GALA: Toward Geometry-and-Lighting-Aware Object Search for Compositing Supplementary Material

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
sizhu@knights.ucf.edu,{zlin,scohen,kuen,zzhang}@adobe.com,chen.chen@crcv.ucf.edu

Overview

In this supplementary material, we provide the following contents for better understanding of the paper:

1. Example on Lighting Compatibility.
2. Difference with Recent Works.
3. Details of UFO on Pixabay.
4. Qualitative results for non-box scenarios.
5. Qualitative results for object retrieval.
6. Qualitative results on harmonization.
7. Example of mask erosion.
8. Detail of geometry and lighting transformation.
9. Qualitative with Shadow Generation.
10. Discussion on Efficiency.
11. Discussion on Constrained Methods.

1 Example on Lighting Compatibility

In Fig. 1, we show an example of lighting compatibility comparison between our method and UFO [7]. It demonstrates that lighting is a critical factor regarding to the compatibility of images and our method can generate more realistic final composite images.

2 Difference with Recent Works

Despite the large number of papers in the related area, there are very few works that are closely related to our setting. We did not add [4] and [5] in experiment because 1) they have very different settings from our work, i.e. [5] focuses on fine-grained retrieval from a specific category (furniture) and [4] is for constrained

[†] This work was done during the first author's internship at Adobe Research.

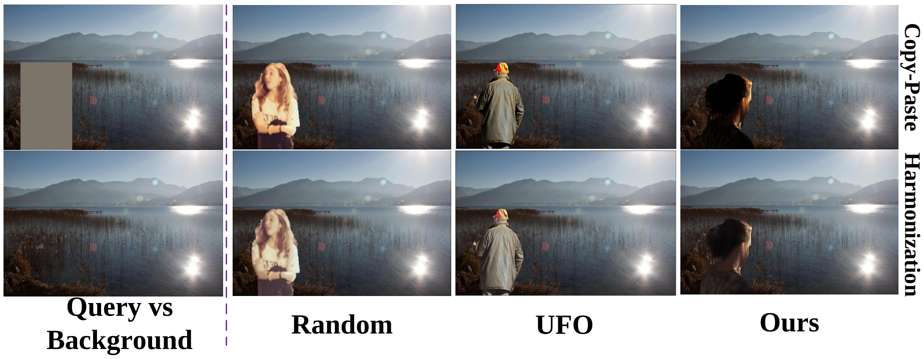


Fig. 1: Comparison between state-of-the-art method (UFO [7]) and the proposed method. Objects retrieved by "Random" and "UFO" do not have good lighting compatibility even after harmonization. Our method better respects the lighting condition (light coming from front-right direction), and thus the final composite image is more realistic.

setting (given categories). However, we focuses on unconstrained setting with diverse categories. 2) [5] trains one model per category which is not scalable for our large-scale setting with hundreds of categories. 3) [4] requires additional annotation for multiple attributes, which is not available for our dataset. 4) [4] and [5] conduct experiments on their own datasets which are not released. Their codes are also not released, so direct comparison is infeasible.

3 Details of UFO on Pixabay

Pixabay is much larger than the datasets used in UFO [7] and it is infeasible to scan all the background images with all the $833,964$ foreground objects. We follow the heuristics proposed in UFO [7] to make it more efficient. We first train a discriminator following UFO [7] to classify positive and negative objects for each background image. The accuracy is about 90% which should be enough to select good positive images. We then find the top retrieved images of each background query using our baseline model. We then scan these images with the discriminator. If the discriminator classify one object image as positive, then we keep it as a candidate positive object for sampling.

4 Qualitative Results for Non-box Scenarios

In Fig. 4, we provide final composite results for non-box scenarios in diverse scenes to demonstrate the generalization ability of the proposed method to open-world scenarios. We generate composite images using the top-2 retrieved objects with automatic location-scale prediction and harmonization [2]. Some of the scenes have extreme lighting conditions, e.g. sunset, while the retrieved objects are still compatible. The location-scale prediction also well respects the geometry of the scene, e.g. the boat is on the water, and the airplane is in the sky. In the first row, the bottom of the tower exactly fits on the surface of the water with realistic lighting. It also works for night scenes and sometimes retrieves long-tail but reasonable objects, e.g. crocodile and road sign.

