TopNet: Transformer-based Object Placement Network for Image Compositing

Sijie Zhu[1], Zhe Lin[2], Scott Cohen[2], Jason Kuen[2], Zhifei Zhang[2], Chen Chen[1]
[1]Center for Research in Computer Vision, University of Central Florida [2]Adobe Research
sizhu@knights.ucf.edu,{zlin,scohen,kuen,zzhang}@adobe.com,chen.chen@crcv.ucf.edu

A. Overview

In this supplementary material, we provide the following content for better understanding of the paper:

B. Broader Impact.

C. Performance in terms of Top-1 IOU.

D. Qualitative Comparison on Real-world Compositing.

E. Qualitative Results on Inpainted Pixabay.

F. Qualitative Results on OPA Dataset.

G. User Study Instruction.

H. Heatmap Comparison w/ Gaussian Assignment Loss.

I. Implementation Details.

J. Failure Case.

K. Diversity of Prediction.

B. Broader Impact

The proposed TopNet shows great potential for future image creation processes with AI assistance, which would be of great interest to researchers in AI applications and the vision community. It could help build more advanced image editing and creation software for better social sharing, advertising, and education.

The authors foresee two potential negative impacts of the proposed method. 1) It might have a bias on certain objects and background combinations, *e.g.* low performance on minority scenes or human categories. This issue could be addressed by using more robust machine learning techniques or better-balanced training data. 2) Although the images in our experiments do not have identity or copyright issues, the trained model might be used for illegal applications like deep fake, *e.g.* generating fake images for public identities. We could address this issue by using certain licenses for our model (TopNet), *i.e.* only allowing the usage of TopNet in legal and ethical applications. We may also train a model to recognize public identities and only allow objects without identity to be processed by our model.

C. Performance in Terms of Top-1 IOU

In Table 1, we show the results in terms of top-1 IOU, which is the IOU between the top-1 predicted bounding box and the ground-truth bounding box. The proposed method significantly outperforms previous methods on both datasets.

D. Qualitative Comparison on Real-world Compositing

Figs. 1 and 2 show comparisons between the proposed method and previous methods on real-world object placement with diverse scenes and object categories. The proposed method generalizes better as compared with previous methods.

E. Qualitative Results on Inpainted Pixabay

Fig. 3 shows the compositing results of the proposed method on the inpainted images as compared with the original images. All the images are from our inpainted Pixabay [1] dataset. We show the top predicted bounding boxes using our local maximum search (usually there are less than 5 boxes). The final compositing is based on the top-1 bounding box.

F. Qualitative Results on OPA

Fig. 4 shows the compositing results of the proposed method on OPA as compared with the annotated positive images. All the images are from the OPA [3] dataset. We show the top predicted bounding boxes using our local maximum search (usually there are less than 5 boxes). The final compositing is based on the top-1 bounding box. In Fig. 5, we also show an example of the predicted heatmaps for 16 scales on OPA. The heatmaps highlight potential candidate locations for different scales.

G. User Study Instruction

Each user is asked to rate each sample on three levels: 0) "Unsatisfactory": The location and scale are clearly wrong.

Method	Infer. Time (s)	Pixabay		OPA	
		$IOU > 0.5$	Mean IOU	$IOU > 0.5$	Mean IOU
Regression [5]	0.08	48.23	0.448	7.24	0.178
†Retrieval [6]	1.69	9.47	0.204	1.88	0.106
Classifier [3]	0.55	6.23	0.145	2.20	0.109
PlaceNet [5]	0.16	4.91	0.149	2.76	0.116
Ours	0.11	**60.70**	**0.506**	**11.55**	**0.197**

Table 1. Evaluation on top-1 IOU, *i.e*. the IOU between the top-1 predicted bounding box and the ground-truth box.

