Shadow-Enlightened Image Outpainting (Supplementary Material)

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

We provide in this document the details of the proposed approach, including the network structures, data selection, evaluation metrics, limitations, and societal impact. We also showcase more visual results on the VG [4] and COCO-stuff [2]. Then, to validate the generalizability of our approach, we show our extension on in-the-wild scene images.

As discussed in the manuscript, the main goal of our work is to show the potential outside the images by utilizing the shadow information, rather than beating the state-of-the-art image outpainting, shadow detection, and shadow removal methods. More sophisticated networks, as long as end-to-end trainable, can be readily applied to substitute the corresponding modules in our approach.

1. Network Structure

Here we give more information about the precise architectural details used to build the components of our model.

ResNet blocks. Our ConvNets are composed of ResNet blocks. The ResNet blocks used are the same as those used in [1] reproduced in Fig. 1. The block may be used to maintain the resolution of the features using an identity layer as opposed to the upsampling layer.

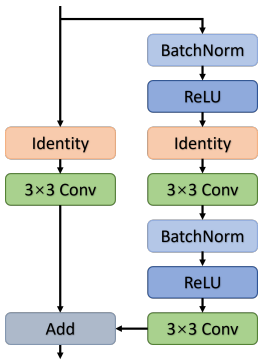

Figure 1. An overview of modified ResNet block replacing the upsample block with an identity block.

*Corresponding author

ConvNet module. ResNet blocks are stacked together to form the embedding network. In particular, we use the setup in Fig. 2. The shadow image with semantic information removed is passed through a ConvNet, resulting in a shadow feature map with a depth of 32.

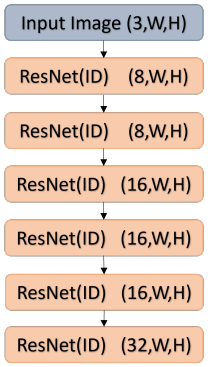

Figure 2. An overview of modified ResNet block replacing the upsample block with an identity block.

Image Outpainting. First, we resize the source image to a width and height of 256. Then, we horizontally randomly mask off half of the width of the image to create the input image. In the outpainting process, we use two encoders E_h and E_l to separately extract the corresponding features for scene layout expansion and layout-to-image conversion. The structure of these two encoders is shown in Tab. 1.

Type	Argument
Input	c=32,W=16,H=16
Conv+BN+ReLU	c=64,k=3,pad=1
Maxpool2d	c=64,k=2,stride=2
Conv+BN+ReLU	c=128,k=3,pad=1
Maxpool2d	c=128,k=2,stride=2
Flatten	c=2048
FC	c=256

Table 1. An overview of shadow encoders for scene layout expansion and layout-to-image conversion.

We modify the diffusion-based method for image conversion. Specifically, we perform a denoising process in the image space, using the same hyperparameters as described in [6, 8].

2. Data Seletion

Tab. 2 shows some statistical properties of the VG [5] and COCO-stuff [2] dataset. We can see that almost half of the proportion has invisible shadows accounting for 10%, and for every 10% increase in the area ratio, the proportion of the quantity decreases by half.

	train set				test set			
	total	>10%	>20%	>30%	total	>10%	>20%	>30%
VG [5]	62565	25310	12947	5825	5096	2640	1405	718
COCO-stuff [2]	25210	12413	7037	4507	3097	1671	843	512

Table 2. The number of images with different proportions of invisible shadows.

3. Metrics

In order to compare the accuracy of the model's predictions for novel / masked objects and relationships, we compare unmasked objects and relationships in the output expanded scene graph $\boldsymbol{S}^{op}$ with the ground truth scene graph $\boldsymbol{S}^{gt}$. We report the metrics of the averaged rank of correct prediction (rAVG) and the top-k accuracy (Hits@ k) for both object and relationship predictions, respectively. Note that we ignore the "empty" relationship in the masked relationships in $\boldsymbol{S}^{gt}$ for accuracy calculation due to the sparsity expected for scene graphs. Besides, we measure the mIoU between the output layout L_{in} and the ground truth L_{gt}. Finally, we use FID [3], which is shown to correlate well with human judgments, to measure similarity by comparing distributions of activations from an Inception network.

4. Limitations

First, our proposed approach can only produce enhancement in scenes with shadows. In no-shadow environments, we provide no complementary information for image outpainting, which is thus unhelpful. Second, our approach depends on the performance of the pretrained shadow processing models, such as shadow detection, instance shadow detection, and shadow removal model. The substandard performance of the shadow processing models may hamper the performance of the proposed approach.

5. Societal Impact

Our approach can extend the scene information, which may be adopted in some safety-related tasks, such as autonomous driving and video surveillance. It should be noted that the generated information may with large difference compared with the real information, which may lead to unexpected outcomes.

6. Additional Visual Results

We give additional qualitative results on VG [4] and COCO-stuff [2] shown in Fig. 3 and Fig. 4. As we can see, our method makes the extension more reasonable and better aligned with the existing shadows on both datasets.

