JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

ObjectStitch: Object Compositing with Diffusion Model

Yizhi Song[1], Zhifei Zhang[2], Zhe Lin[2], Scott Cohen[2], Brian Price[2], Jianming Zhang[2], Soo Ye Kim[2], Daniel Aliaga[1]

[1]Purdue University, [2]Adobe Research

Poster: THU-AM-175

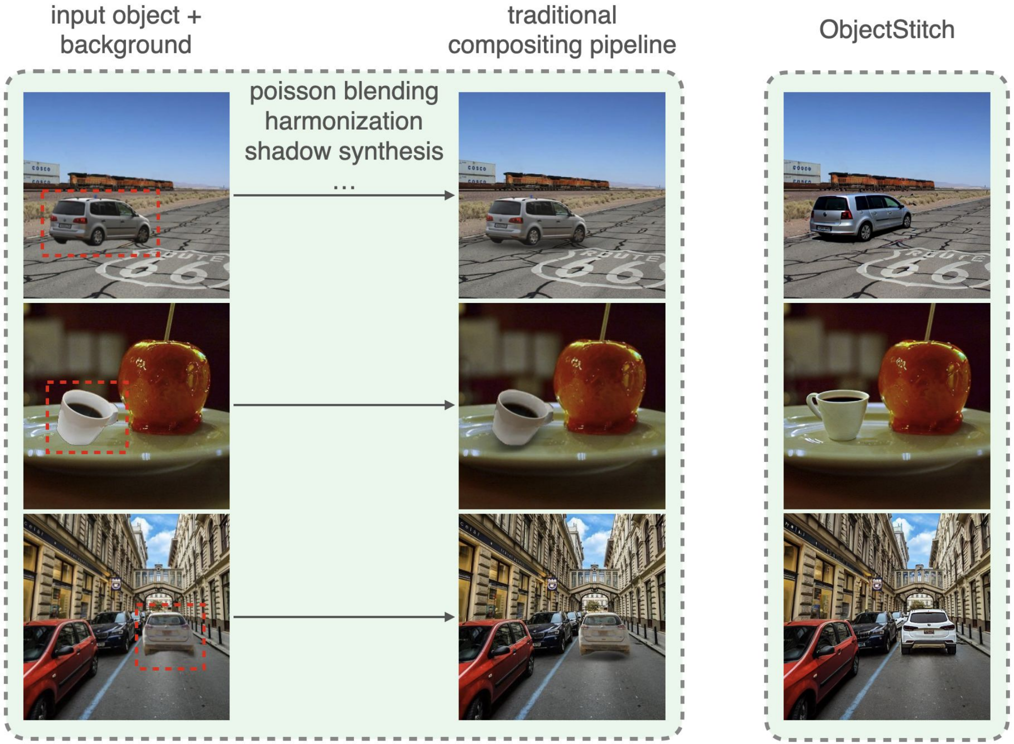
Paper: https://arxiv.org/pdf/2212.00932.pdf

