

The Supplementary of Learning Object Placement via Convolution Scoring Attention

Yibin Wang
yibinwang1121@163.com

Yuchao Feng
fyc@zjut.edu.cn

Jianwei Zheng*
zjw@zjut.edu.cn

College of Computer Science and
Technology
Zhejiang University of Technology
Hangzhou, China

1 Affine Transformation

Given the prepared parameter \mathbf{t} , we follow Spatial Transformer Network [1] to achieve affine transformation T on foreground objects, which can be further used for the final composition. Note $\mathbf{t} = [t_r, t_x, t_y] \in \mathcal{R}^3$. t_r is defined as the scale ratio for the foreground object, whose span is within $(0,1)$. The height and width of the transformed foreground region are then $h = t_r H$ and $w = t_r W$. On that basis, we define $t_x = \frac{x}{W-w} \in (0,1)$ and $t_y = \frac{y}{H-h} \in (0,1)$ to represent the relative vertical and horizontal locations of the foreground placement over the background, where (x, y) denotes the background coordinate for the left top pixel point of the scaled foreground region.

After a simple derivation, the parameter Θ of T can be formulated as

$$\Theta(\mathbf{t}) = \begin{pmatrix} 1/t_r & 0 & (1-2t_x)(1/t_r-1) \\ 0 & 1/t_r & (1-2t_y)(1/t_r-1) \end{pmatrix}. \quad (1)$$

Subsequently, the transformed image $\mathbf{I}_{fg}^t = T(\mathbf{I}_{fg}; \Theta)$ with the counterpart mask $\mathbf{M}_{fg}^t = T(\mathbf{M}_{fg}; \Theta)$ can be obtained.

In the end, the composite image \mathbf{I}_c is generated as

$$\mathbf{I}_c = \mathbf{M}_{fg}^t * \mathbf{I}_{fg}^t + (1 - \mathbf{M}_{fg}^t) * \mathbf{I}_{bg}. \quad (2)$$

2 The Functionality of the Pyramid Pooling

The efficacy of the pyramid pooling method in our work lies in its ability to capture essential multi-scale contextual information, a crucial aspect for accurately placing objects within their proper scale. Considering the significance of the relative size relationship between objects and the overall scene, understanding both foreground and background scales becomes

*Corresponding author.

imperative. This method allows our model to skillfully aggregate information across diverse scales, facilitating the exploration of reasonable object scales within a background scene.

3 The Competing Methods

Since OP is still an emerging topic, to the best of our knowledge, only four latest benchmarks are closely related to our work, including TERSE [1], PlaceNet [2], GracoNet [3], and CA-GAN [4].

TERSE (CVPR, 2019)[1]: With a shared backbone, TERSE models heterogeneous features for foreground and background via two separate branches, whose outputs are then concatenated and regressed to predict the transformation parameters.

PlaceNet (ECCV, 2020)[2]: The core of PlaceNet lies in two independent encoders extracting features respectively from foreground and background. Then, the extracted features are combined with different random vectors, thereby outputting various object placements via a shared decoder network.

GracoNet (ECCV, 2022)[3]: GracoNet treats object placement task as a graph completion problem, which considers the background as multiple nodes with different locations and the foreground as a unique node. In addition, the cross-attention module is employed between nodes of background and foreground. Finally, random vectors are also introduced to enhance its diversity.

CA-GAN (ICME, 2023)[4]: CA-GAN proposes a coalescing attention module to extract salient feature interaction between background and foreground. Assisted by a purely VAE-based supervised path, it achieves state-of-the-art performance compared with other models. Analogously, random vectors are used for better diversity.

The source codes of the competing methods are publicly available. For our CSANet, the model is trained on a single RTX 3090 GPU with batch size 32 and epoch 18. The discriminator in our work follows a similar architecture to [5].

4 Evaluation Metrics

User Study: 20 voluntary participants are invited to subjectively estimate the placement results of all methods. Concretely, all methods generate composite images given the same pairs of foreground and background, which are then delivered to the participants for the decisions of compositing quality. The score of this metric for each method is obtained using the average result on all test samples.

FID: Fréchet Inception Distance (FID) is born for measuring the similarity between two group images, which is the most common measure for the performance evaluation of generative adversarial networks. Analogously, we generate FID between one set of ground-truth composite images of the OPA test set and the other set of composite images produced by the competing methods.