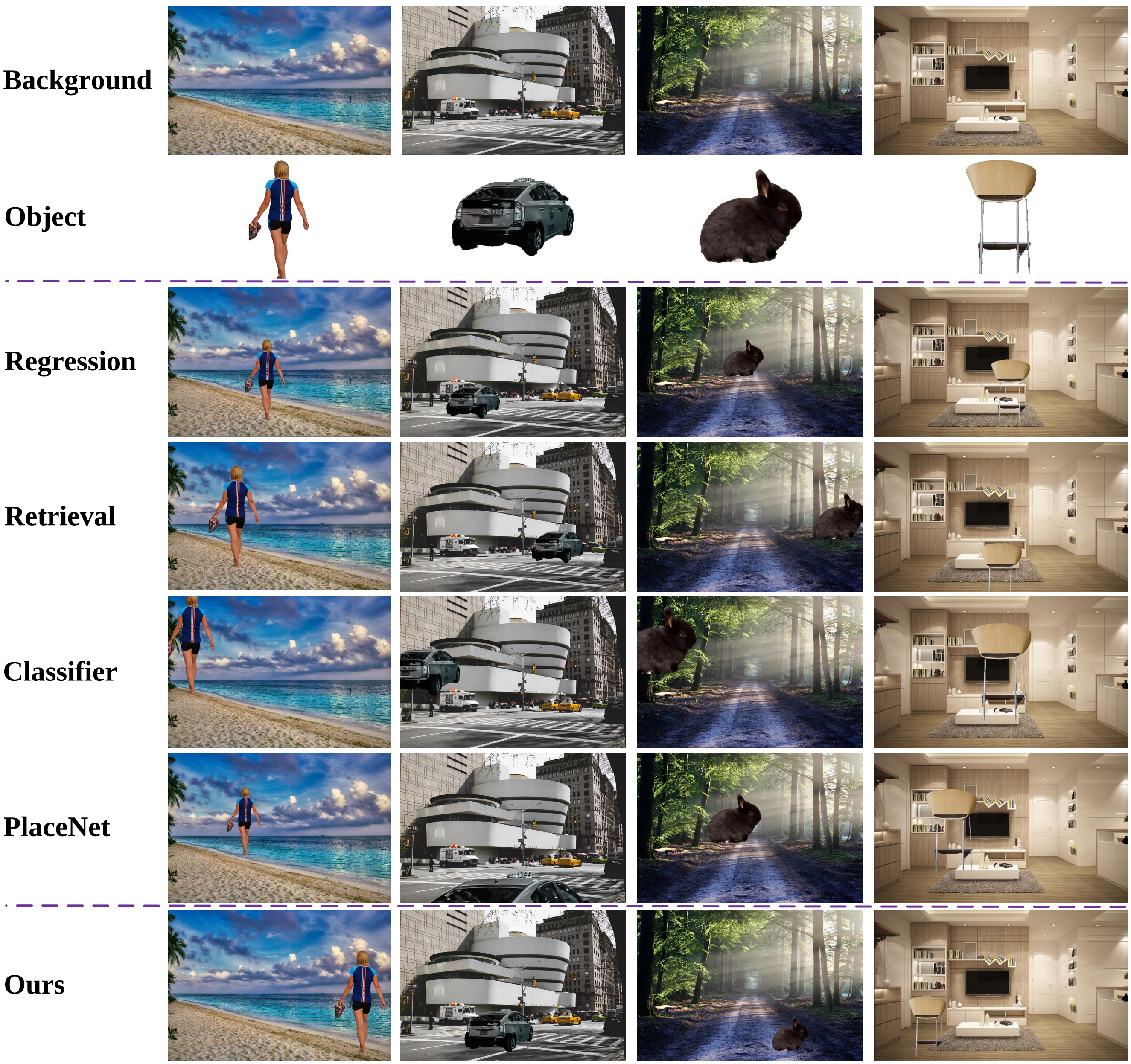

Figure 1. Qualitative comparison with previous methods on object placement for compositing.

