UCF Center for Research in Computer Vision

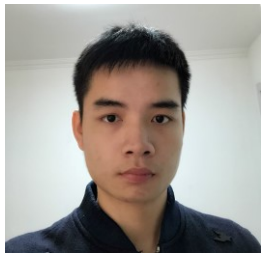
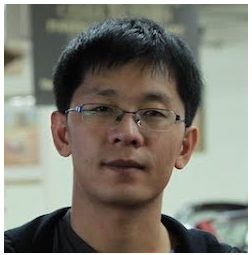
Adobe

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

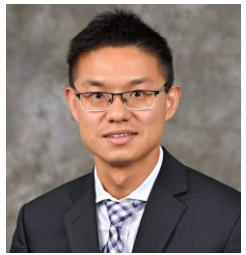
TUE-AM-175

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



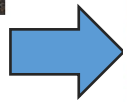


UCF

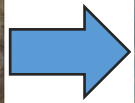
Object Placement for Compositing

Background Image

Foreground Object



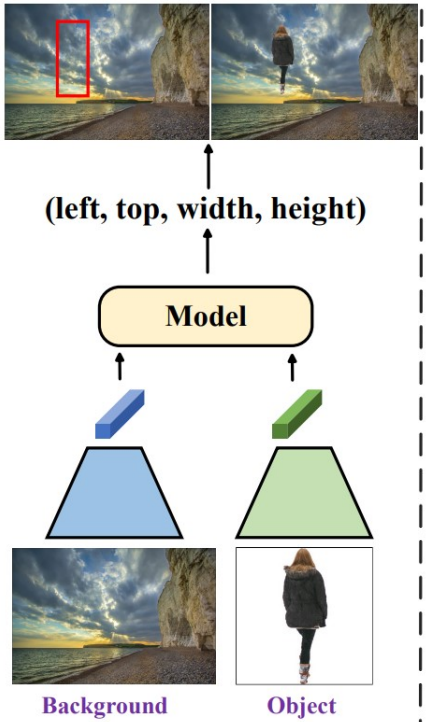
Location and Scale



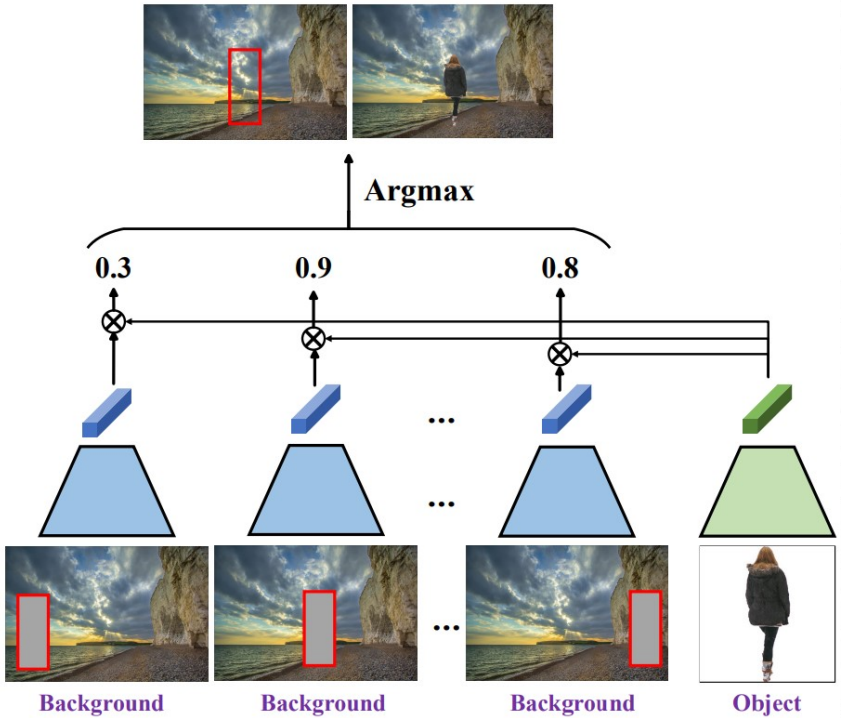
Composite Image

UCF

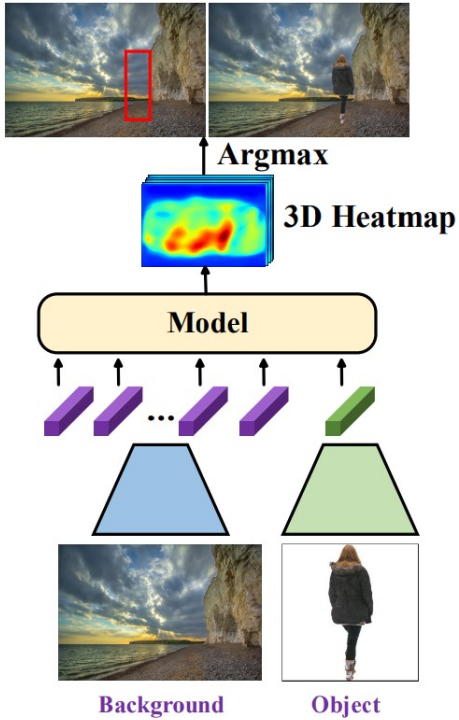
TopNet vs Previous Works



Direct Prediction [26]
(Sparse Evaluation, Fast)

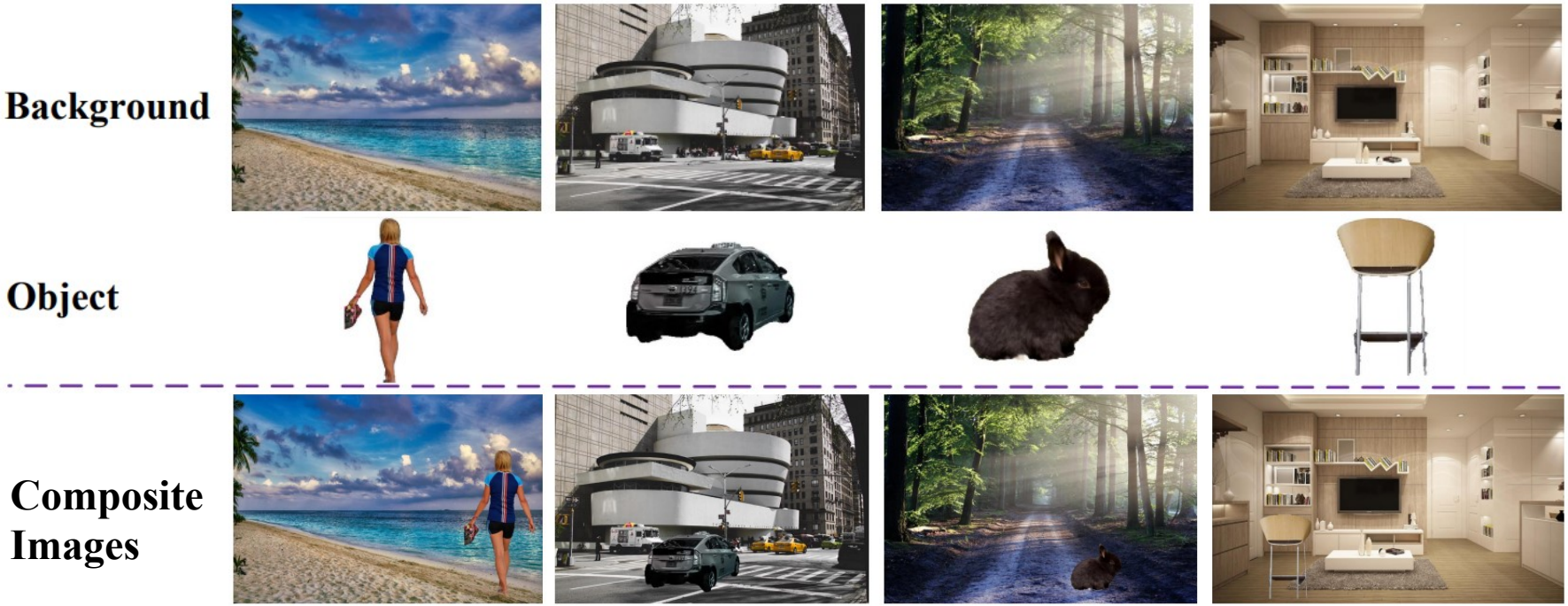


Sliding-Window [29]
(Dense Evaluation, Slow)