PURDUE UNIVERSITY

Adobe

ObjectStitch Overview

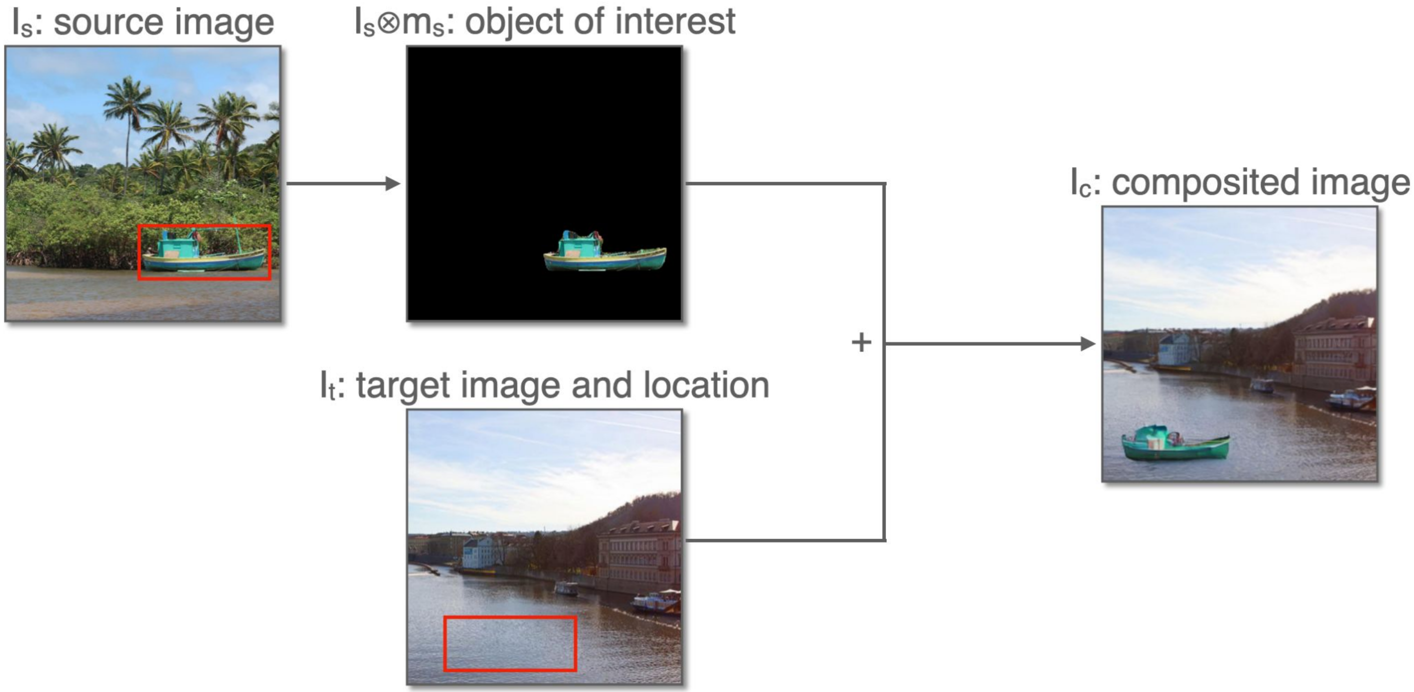
- ObjectStitch simultaneously handles multiple aspects of 2D object compositing
- ObjectStitch is a self-supervised model that does not require task-specific annotation
- We use a content adaptor to maintain categorical semantics and object appearance



2D Object Compositing

Task definition:

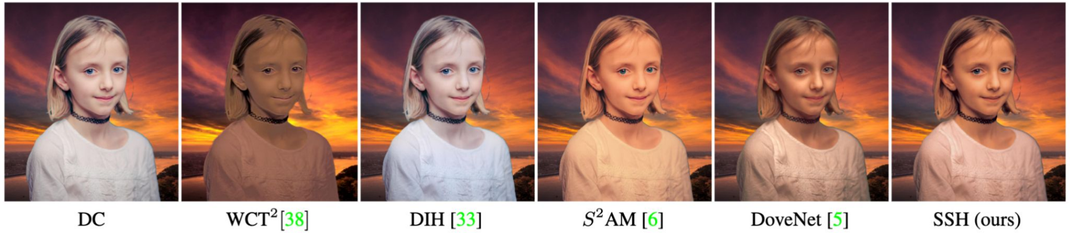
- A common scenario in image editing: foreground image A + background image B = ?
- Given the location and scale in the target image, how to generate a realistic composite image that preserves the identity of the source object?