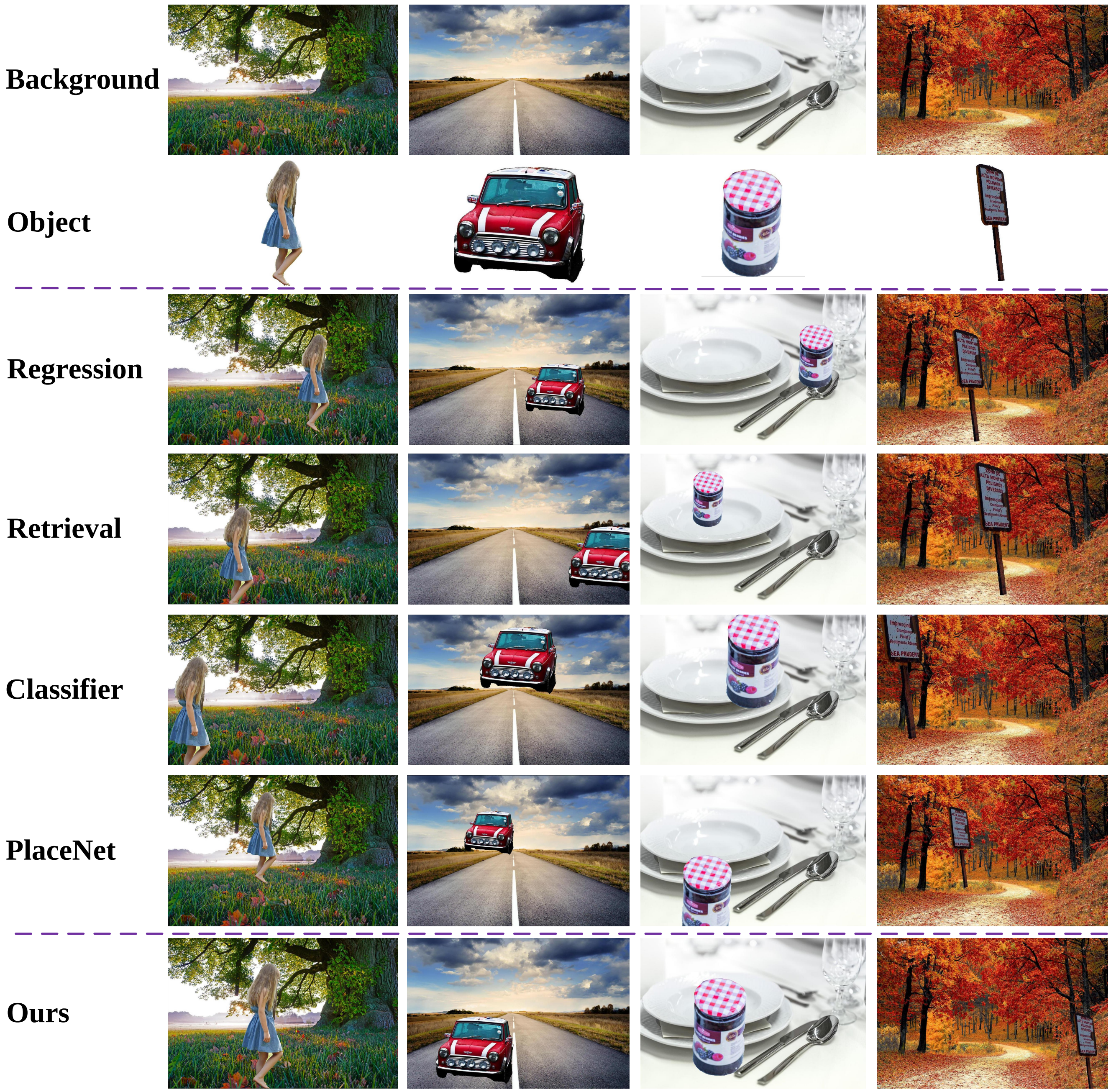

Figure 2. Qualitative comparison with previous methods on object placement for compositing.

1) "Borderline": The location and scale are OK but somewhat unrealistic. 2) The location and scale are clearly reasonable. One example interface is shown in Fig. 6.

H. Heatmap Comparison with Gaussian Assignment Loss

Fig. 7 shows the comparison between the heatmaps of Gaussian assignment loss (denoted as "Gaussian") and the proposed loss. As expected, Gaussian assignment loss generates a single-peak distribution on all dimensions, while the proposed sparse contrastive loss recommends multiple possible candidate placements with a multi-peak 3D heatmap. Gaussian assignment loss suppresses all locations/scales that are far away from ground truth, and such supervision could be unreasonable when multiple good locations/scales exist. The heatmaps of "Gaussian" in Fig. 7 highlight only the land regions, which is not reasonable for