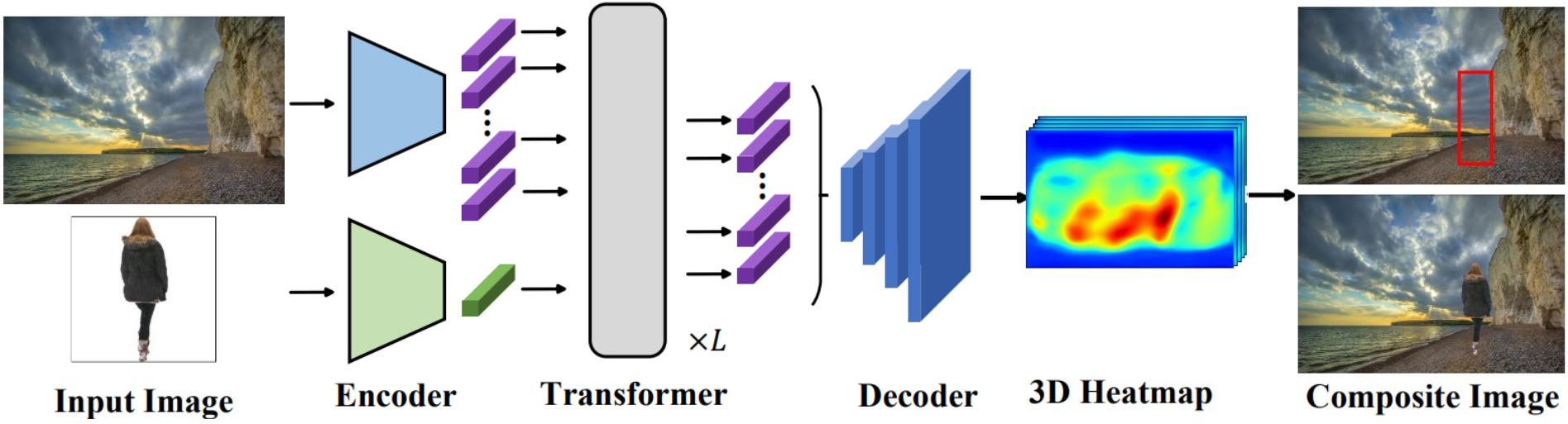


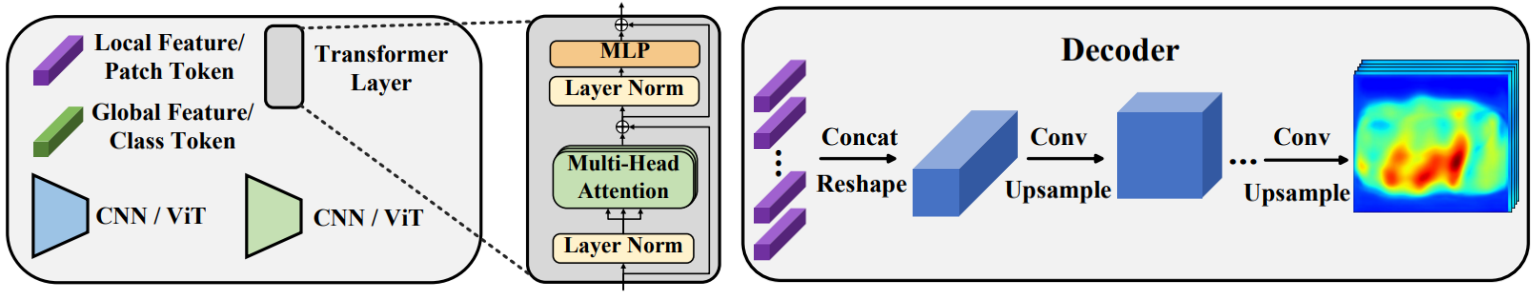
Dense Prediction (Ours)
(Dense Evaluation, Fast)

