Thinking Outside the BBox: Unconstrained Generative Object Compositing

Gemma Canet Tarrés[1], Zhe Lin[2], Zhifei Zhang[2], Jianming Zhang[2], Yizhi Song[3], Dan Ruta[1], Andrew Gilbert[1], John Collomosse[1,2], Soo Ye Kim[2]

[1] University of Surrey, [2] Adobe Research, [3] Purdue University

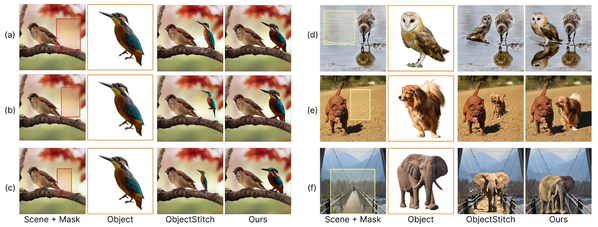
Motivation

Recent Generative Compositing Methods require a mask as input, defining the region of generation. This leads to several limitations:

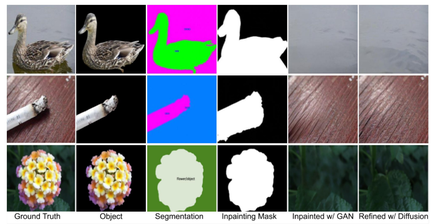
- Drawing an **accurate mask** can be non-trivial, leading to unnatural composite images.
- It limits the ability to synthesize appropriate **object effects** (long shadows, reflections, …).
- **Background areas** around the object tend to be inconsistent with the original background.

We propose:

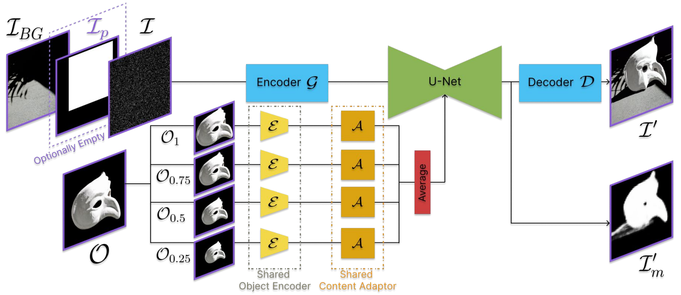
- Introduce novel task: ***"Unconstrained Image Compositing"***
- **Diffusion model** for unconstrained image compositing, trained on synthesized paired data

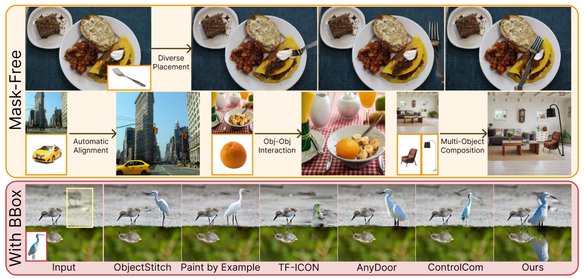


Data Generation



1. Segment **foreground objects** and filter out those too large/ small.
2. Detect shadows using instance shadow detector.
3. Use heuristics for approximating reflection masks.
4. Define inpainting mask as union of object, shadow and reflection masks.
5. Apply GAN-based inpainting model followed by Diffusion-based inpainting model for obtaining a clean background image.





Applications

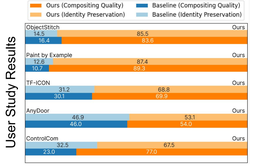
Comparison to SoTA Generative Object Compositing

Benefits of unconstrained compositing approach:

(i) Better **background preservation** (rows 3-6).

(ii) More natural object effects (i.e. **shadows and reflections**) beyond the bounding box (rows 3-4).

(iii) Can adjust any **misaligned bounding box** (rows 1-2).

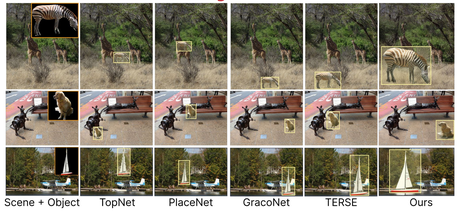


Quantitative Metrics

Method	DreamBooth CLIP-Score↑	DreamBooth DINO-Score↑	DreamBooth DreamSim↓	FID↓	Pixabay-Comp CLIP-Score↑	Pixabay-Comp DINO-Score↑	Pixabay-Comp DreamSim↓
ObjectStitch† [56]	78.018	85.247	0.342	70.111	74.964	77.506	0.488
PaintByExample† [62]	77.782	79.887	0.438	82.923	76.604	75.707	0.515
TF-ICON* [38]	79.094	81.781	0.341	77.368	75.694	77.810	0.485
AnyDoor† [9]	80.619	83.632	**0.272**	72.996	**80.284**	80.829	0.399
ControlCom† [68]	74.312	70.497	0.424	66.071	72.006	67.476	0.614
Ours (w/ bbox)	**80.946**	**85.646**	0.285	**62.406**	77.129	**80.896**	**0.395**

Table 1: Quantitative comparison of composition quality and identity preservation. FID is only computed on Pixabay-Comp, which has ground truth images. †: Model finetuned on the same data as Ours. ‡: Paper version, already includes diverse video and multiview data. *: Paper version, inference-based model that does not require training. ^: Paper version, no available training code.

Comparison to SoTA Object Placement Prediction



Quantitative Metrics

Method	OPA SimOPA ↑	OPA LPIPS↑	OPA IoU > 0.5↑	OPA mean-IoU↑	Pixabay-Comp IoU > 0.5↑	Pixabay-Comp mean-IoU↑	Pixabay-Comp LPIPS↑
TopNet [74]	0.256	2.758	16.8 %	0.094	48.0 %	0.246	1.218
GracoNet [73]	**0.395**	0.836	12.2 %	0.189	30.2 %	0.327	2.832
PlaceNet [69]	0.197	0.746	11.2 %	0.194	8.6 %	0.237	2.072
TERSE [53]	0.319	0.000	10.8 %	0.123	12.2 %	0.230	0.000
Ours (w/o bbox)	0.382	**5.619**	**31.4 %**	**0.196**	**65.4 %**	**0.562**	**3.158**

Table 2: Quantitative evaluation of predicted location and scale of our model compared to state-of-the-art object placement prediction models. LPIPS is $\times 10^{-3}$.