7. Outpainting on In-The-Wild Scene Images

We show our extension on in-the-wild scene images shown in Fig. 5 to validate the generalizability of our approach. As we can see, our method makes the outpainting region more reasonable and better aligned with the existing shadows on in-the-wild scene.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1, 2, 4

[3] 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. *Advances in neural information processing systems*, 30, 2017. 2

[4] 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. 1, 2, 3

[5] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 2

[6] 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

[7] 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. 3, 4, 5

[8] 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

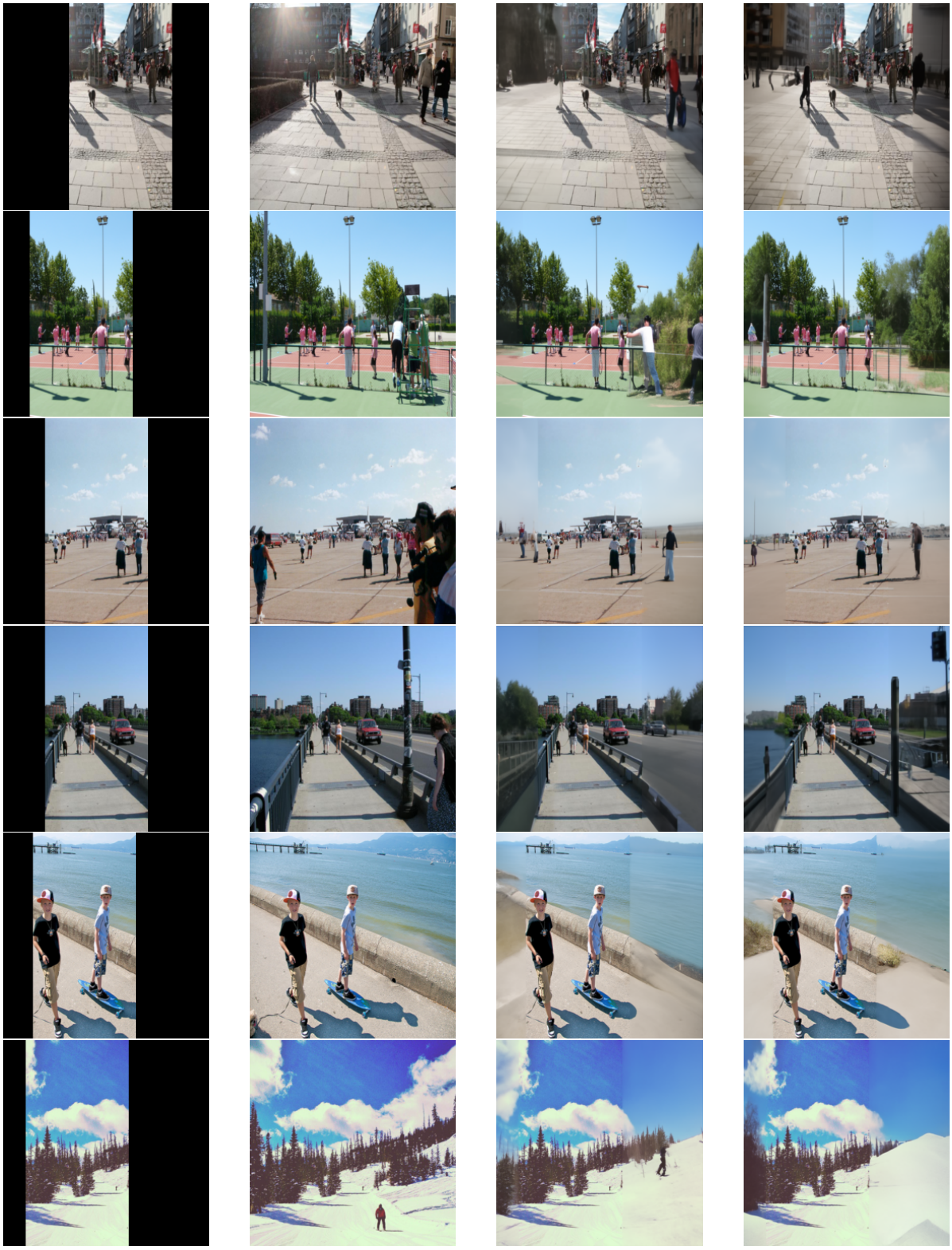

Figure 3. Visual results from SGT [7] on VG [4] dataset. The first image denotes the observed image, the second one denotes the ground-truth image, the third one denotes the outpainting result generated by SGT [7], and the last one denotes the outpainting result after adopting our approach as a plug-in module.

Figure 4. Visual results from SGT [7] on COCO-stuff [2] dataset. The first image denotes the observed image, the second one denotes the ground-truth image, the third one denotes the outpainting result generated by SGT [7], and the last one denotes the outpainting result after adopting our approach as a plug-in module.

Figure 5. Example outpainting results on in-the-wild scene images. The first image denotes the observed image, the second one denotes the ground-truth image, the third one denotes the outpainting result generated by SGT [7], and the last one denotes the outpainting result after adopting our approach.