Generalize on Natural Images



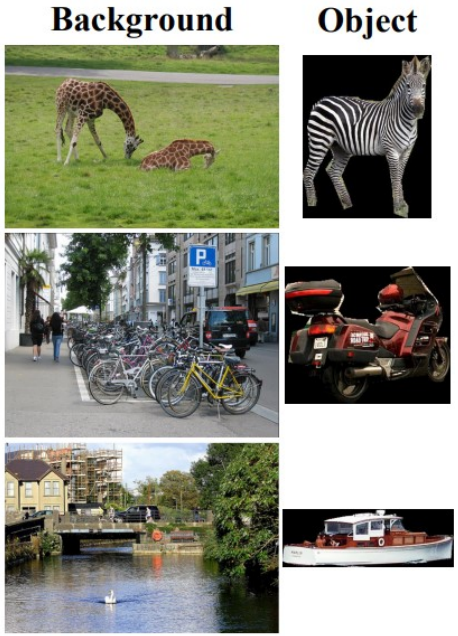
Architecture





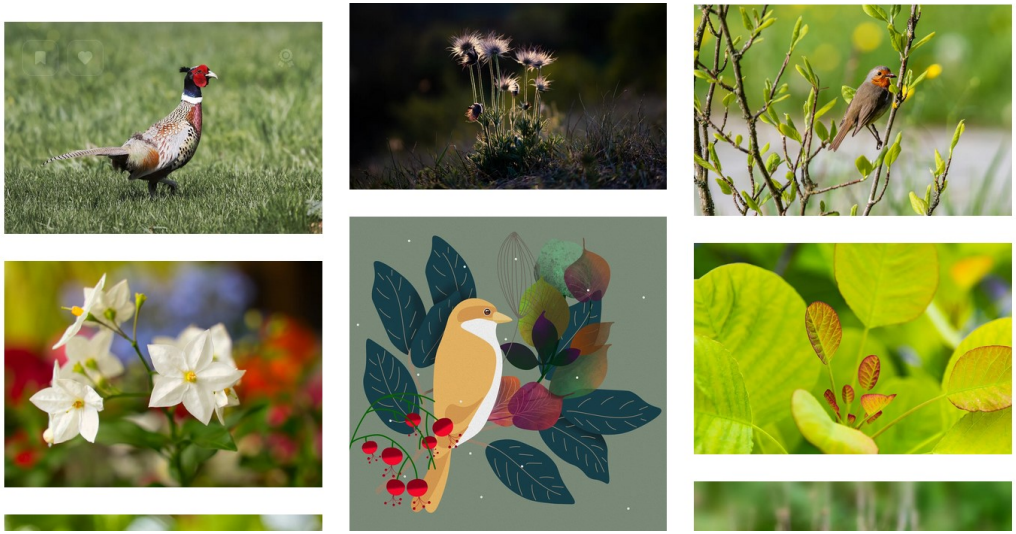
Training Data

OPA Dataset (Small-scale, Annotated)



Liu, Liu, et al. "OPA: object placement assessment dataset." arXiv preprint arXiv:2107.01889 (2021).

Pixabay Dataset (Large-scale, Not Annotated)



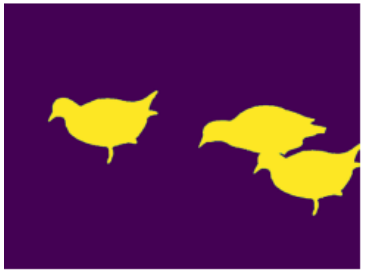
https://pixabay.com/images/search/?order=ec