Related Works - Image Compositing

- ST-GAN[1]: geometric correction
- SSH[2]: harmonization
- SSN[3]: shadow synthesis

Each of them focuses on a single sub-task
They cannot generate novel view

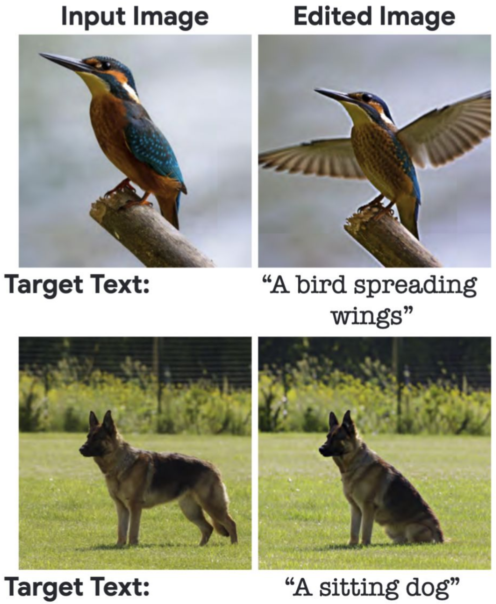


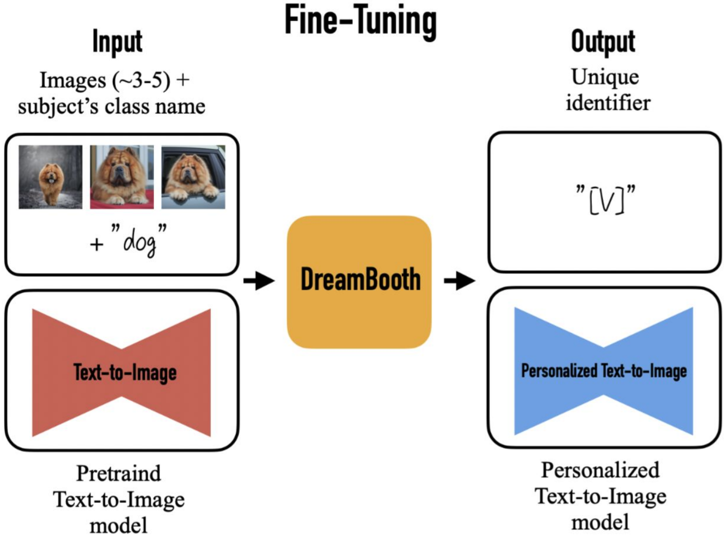
(a) object cutouts (2D masks) (b) soft shadows generated by our SNN from the image-based light map above

[1] Lin, Chen-Hsuan, et al. "St-gan: Spatial transformer generative adversarial networks for image compositing." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
[2] Jiang, Yifan, et al. "SSH: a self-supervised framework for image harmonization." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
[3] Sheng, Yichen, Jianming Zhang, and Bedrich Benes. "SSN: Soft shadow network for image compositing." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Related Works - Object Personalization with Diffusion Models

- DreamBooth[4], Imagic[5]
- Limitation: the model needs to be fine-tuned for each subject on paired images