LPIPS: LPIPS is a measure of perceptual similarity between two images. During the inference stage, we first generate 10 composite images by sampling random vectors 10 times given a pair of foreground and background. Afterwards, we pair-wisely compute LPIPS from the 10 composite images. Then, the average LPIPS score on all test samples can be generated for the final evaluation. On account of LPIPS exposing the difference between two images, a larger LPIPS score demonstrates a better generation diversity.

Table 1: Ablation study on loss functions

Methods	<i>Credibility</i>	<i>Diversity</i>
	FID ↓	LPIPS ↑
$L_{adv}^u + L_{adv}^s$	39.31	0.057
$L_{adv}^u + L_{adv}^s + L_{bce}$	29.31	0.073
$L_{adv}^u + L_{adv}^s + L_{bce} + L_{kld}$	23.69	0.238
$L_{adv}^u + L_{adv}^s + L_{bce} + L_{kld} + L_{rec}$	20.88	0.274

Accuracy: We extend the SimOPA [2] model to check the accuracy of object placement generation results. The extended model functions as a binary classifier that distinguishes between reasonable and unreasonable object placements. We define accuracy as the proportion of the generated composite images that are classified as positive by the binary classifier during inference.

5 Ablation Study

5.1 Different Loss Functions

Our loss function consists of several subassemblies, all of which collaboratively assist in the training of our generator. By removing certain subassemblies from the overall loss, the ablation results are given in Table 1. As can be seen, with the participation of L_{bce} , the diversity and credibility both increase sharply, demonstrating its effectiveness in facilitating a better discriminator. The addition of L_{kld} and L_{rec} further brings considerable improvements, manifesting their effectiveness in approximating the generated images with more details to ground truth. Clearly, the joint collaboration of all the involved losses reveals the best result, ensuring the superior performance of our proposal.

5.2 Degree of Supervision

Empirically, involving supervision mechanisms in originally unsupervised models often leads to better performance. To validate this, the results caused by different degrees of supervision are ablated in Table 2. Clearly, with the supervised path removed, the performance confronts an evident decrease, which can be attributed to the “model collapse” issue of generative adversarial network. This issue can be sharply ameliorated with the addition of binary classification loss L_{bce} . We consider the reason may lie in its power in enabling the discriminator to latch the view of deeper data distribution so that it can be more distinguishable for images generated by the generator. Moreover, a further performance promotion arises with the aid of the complete supervised path, in which PCSSA and VAE collaboratively strive for prior property extraction. Astonishingly, not only the model collapse issue has been addressed, but also a great balance of credibility and diversity is obtained. Therefore, the involved supervised path is assumed to be contributive in guiding the generator to learn more reasonable and diverse object placements.

Table 2: Ablation study on degree of supervision

Supervision Degree	<i>Credibility</i>	<i>Diversity</i>
	FID ↓	LPIPS ↑
P_u	43.52	0.025
$P_u + L_{bce}$	30.79	0.143
$P_u + P_s$	20.88	0.274

Table 3: Choices of hyper-parameter C_p and kernel size k

C_p	<i>Credibility</i>	<i>Diversity</i>	k	<i>Credibility</i>	<i>Diversity</i>
	FID ↓	LPIPS ↑		FID ↓	LPIPS ↑
256	21.29	0.257	7×7	24.62	0.221
1024	23.78	0.272	9×9	21.03	0.268
2048	27.64	0.279	13×13	23.44	0.254
512	20.88	0.274	11×11	20.88	0.274

6 Analysis of Hyper-parameters

6.1 Dimension C_p of Prior/Random Vectors

The imposed prior and random vectors on the generator play a positive role in diversity. Empirically, a larger vector dimension C_p provides the higher possibility that leads to better diversity, yet it would inevitably consume more hardware resources. To achieve a balance, Table 3 lists the results from different selections of C_p in span [256, 2048]. As can be seen, the values of LPIPS present an increasing trend consistently along with a larger C_p . However, when C_p increases, FID initially increases, reaches a peak at 512, and then drops. By considering both the credibility and diversity, we select $C_p=512$ in our experiments.