UCF

Inpainted Pixabay

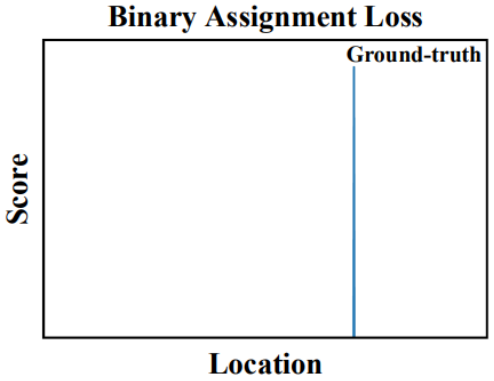
Original

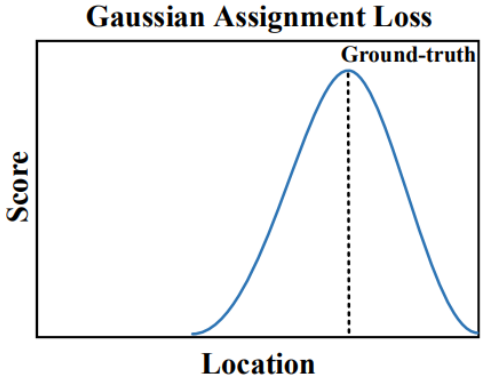
Inpainting Mask

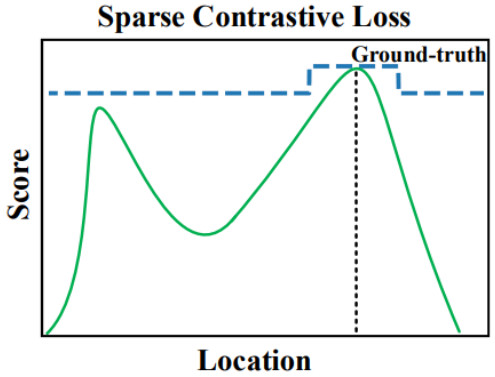


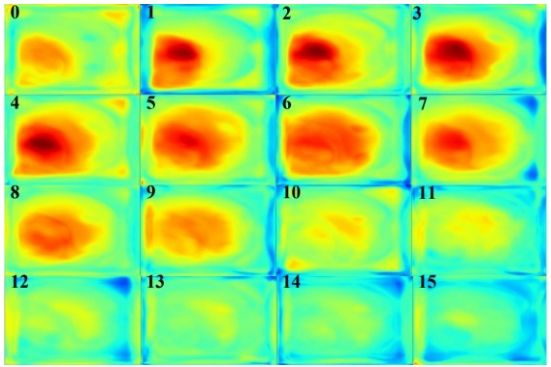
Inpainted Image