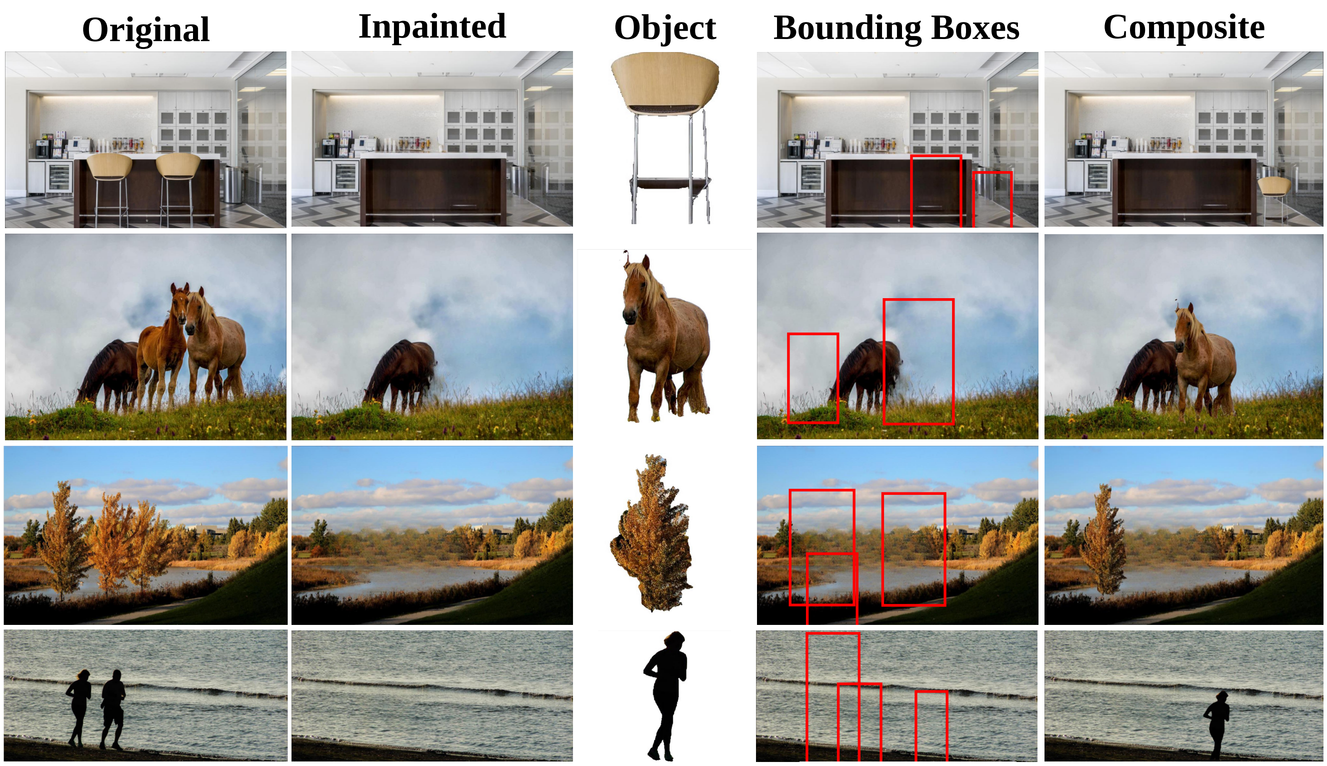

Figure 3. Qualitative result of the proposed method on inpainted Pixabay.