6.2 Kernel size k in DConv

Recent work [9] shows that it brings no performance gain but computational burden when employing standard depthwise convolutions with kernel size larger than 9×9 . Interestingly, CSANet benefits more from the convolutions with larger kernels, i.e., 11×11 . we analyzes different choices of k in the range of [7×7 , 11×11] as shown in Table 3. When k increases, FID and LPIPS also increase and reach a peak at 11×11 , and then decrease. Accordingly, we choose $k = 11$ in our implementation.

7 Limitation

Until recently, the mentioned dataset in our paper is the only released benchmark to evaluate the object placement task. In our future work, we are committed to exploring and potentially constructing additional datasets to further enhance the evaluation. Besides, another limitation in our current object placement model lies in its sensitivity to lighting conditions. The model may struggle to consistently produce visually coherent results when foreground objects are placed in scenes with varying lighting intensities or directions. This limitation arises due to the absence of a dedicated lighting adaptation mechanism, which could allow

the model to adjust the appearance of foreground objects to match the lighting nuances of the background. Incorporating a module that explicitly accounts for and adapts to different lighting scenarios could substantially enhance the model’s robustness and its ability to seamlessly integrate foreground objects into diverse environments. Moreover, the current model may encounter difficulties when tasked with placing objects in scenes with irregular perspectives or complex spatial configurations. The absence of a perspective-aware module hampers the model’s capacity to accurately consider depth and spatial relationships, leading to potential distortions in the final composite images. Addressing this limitation by incorporating a module that comprehensively understands and adjusts for varying perspectives would be instrumental in achieving more realistic and visually convincing object placements.

8 Societal Impacts

This task has the potential to revolutionize several industries, leading to tangible societal changes. In the realm of advertising and design, for instance, graphic designers can leverage this technology to seamlessly integrate products into visual compositions, saving time and allowing for more creative exploration in marketing campaigns. In film and video production, filmmakers can automate the placement of characters or objects within scenes, streamlining the editing process and enhancing overall visual appeal. Moreover, in manufacturing, the precise positioning of components during assembly processes can be automated, leading to increased production efficiency and a reduction in errors. This has direct implications for industries such as automotive manufacturing, where the proper placement of intricate parts is critical to the functioning of the final product. In the context of virtual and augmented reality, automatic foreground object placement can significantly enhance user experiences. Virtual tourism applications can automatically position avatars or objects within virtual environments, creating more realistic and immersive simulations. This, however, raises questions about the authenticity of virtual experiences and the ethical considerations surrounding the use of such technology in shaping digital realities.

9 Additional Qualitative Comparison Results

More visual results are shown in Figs. 1 and 2 to further manifest the superiority of our model in different aspects.

Figs. 1 and 2 provide a direct contrast in terms of the generation credibility. Two cases, i.e., placing the same foreground object over different background scenes and placing various foreground objects over an identical background scene, are respectively considered. It is evident that CSANet shows outstanding robustness in placing specified foregrounds over different background scenes. With the more credible locations and sizes, the performance for positing various foreground objects over the given background scene is also better than other competitors. We attribute this to the beneficial cooperation between PCSUA and PCSSA. On that basis, the generator is empowered to take multi-scale features and interactive information into account, which accompanied by prior guidance leads to more reasonable predictions.

Besides, we also provide Fig. 3 to reveal the generation diversity of several competing methods. Note TERSE is omitted in this experiment since it is unable to produce multiple placements given the same background-foreground pair. Again, provided the same environ-



Figure 1: Visualization of object placement with the same foreground object and different background scenes. The placed foreground objects are highlighted by the red outline.

ment, our proposal shows more possible placements with lossless credibility. For example, the sailboat predicted by our proposal is posited in various directions surrounding the surfing guy, while the placements of all other methods are relatively restricted.



Figure 2: Visualization of object placement with the same background scene and different foreground objects. The placed foreground objects are highlighted by the red outline.