5 Qualitative Results for Object Retrieval

In Figs. 5 and 6, we show top-5 retrieved objects of three variants, i.e. “Baseline Model”, “Baseline w/ AT”, and “Our Full Model w/ AT,GL”, on Pixabay [1] and Open Images [3]. “AT” denotes the proposed alternating training strategy and “GL” means lighting and geometry contrastive learning. In Figs. 7 and 8, we also show qualitative comparison with UFO [7] and our baseline on CAIS [6]. The proposed method achieves better results in terms of geometry and lighting as compared with the baseline and UFO, while maintaining semantic compatibility.

6 Qualitative Results for Harmonization

In Fig. 9, we show qualitative comparison between “Copy-Paste” and “Harmonization”, both with our top-3 retrieved foreground object images and a bounding box given by the user. Copy-paste results already look realistic and harmonization results have little difference with the copy-paste results, indicating that our retrieved objects are already highly compatible with the background image.

7 Example of Mask Erosion

In Fig. 2, we provide an example on the mask erosion augmentation, which reduces the background edge pixels in the foreground objects. In our experiment, we randomly erode the segmentation mask for 3 to 10 pixels. And the bounding box on background image is randomly extended with 1 to 20 pixels.

8 Geometry and Lighting Transformation

For the lighting transform, the radius for Gaussian blur is 100. For the exponential function, we first normalize the lighting map with the average value, then the maximum value x_{max} is enhanced to β by applying $y = x^{log(\beta)/log(x_{max})}$ on all pixels. We select $\beta = 5$ as it generates the best performance. For geometry transform, we use random perspective transform with a distortion scale $\alpha = 0.8$, and the left-right flip is adopted with 50% chance. The performance comparison in provided in Tables 1 and 2. The transformation is stronger when using a large α or β. When α or β is too small, the transformed object images still have similar geometry or lighting with the original object images, resulting in performance

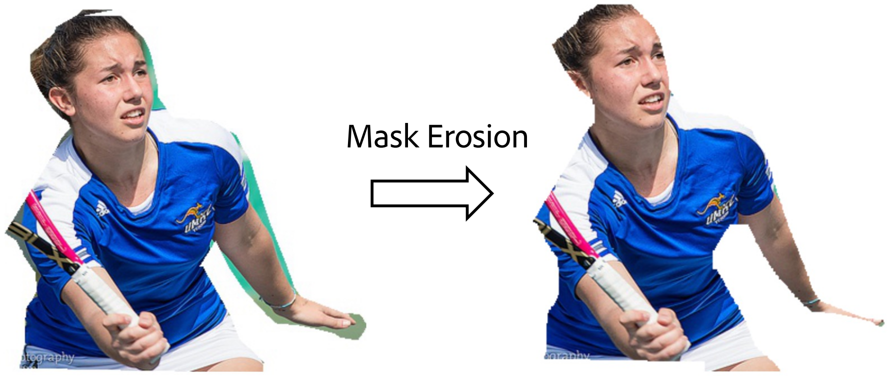


Fig. 2: Example of mask erosion on CAIS.

drop if they are considered as negative samples in contrastive learning. When the transform is strong enough, the performance is not sensitive to α, β. The learning rate is reduced by 10× for the second round of alternative training.

	mAP
$\alpha = 0.6$	31.80
$\alpha = 0.8$	32.67
$\alpha = 0.9$	32.33

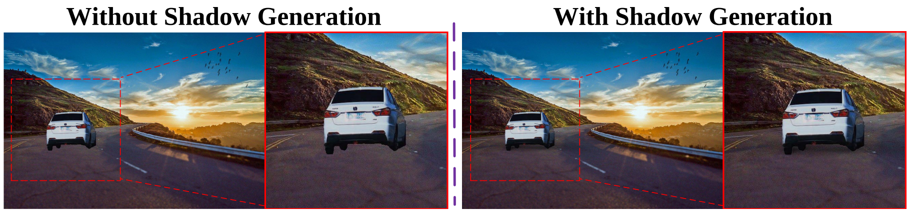
Table 1: Retrieval results with different α values (β is set to 5).

	mAP
$\beta = 3$	30.88
$\beta = 4$	32.63
$\beta = 5$	32.67

Table 2: Retrieval results with different β values (α is set to 0.8).