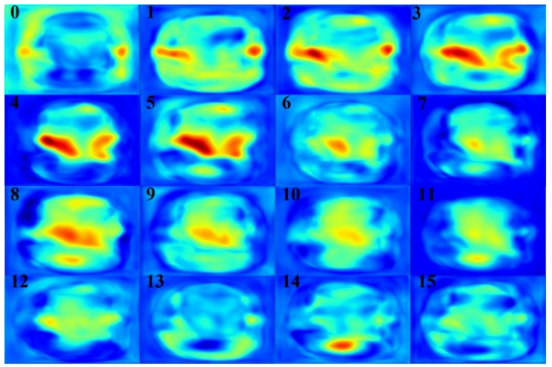
Loss Function







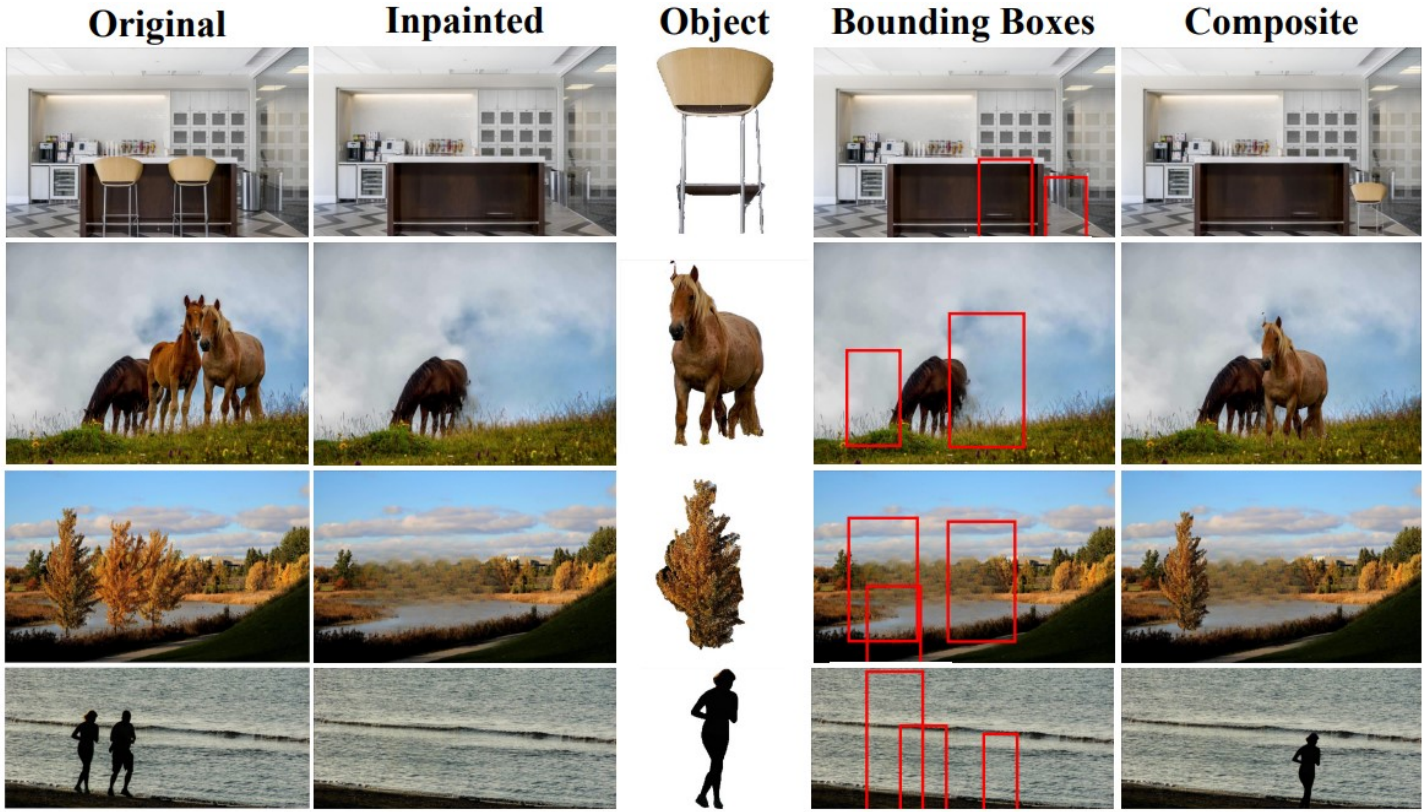




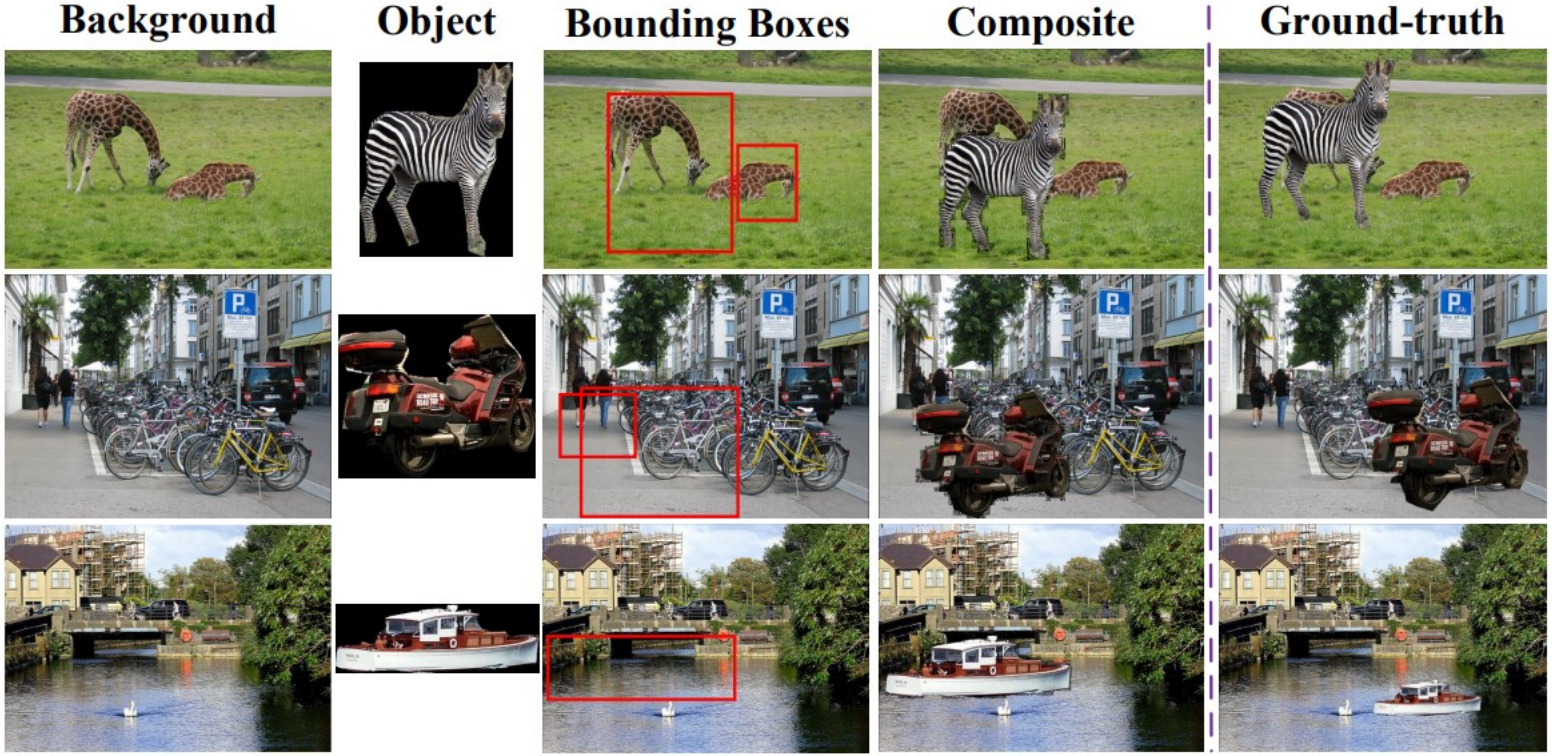
Evaluation with IOU

Method	Infer. Time (s)	Pixabay		OPA	
		$IOU > 0.5$	Mean IOU	$IOU > 0.5$	Mean IOU
Regression [26]	0.08	48.23	0.448	7.24	0.178
†Retrieval [29]	1.69	11.91	0.220	2.08	0.112
Classifier [12]	0.55	6.82	0.147	2.54	0.115
PlaceNet [26]	0.16	19.44	0.308	10.09	0.225
Ours	0.11	**74.74**	**0.620**	**15.95**	**0.241**