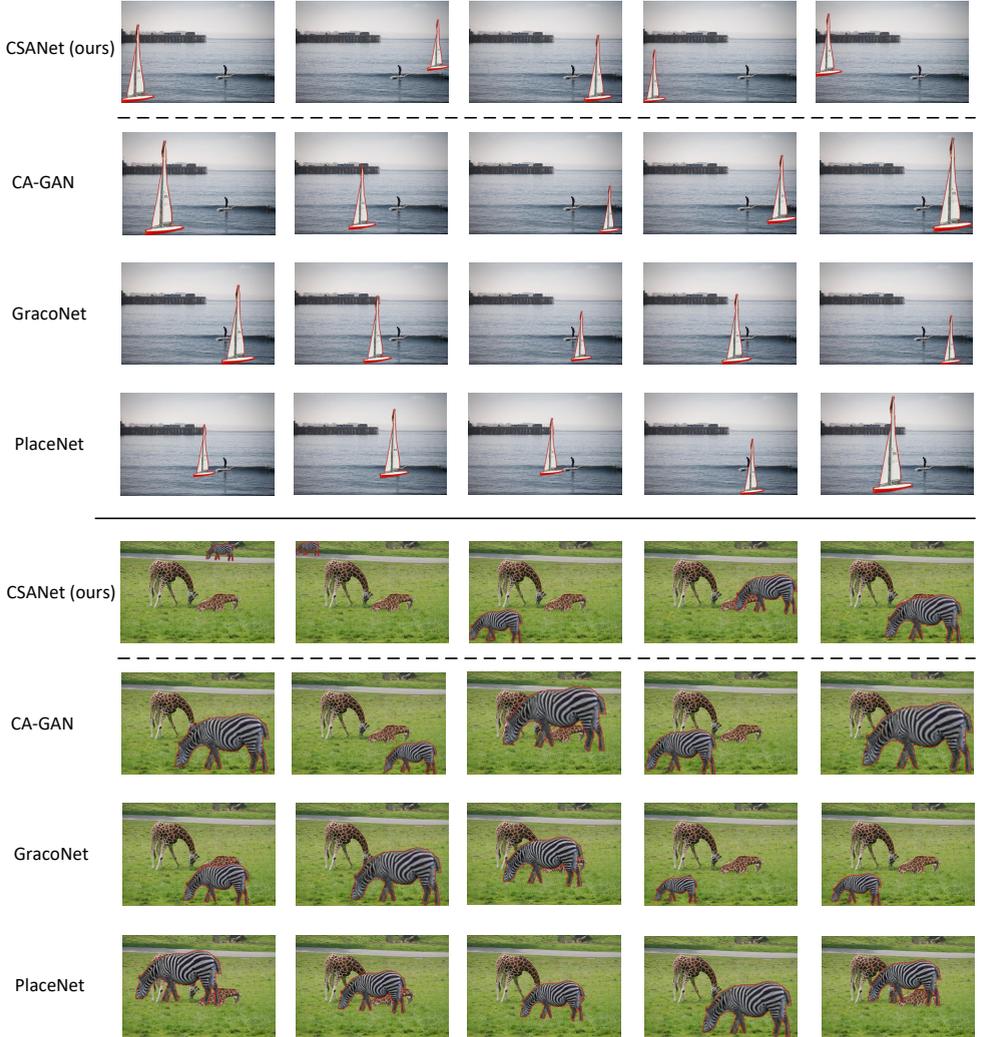


Figure 3: Comparisons of placement diversity by sampling different random vectors.

References

- [1] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Neural Information Processing Systems (NerulIPS)*, pages 28–34, 2015.
- [2] Liu Liu, Zhenchen Liu, Bo Zhang, Jiangtong Li, Li Niu, Qingyang Liu, and Liqing Zhang. Opa: object placement assessment dataset. *arXiv preprint arXiv:2107.01889*, 2021.
- [3] Zhuang Liu, Hanzi Mao, Chaoyuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.
- [4] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Amrith Tyagi, James M Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 461–470, 2019.
- [5] Tingchun Wang, Mingyu Liu, Junyan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018.
- [6] Yibin Wang, Yuchao Feng, Jie Wu, Honghui Xu, and Jianwei Zheng. Ca-gan: object placement via coalescing attention based generative adversarial network. In *International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2023.
- [7] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *European Conference on Computer Vision (ECCV)*, pages 566–581, 2020.
- [8] Siyuan Zhou, Liu Liu, Li Niu, and Liqing Zhang. Learning object placement via dual-path graph completion. In *European Conference on Computer Vision (ECCV)*, pages 373–389, 2022.