9 Qualitative with Shadow Generation

This work focuses on the first and important step of compositing pipeline, i.e. finding compatible object images, thus only applies simple harmonization. Real-world compositing need more steps, e.g. shadow generation and fine-grained 3D adjustment, to generate more realistic images. The car could look like flying because there is no shadow, but it looks more realistic when off-the-shelf auto-shadow generation is applied (see Fig. 3). Our compositing-aware retrieval is first key step, and can be combined with other post-processing steps to generate realistic results for real-world application.


Without Shadow Generation

With Shadow Generation

Fig. 3: Shadow generation example. Zoom-in for details.

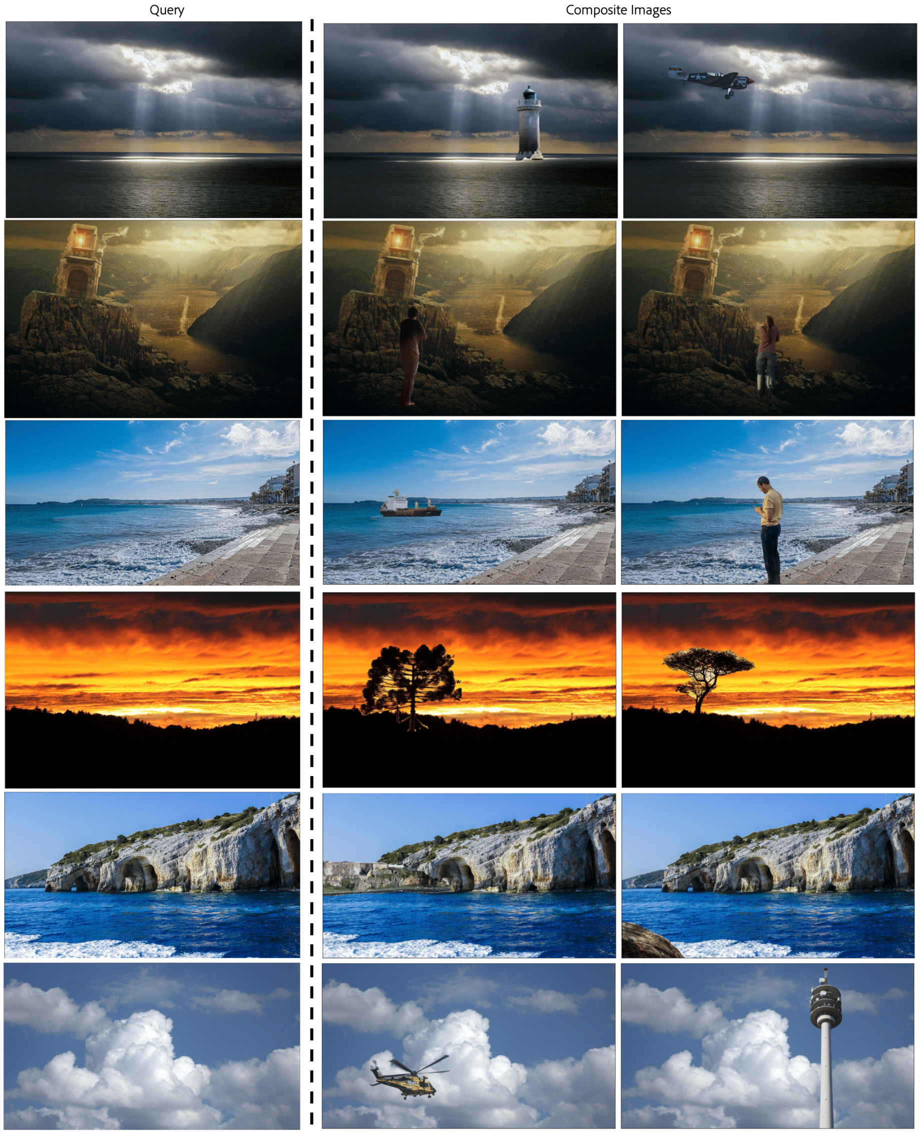
10 Discussion on Efficiency

Our method uses the same backbone as UFO for both foreground and background branches, so the model complexity and inference time is exactly the same as UFO. The inference time per query is 10.2ms for network forward and 3.0ms for retrieval from 83,691 reference images, which is satisfactory for real-time application. As for training, the proposed transformations can be efficiently computed on the fly, but UFO requires training of an additional discriminator and a complete scan of all the training samples before standard training. Discussion will be added in the final version.