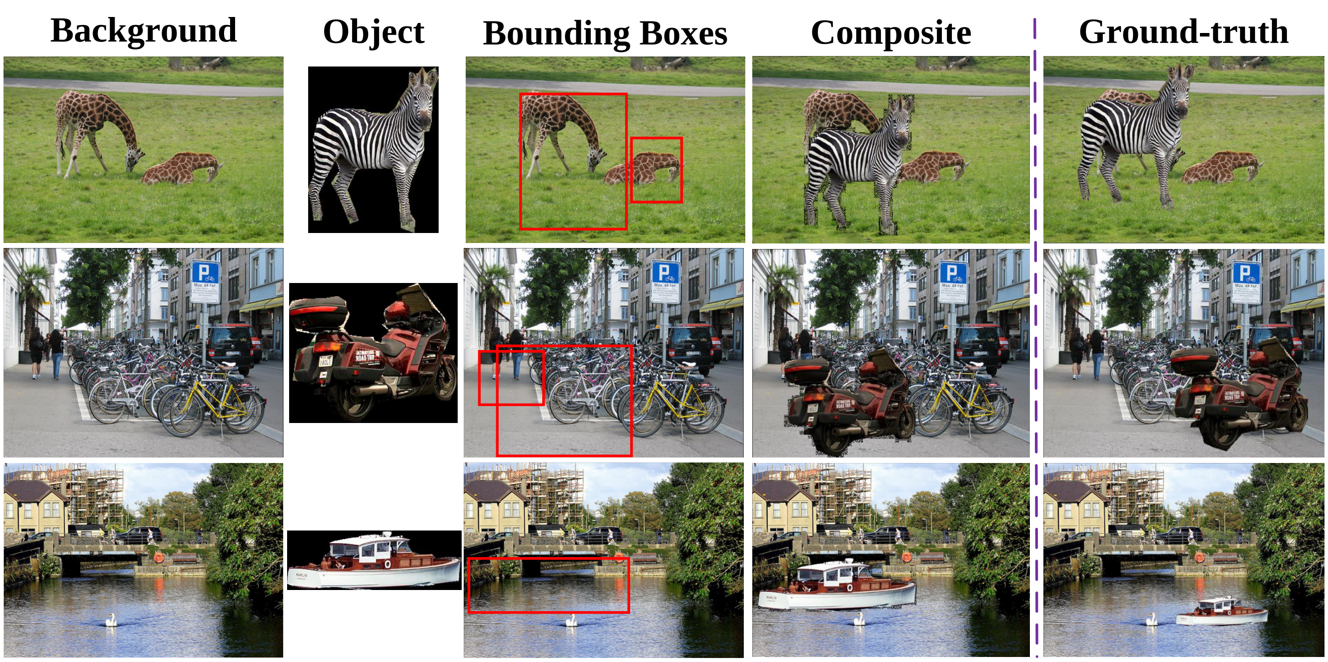

Figure 4. Qualitative result of the proposed method on the OPA dataset.

the boat object. Our method highlights multiple suitable regions with different scales, and the boat is mostly placed in the water.

I. Implementation Details

Local Maximum. To find the local maximum, we first filter out locations/scales with low scores using a threshold.