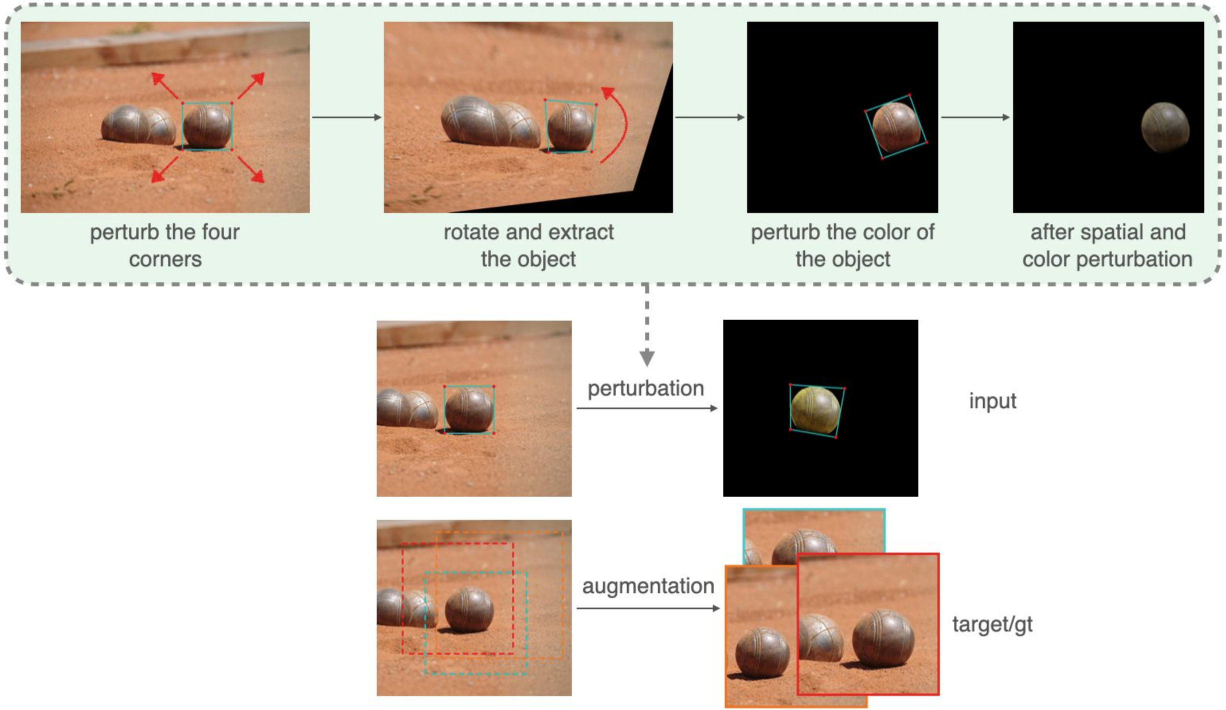




[4] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
[5] Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

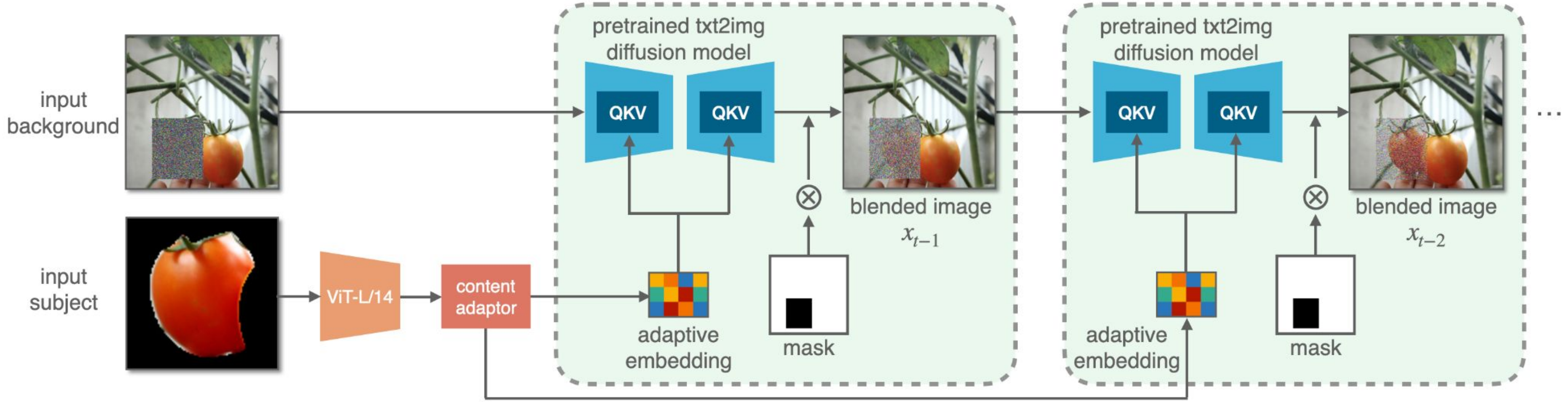
Data Preparation and Self-Supervision

- Task-specific training data is expensive to obtain
- Source: Pixabay
- Augmentations: warping → rotation → color shifting → crop
- Training pairs: segmented object (augmented) + original image



