TopNet: Transformer-based Object Placement Network for Image Compositing

Sijie Zhu, Zhe Lin, Scott Cohen, Jason Kuen, Zhifei Zhang, Chen Chen
1Center for Research in Computer Vision, University of Central Florida 2Adobe 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
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.
<table>
<thead>
<tr>
<th>Method</th>
<th>Infer. Time (s)</th>
<th>Pixabay</th>
<th>OPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression [5]</td>
<td>0.08</td>
<td>48.23</td>
<td>7.24</td>
</tr>
<tr>
<td>†Retrieval [6]</td>
<td>1.69</td>
<td>9.47</td>
<td>1.88</td>
</tr>
<tr>
<td>Classifier [3]</td>
<td>0.55</td>
<td>6.23</td>
<td>2.20</td>
</tr>
<tr>
<td>PlaceNet [5]</td>
<td>0.16</td>
<td>4.91</td>
<td>2.76</td>
</tr>
<tr>
<td>Ours</td>
<td>0.11</td>
<td>60.70</td>
<td>11.55</td>
</tr>
</tbody>
</table>

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.
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
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.
Given a background image and an object image, the composite image is generated by inserting/placing the object on the background image.

Please check the location/scale of the object carefully. Then rate the placement in the following 3 levels:

0. The location and scale are clearly wrong.
1. The location and scale are OK, but somewhat unrealistic.
2. The location and scale are clearly reasonable.

Figure 6. An example of user study interface.

It is computed as \( th = \text{mean}(\hat{H}) + 2 \times \text{std}(\hat{H}) \), 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. “\( \hat{\text{Retrieval}} \)” 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 “\( \hat{\text{Retrieval}} \)” [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.

<table>
<thead>
<tr>
<th></th>
<th>Retrieval [29]</th>
<th>PlaceNet</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity</td>
<td>0.132</td>
<td>0.123</td>
<td>0.154</td>
</tr>
</tbody>
</table>

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