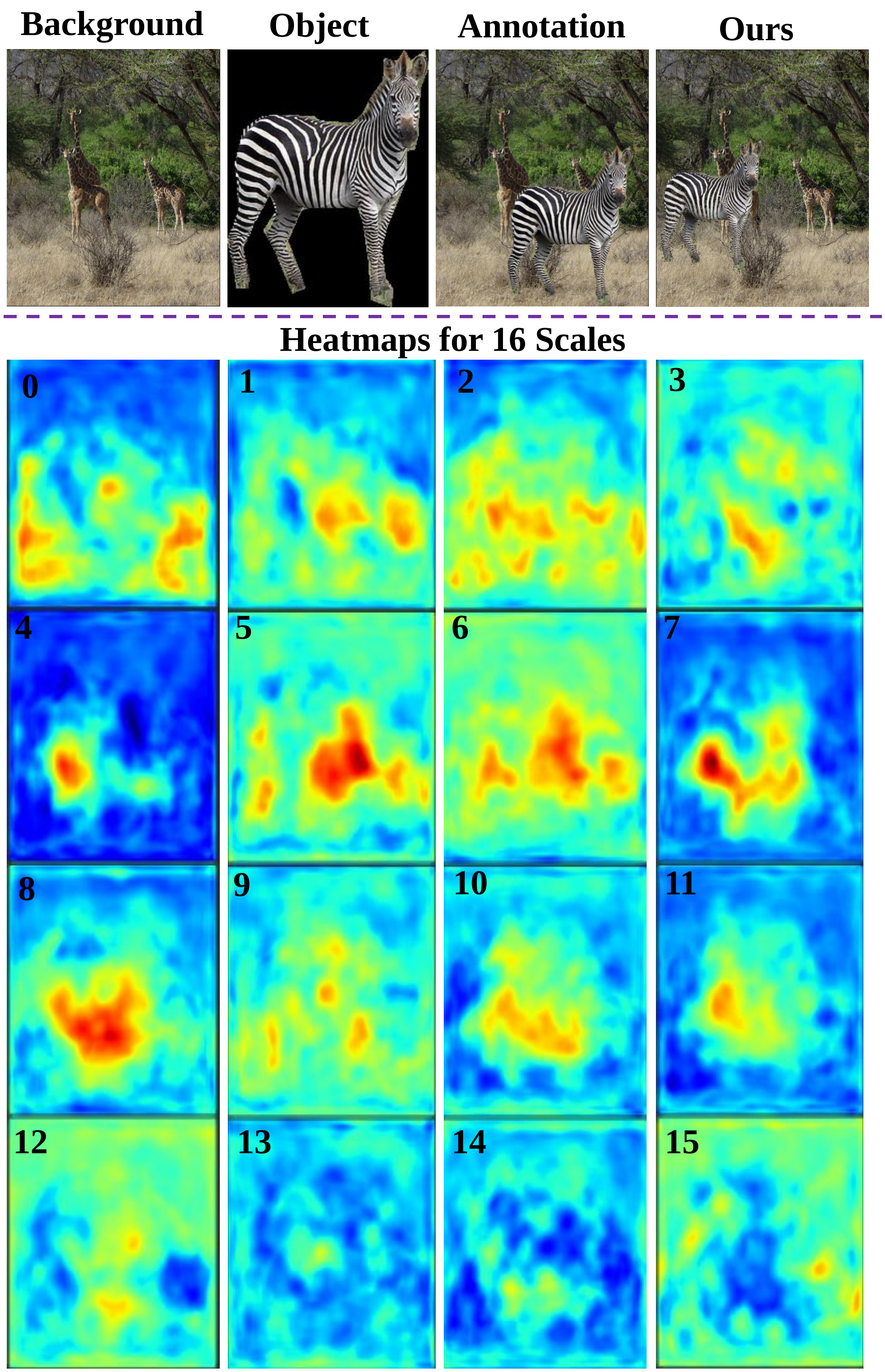

Figure 5. An example of the predicted heatmaps for all scales on the OPA dataset.