The random seed and greedy search algorithm is only for non-box scenarios. The goal is to simultaneously find compatible objects and recommend the location/scale to place it. To the best of our knowledge, there is no prior work or straightforward way to achieve the same goal. We agree that the inference speed could be an issue if the greedy search space is very large, and we will address this with a specifically trained end-to-end network in the future.

11 Discussion on Constrained Methods

The CFO-C and CFO-D are constrained search methods with one model for each category and additional classifier or discriminator, which is not scalable in practice. A specifically trained model for a certain category (e.g. Dog) could be better than an unconstrained model for all categories, because object images in certain categories could be highly similar and hard to distinguish. Dog and Person have the largest number of object images in CAIS dataset, this might be the reason that CFO could perform well on specific categories.


Query
Composite Images

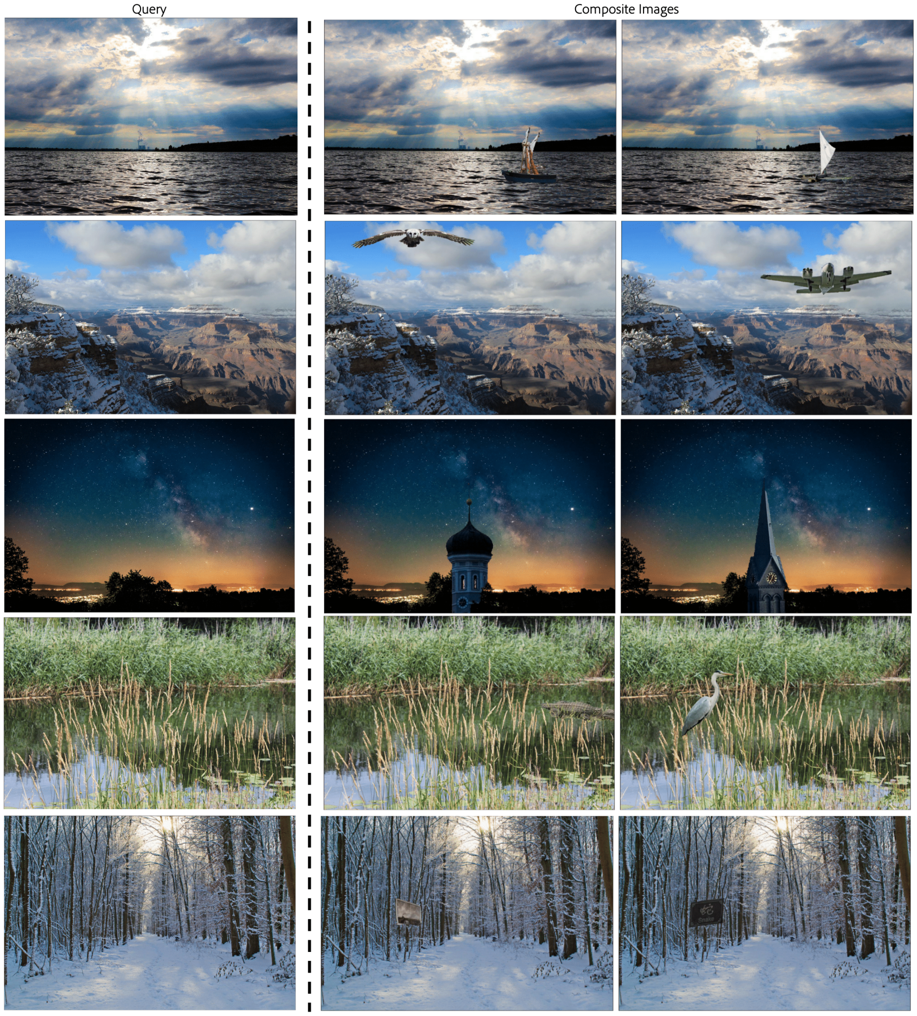
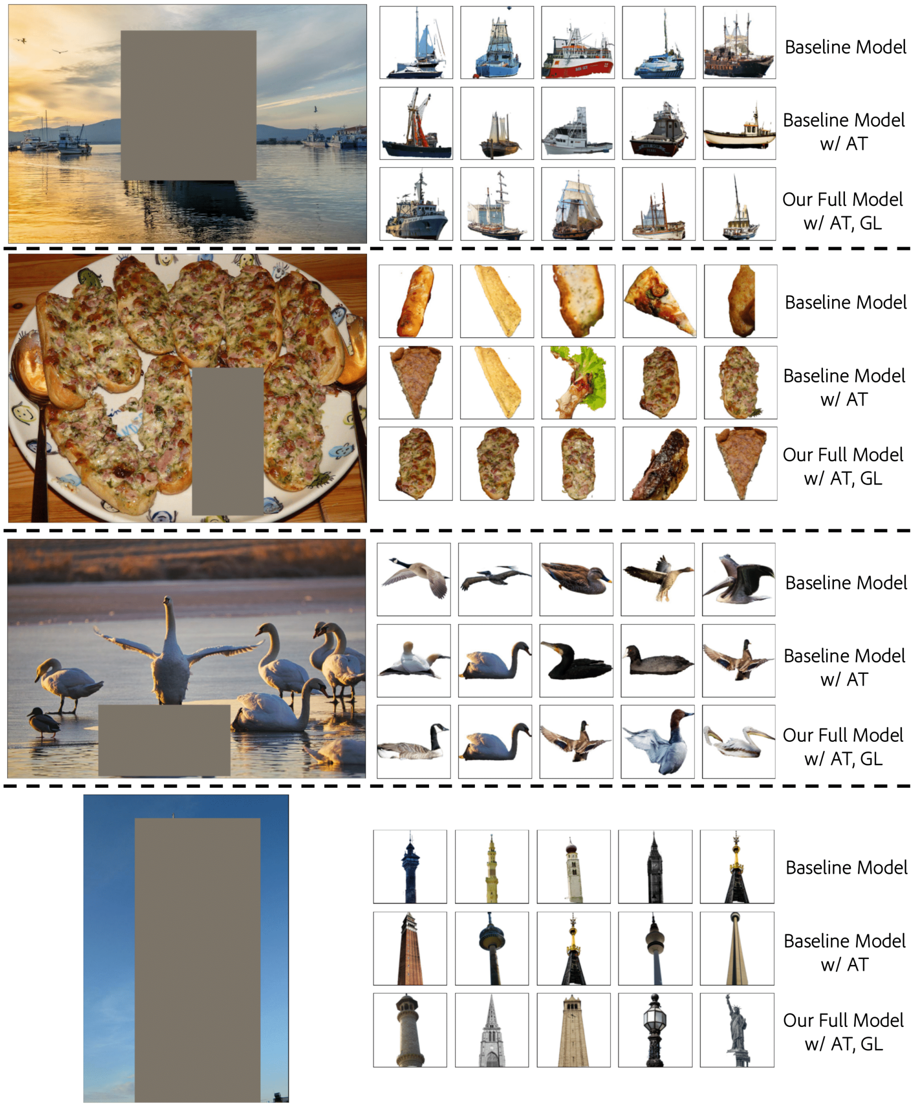