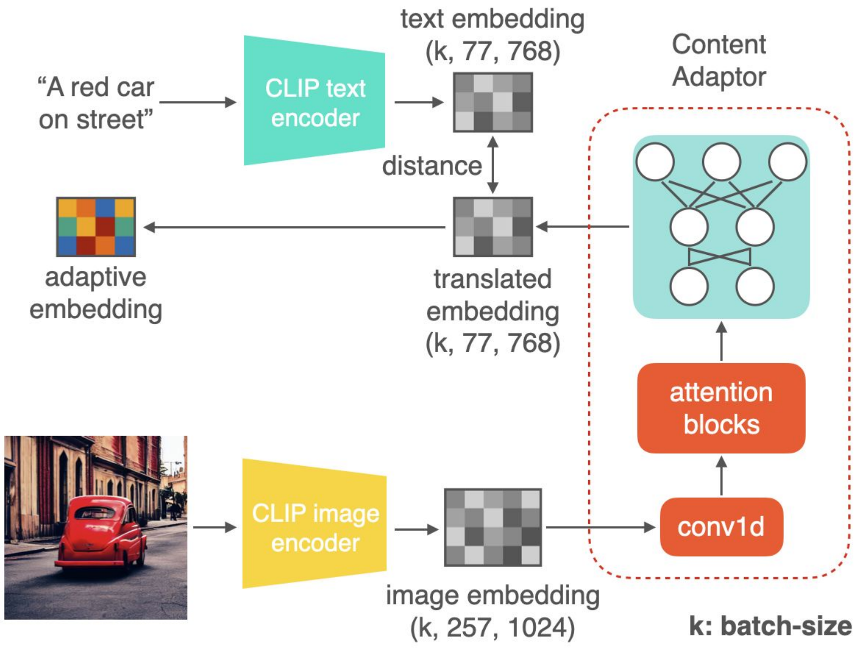
ObjectStitch: Architecture

- Consist of: a generator (pretrained T2I DM) + a content adaptor
- The content adaptor generates adaptive embedding that preserves details
- The mask is applied on the generated image at each iteration



Content Adaptor

- Motivation:
  - Bridge the domain gap between text embedding and image embedding
  - Resolve the mismatch in their dimensions
  - Trained on LAION image-caption pairs



$$\mathcal{L}_{dist} = \|T(\tilde{E}) - E\|_1$$

where T is the content adaptor,
E is the target text embedding

Training

- Content adaptor pretraining

$\mathcal{L}_{dist} = \|T(\tilde{E}) - E\|_1$ where T is the content adaptor

- Content adaptor fine-tuning

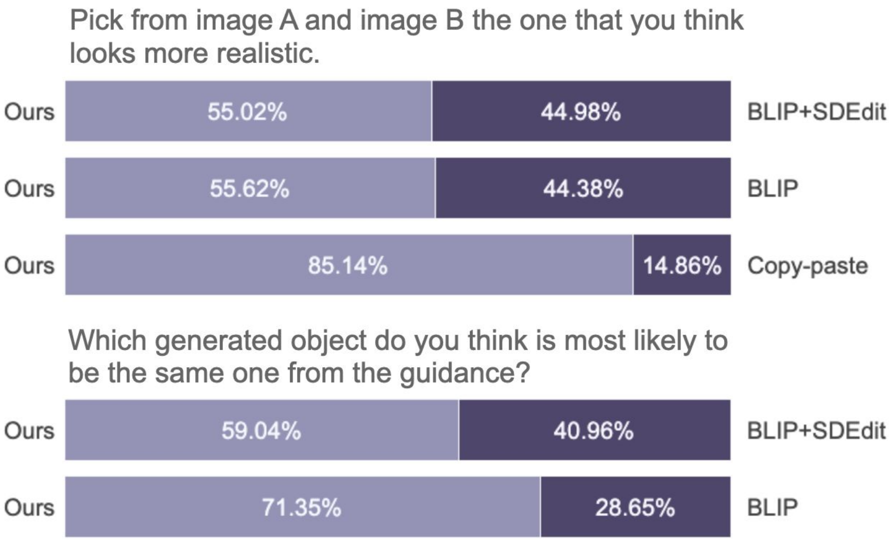
$\mathcal{L}_{adapt} = \mathbb{E}_{T,\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(I_t \circ M, t, T(\widetilde{E}))\|_2^2]$ where I_t is a noisy version of I at step t