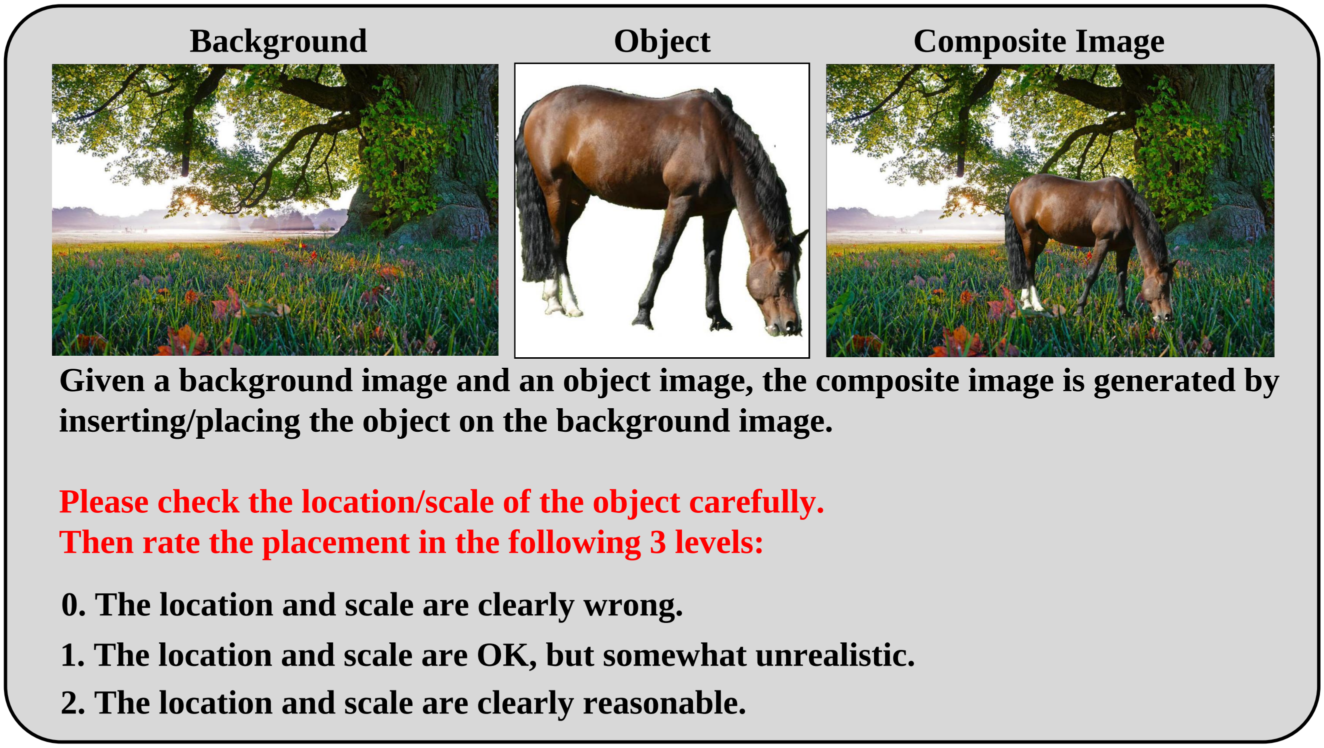

Figure 6. An example of user study interface.

It is computed as $th = mean(\hat{H}) + 2 * std(\hat{H})$, here $\hat{H}$ is the normalized heatmap defined in Sec. 3.1. Then we find the maximum point of each connected region as the local maximum. If there are more than 5 local peaks, we only keep the top-5 peaks with the highest scores. Otherwise, we keep all the local maximums as candidate bounding boxes. We apply the same procedure when computing top-5 boxes for the sliding-window methods, *i.e.* "†Retrieval" and "Classifier". "Regression" only generates one bounding box. "PlaceNet" [5] generates top-5 bounding boxes with 5 network forward passes. The transformer layers have a dimension of 384 and 16 heads. The decoder contains 4 transformer layers.

Previous Methods. The previous methods [3,5,6] do not provide the code, so we implement them by following the description in their paper. We use ResNet50 as backbone for "PlaceNet" [5] and "Regression". The prediction head contains 3 fully connected layers, along with batch normalization and ReLU [2] activation. "Classifier" uses the GRB image and the composite mask as input using ResNet18 [2] following [3]. The "†Retrieval" [6] on Pixabay is obtained from the authors, and we follow the same architecture (two-branch VGG-19 [4]) and loss to train it on OPA [3].

J. Failure Case

We provide a failure case example here for analysis. As shown in Fig. 8, our method could fail when the scene is complex with dense objects that overlap with multiple backgrounds, *e.g.* water, branches, and mountains. To tackle this example, the model needs to understand the object and the tree branches correctly. Although our model generates three candidate placements and two of them are close to the branches, the location and scale are still not accurate enough to have a realistic composting.

K. Diversity of Prediction

In Table 2, we compute the variance of predicted candidate bounding boxes (normalized with height/width to $[0, 1]$) as diversity on Pixabay dataset. Three methods have similar diversity, but the diversity of our method is slightly higher.

	Retrieval [29]	PlaceNet	Ours
Diversity	0.132	0.123	**0.154**

Table 2. Candidate placements diversity on Pixabay.

