai Yao, Penglei Gao, Xi Yang, Jie Sun, Rui Zhang, and Kaizhu Huang. Outpainting by queries. In *European Conference on Computer Vision*, pages 153–169. Springer, 2022. 1, 2

[75] Guangcong Zheng, Shengming Li, Hui Wang, Taiping Yao, Yang Chen, Shouhong Ding, and Xi Li. Entropy-driven sampling and training scheme for conditional diffusion generation. In *European Conference on Computer Vision*, pages 754–769. Springer, 2022. 2

[76] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 2, 5, 6

[77] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson WH Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4702–4711, 2021. 2, 6