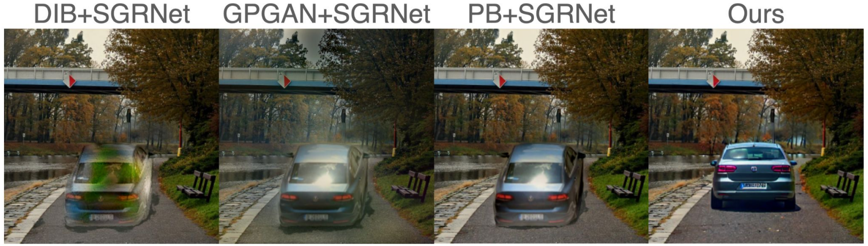
- Generator fine-tuning

$\mathcal{L}_{gen} = \mathbb{E}_{\hat{E},\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(I_t \circ M, t, \hat{E})\|_2^2]$

User Studies

- Collect a real test dataset of 503 object-background pairs
- A side-by-side comparison of the results
- Diffusion model-based baselines: BLIP[6] and SDEdit[7]
- Image blending-based baselines + shadow synthesis





Method	DIB+SGRNet	GPGAN+SGRNet	PB+SGRNet
Ours	82.93%	84.74%	76.91%

BLIP: Li *et al.* 2022; SDEdit: Meng *et al.* 2021; DIB: Zhang *et al.* 2020;
GPGAN: Wu *et al.* 2019; PB: Pérez *et al.* 2003; SGRNet: Hong *et al.* 2022

[6] Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." *International Conference on Machine Learning*. PMLR, 2022.
[7] Meng, Chenlin, et al. "Sdedit: Image synthesis and editing with stochastic differential equations." *arXiv preprint arXiv:2108.01073* (2021).

Quantitative Results

- Metrics: FID[8], LPIPS[9], modified CLIP[10] scores
  - CLIP text score and CLIP image score

$$\mathcal{C}_{txt} = E\left[s \cdot f(I_{pred}) \cdot g\left(B(I_{gt})\right)\right]$$ where B is pretrained BLIP