Qualitative Results on Inpainted Pixabay



Qualitative Results on OPA

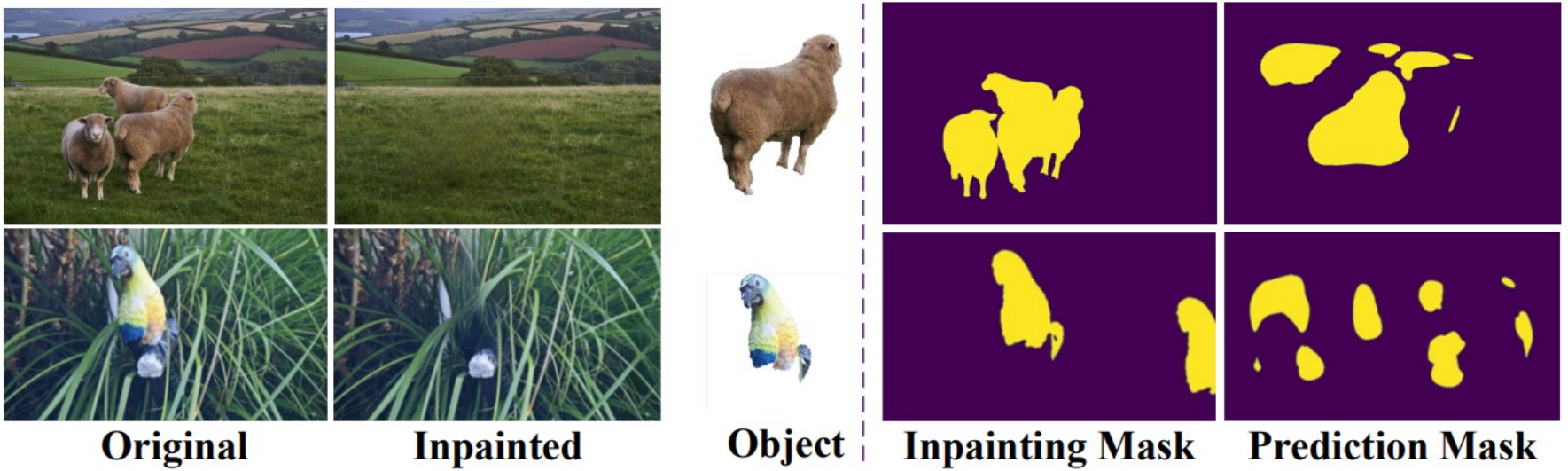


Generalize on Natural Images

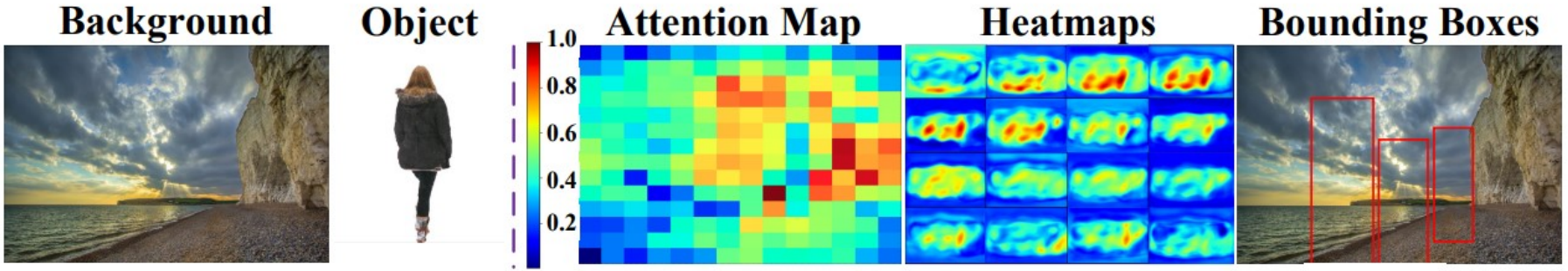
User Study

Method	Unsatisfactory ↓	Borderline	Satisfactory ↑
Regression [26]	46.8	17.4	35.8
†Retrieval [29]	45.4	22.6	32.0
Classifier [12]	72.0	9.6	18.4
PlaceNet [26]	69.0	12.6	18.4
Ours	**42.8**	17.8	**39.4**

Overfitting Inpainting Artifact?



Visualization



Summary

- A novel transformer-based architecture to model the correlation between object image and local clues from the background image, and generate dense object placement evaluation > 10× faster than previous sliding-window method.

- A sparse contrastive loss to effectively train a dense prediction network with sparse supervision.

- Experiments on both manually annotated dataset and large-scale inpainted dataset show significant improvements over previous state-of-the-art methods. It also generalizes well to challenging real-world cases.

Thank You!

UCF