Fig. 4: Qualitative results on non-box scenarios, where the users do not provide a bounding box. The results are based on top-2 retrieved objects with automatic location-scale prediction and harmonization[2].


Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL
Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL
Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL
Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL

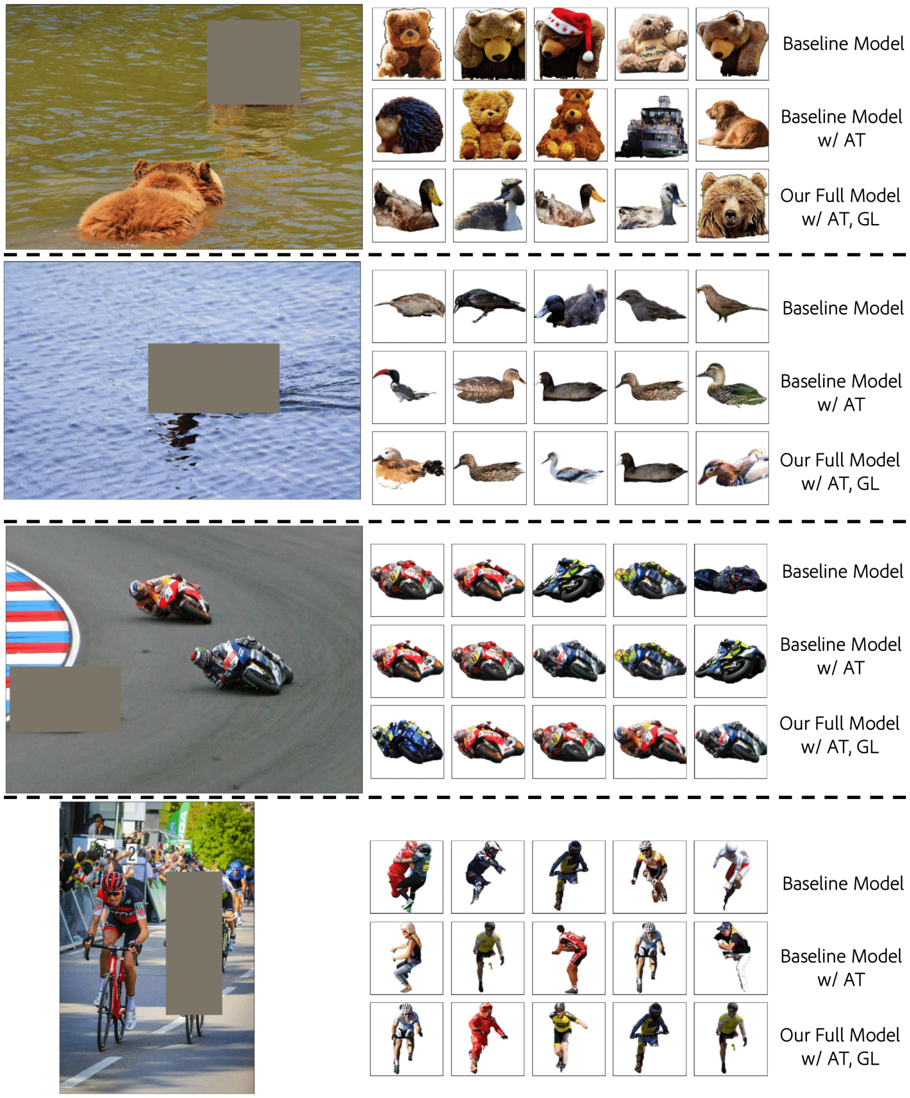
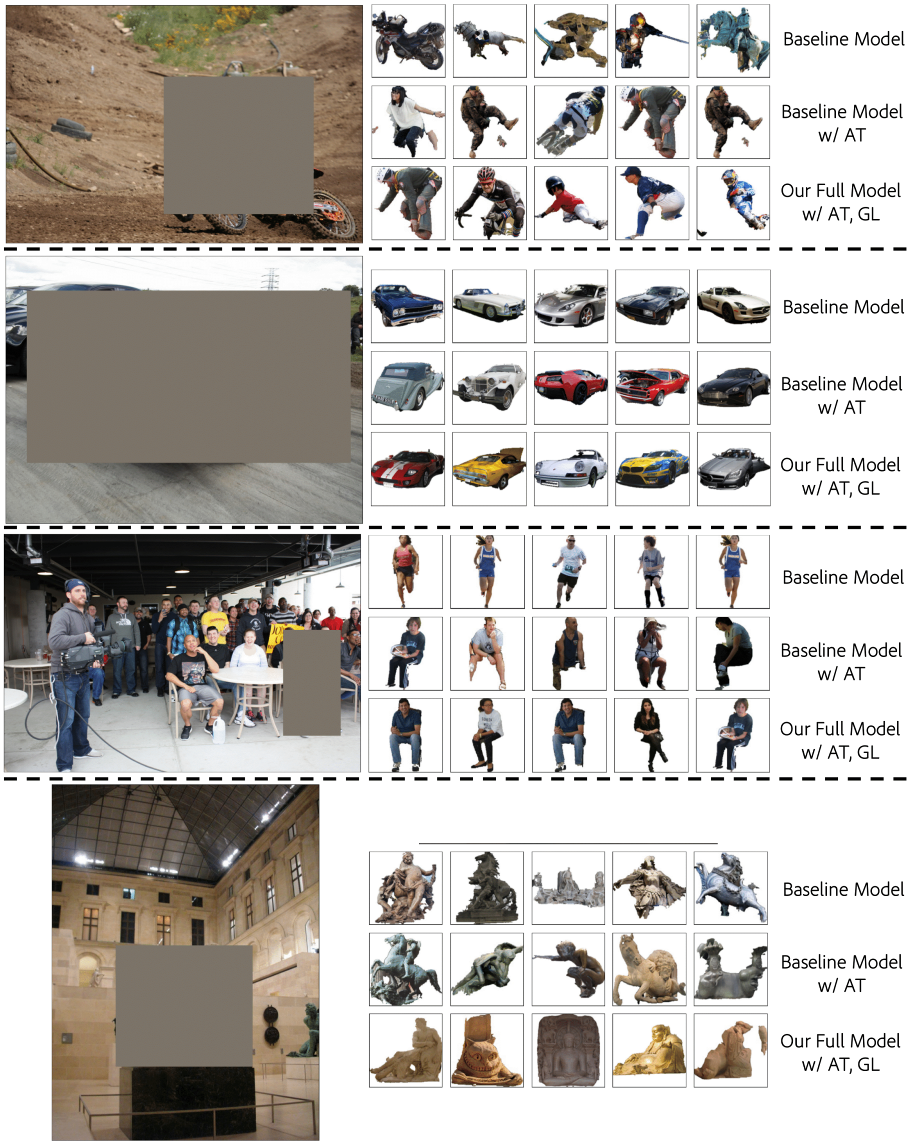