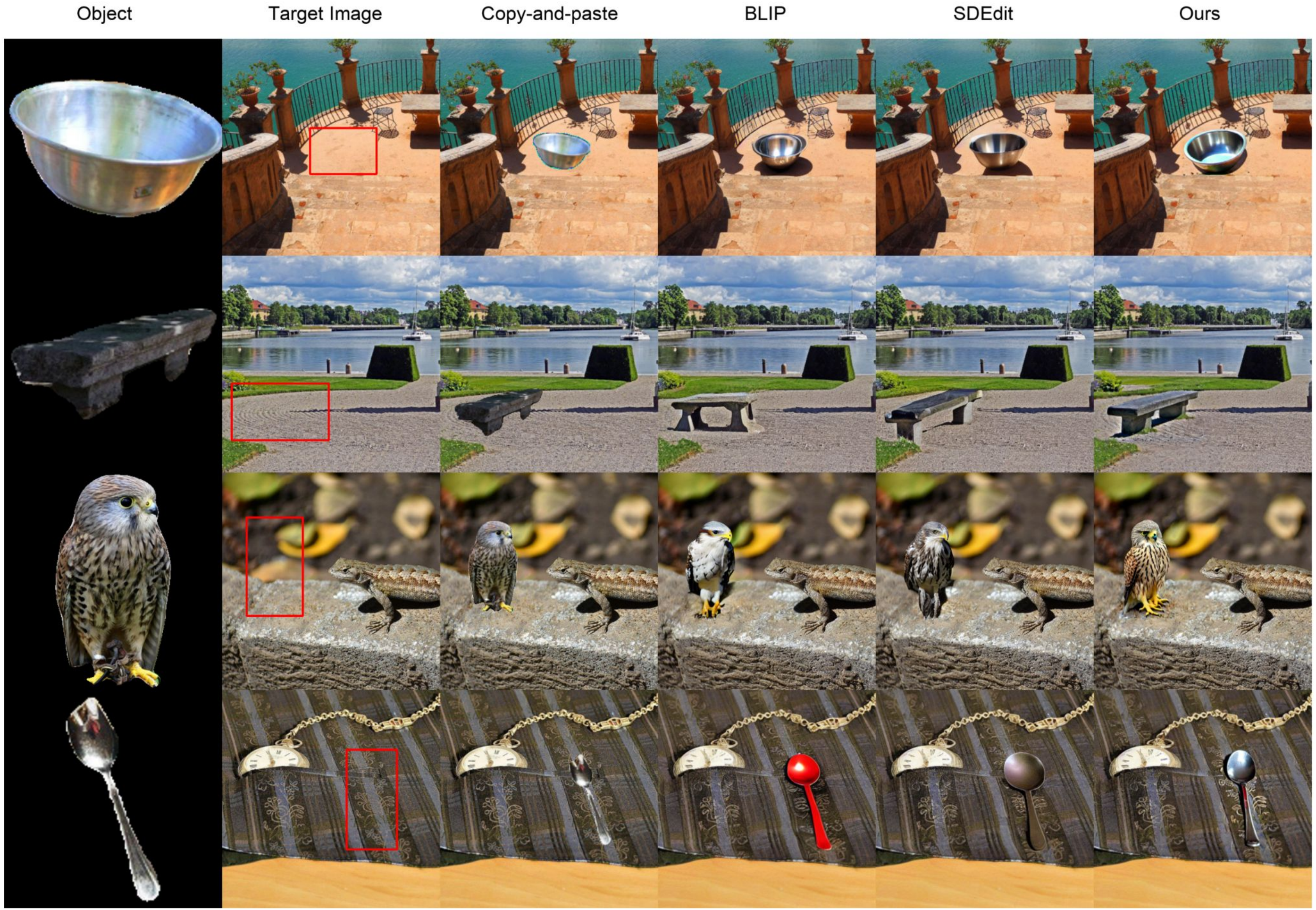
$$\mathcal{C}_{img} = E\left[s \cdot f(I_{pred}) \cdot f(I_{gt})\right]$$

Method	Crop	FID ↓	LPIPS ↓	CLIP text score ↑	CLIP image score ↑
BLIP	✗	18.3673	0.0923	29.6719	95.5625
SDEdit	✗	17.4963	0.0870	29.6563	96.1250
Ours	✗	**15.8191**	**0.0835**	**29.8594**	**97.0000**
BLIP	✓	28.0690	0.2463	29.0313	91.1250
SDEdit	✓	27.0630	0.2312	29.0625	91.8750
Ours	✓	**24.4719**	**0.2223**	**29.4844**	**93.7500**

[8] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." *Advances in neural information processing systems* 30 (2017).
[9] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[10] Hessel, Jack, et al. "Clipscore: A reference-free evaluation metric for image captioning." *arXiv preprint arXiv:2104.08718* (2021).

Qualitative Results

Qualitative Results



Conclusions

- We propose the first diffusion-based method to tackle object compositing
- We introduce a novel content adaptor module
- We present a fully self-supervised framework with data augmentation
- Our model outperforms the baselines on real-world examples