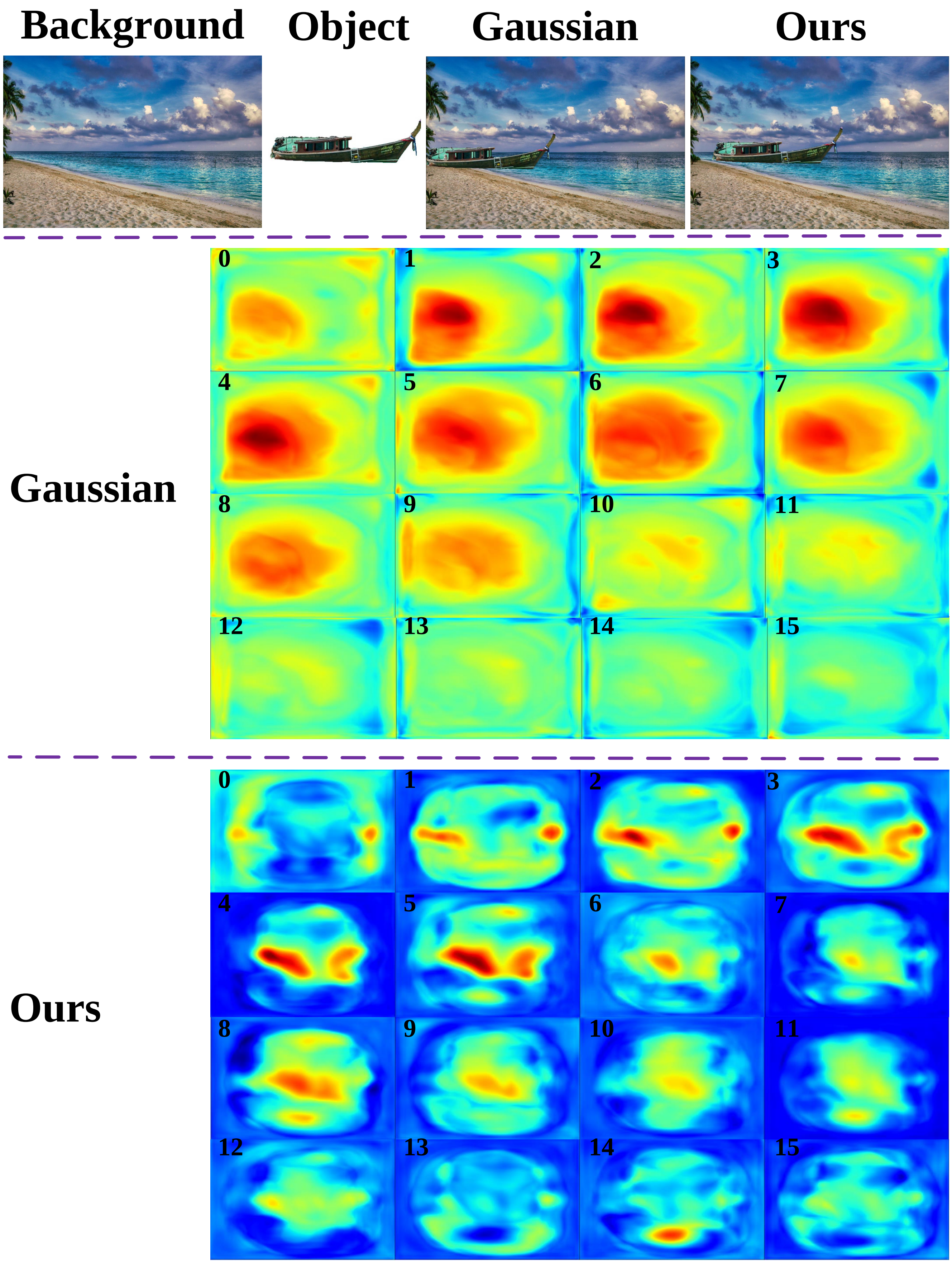

Figure 7. Heatmap comparison between the Gaussian assignment loss and the proposed sparse contrastive loss.

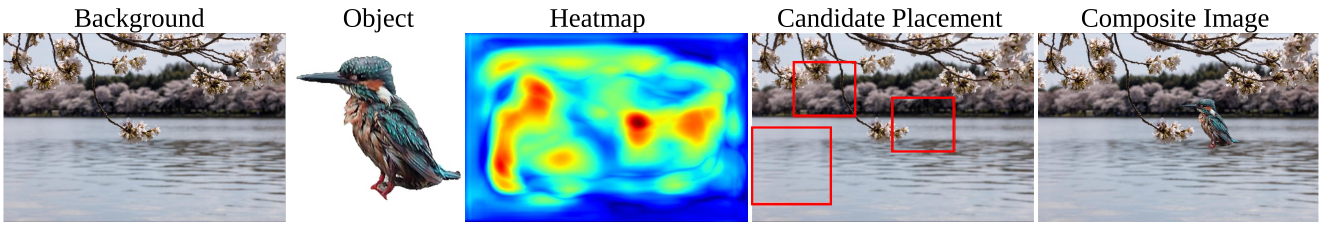

Figure 8. A failure case of our method. View with zoom-in.

References

[1] *https://pixabay.com/*. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[3] Liu Liu, Bo Zhang, Jiangtong Li, Li Niu, Qingyang Liu, and Liqing Zhang. Opa: Object placement assessment dataset. *arXiv preprint arXiv:2107.01889*, 2021. 1, 2, 6

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[5] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 566–581. Springer, 2020. 2, 6

[6] Sijie Zhu, Zhe Lin, Scott Cohen, Jason Kuen, Zhifei Zhang, and Chen Chen. Gala: Toward geometry-and-lighting-aware object search for compositing. *arXiv preprint arXiv:2204.00125*, 2022. 2, 6