Fig. 5: Retrieval results on Pixabay. "AT" means alternating training and "GL" means lighting and geometry contrastive learning.


Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL
Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL
Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL
Baseline Model
Baseline Model w/ AT
Our Full Model w/ AT, GL

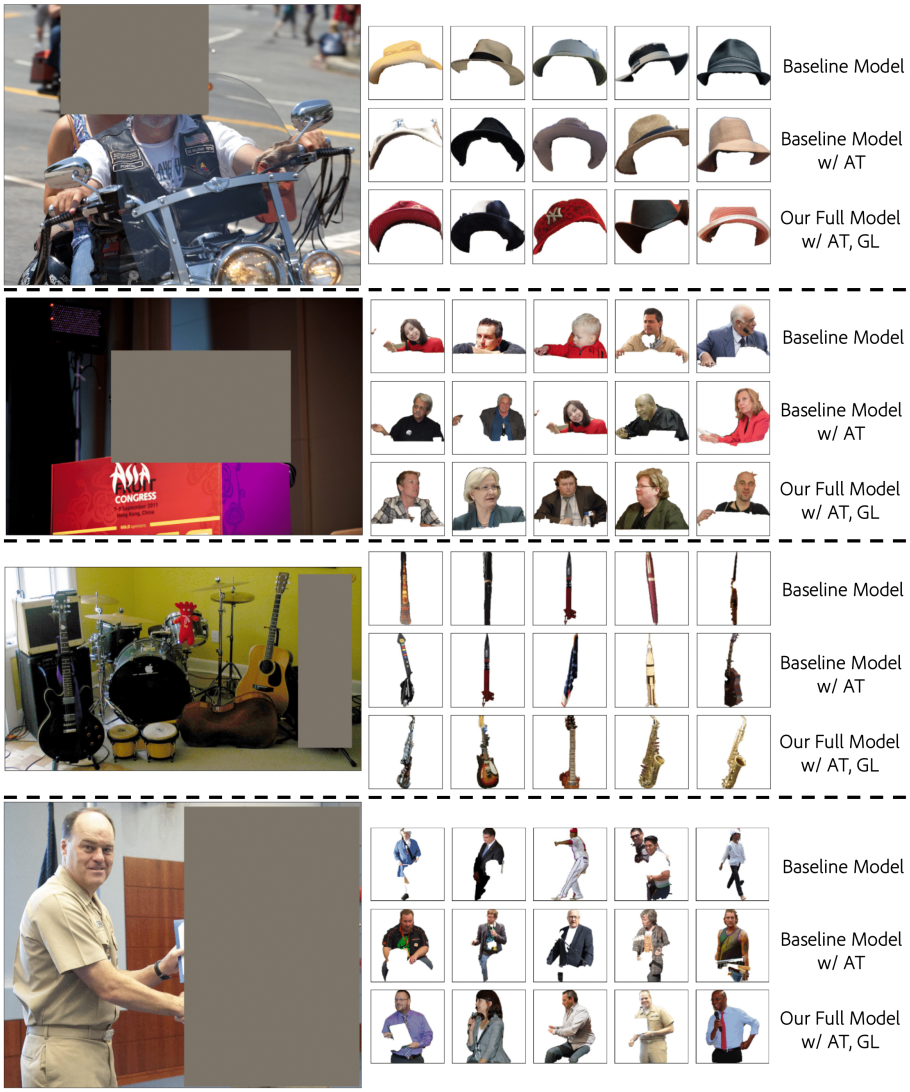


Fig. 6: Retrieval results on OpenImages. "AT" means alternating training and "GL" means lighting and geometry contrastive learning.

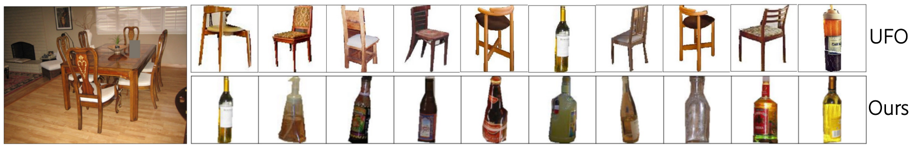


Fig. 7: Qualitative comparison with UFO[7] on CAIS.

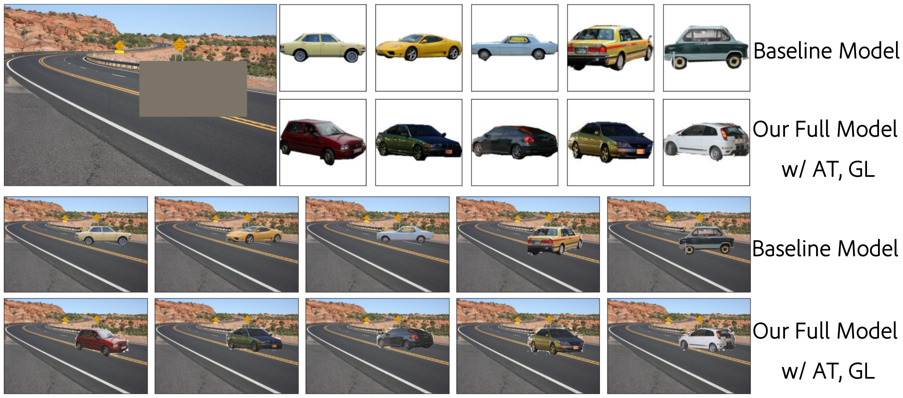


Fig. 8: Qualitative results on CAIS. "AT" means alternating training and "GL" means lighting and geometry contrastive learning.

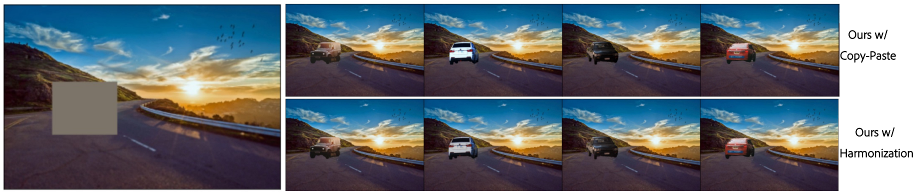


Fig. 9: Qualitative comparison between copy-paste and harmonization on Pixabay.

References

1. https://pixabay.com/, https://pixabay.com/ 3
2. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4832–4841 (October 2021) 2, 7
3. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 3
4. Li, B., Zhuang, P.Y., Gu, J., Li, M., Tan, P.: Interpretable foreground object search as knowledge distillation. In: European Conference on Computer Vision. pp. 189–204. Springer (2020) 1, 2
5. Wu, Z., Lischinski, D., Shechtman, E.: Fine-grained foreground retrieval via teacher-student learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3646–3654 (2021) 1, 2
6. Zhao, H., Shen, X., Lin, Z., Sunkavalli, K., Price, B., Jia, J.: Compositing-aware image search. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 502–516 (2018) 3
7. Zhao, Y., Price, B., Cohen, S., Gurari, D.: Unconstrained foreground object search. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2030–2039 (2019) 1, 2, 3, 11