BootPlace: Bootstrapped Object Placement with Detection Transformers

Hang Zhou[1] Xinxin Zuo[2] Rui Ma[3] Li Cheng[1]

[1]University of Alberta [2]Concordia University [3]Jilin University

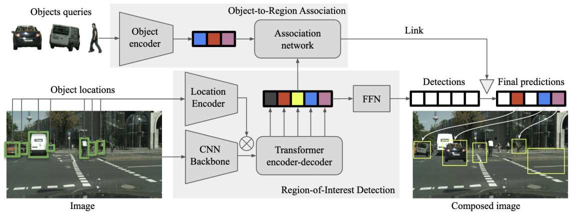
Motivation

- ❑ **Generative models** may limit their ability to model complex data distributions for object placement.
- ❑ **Transformer networks** with a sparse contrastive loss leads to imprecise placement overrelaxed regularization.

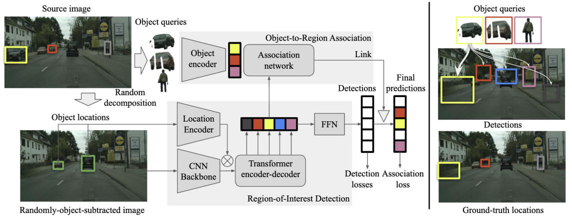
Solution

- ❑ A novel paradigm that formulates object placement as a **placement-by-detection** problem.
- ❑ First identifies regions of interest suitable for object placement by training a **dedicated detection transformer** on **object-subtracted backgrounds with multi-object supervisions**.
- ❑ It then **associates** each target compositing object with detected regions based on semantic complementary.
- ❑ Using a **boostrapped training approach** on randomly object-subtracted images, our model regularizes meaningful placements through richly paired data augmentation.
- ❑ Experimental results on standard benchmarks demonstrate the superior performance of our method in object reposition, significantly outperforming state-of-the-art baselines on Cityscapes and OPA datasets with notable improvements in **IOU scores**.
- ❑ Additional ablation studies further showcase the compositionality and generalizability of our approach, supported by user study evaluations.

BootPlace



Inference framework.



Training framework.

- ❑ Association over all regions of interest in the image for each object patch:

$$P_A(\alpha = i|F) = \frac{\exp\left(g_i(q_k, F)\right)}{\sum_{j\in\{\emptyset,1,2,\dots,N\}} \exp\left(g_i(q_k, F)\right)}.$$

- ❑ Semantic complementary:

$$g_i(q_k, F) = -q_k \cdot F_i/\mu,$$

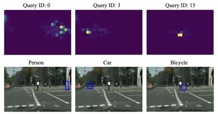
- ❑ Train jointly with standard detection losses, including classification loss and bounding-box regression loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{box} + \beta\mathcal{L}_{asso}.$$

Experimental Results

	Cityscapes				OPA			
	IOU50@1	IOU@1	IOU50@5	IOU@5	IOU50@1	IOU@1	IOU50@5	IOU@5
PlaceNet (ECCV'20) [56]	0	0.045	0	0.045	2.76	0.116	10.09	0.225
GracoNet (ECCV'22) [60]	—	—	—	—	2.49	0.131	16.60	0.248
SAC-GAN (IEEE TVCG'22) [59]	0.806	0.082	1.08	0.085	—	—	—	—
TopNet (CVPR'23) [61]	0.807	0.045	1.61	0.070	11.55	**0.197**	15.95	0.241
BootPlace (ours)	**3.50**	**0.097**	**5.91**	**0.190**	**11.60**	**0.197**	**22.41**	**0.281**

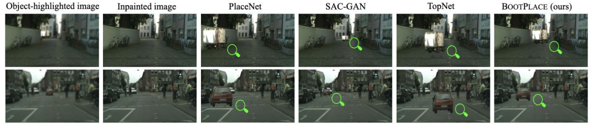
Table 1. **Quantitative results of object reposition** on Cityscapes and OPA datasets, evaluated by IOU50 (%), top-1 and top-5 IOU.



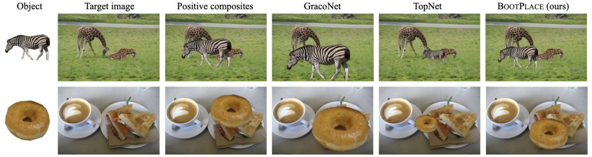
Attention visualization.

	PlaceNet [56]	SAC-GAN [59]	TopNet [61]	Ours
Cityscapes	0.183	0.269	0.246	**0.303**
Mapillary Vistas	0.133	0.285	0.260	**0.323**

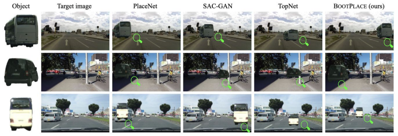
Table 2. **Quantitative comparisons of car placement** on Cityscapes and Mapillary Vistas datasets, evaluated by user study.



Object replacement on Cityscapes dataset.



Object replacement on OPA dataset.



Object placement on Mapillary Vistas dataset.