# No Cameraman Left Behind

Michael McNally and Julia Thorne

## Abstract

We implemented No Cameraman Left Behind, which deals with the problem of how to include the photographer in a group photo. We explored both traditional and machine learning methods for segmenting the cameraman, placing them in an appropriate location in the group image, and blending them into a realistic composite image result. Ultimately, we found that a mix of deep learning and traditional methods allowed us to create an efficient and effective pipeline to perform this task. Quantitative assessment of result image quality was challenging, but qualitative assessment was adequate for development of the pipeline. Further research is necessary to handle challenging cases of illumination differences and image quality differences between the two photos.

## Introduction

A common problem in group photos is the difficulty of including the cameraperson in the final image. There are a few typical suboptimal solutions to this problem: require the photographer to use a tripod and set a timer, then run into the frame; ask a stranger to take a photo; or leave the photographer out entirely. This project takes separate photos of a group and the cameraperson in the same location and combines them into one realistic final image of the entire group using image segmentation and blending techniques.

This is essentially a problem of extracting the person in the foreground of the image of the cameraperson (image segmentation) and blending it into the picture of the group (image compositing). A state-of-the-art approach might involve training a deep learning model to do the process completely automatically, but in the end this approach requires a significant amount of time, data, and computational resources. Our process combines a pre-trained deep learning-based human segmentation model with more traditional object placement and image blending techniques to achieve a realistic result with reduced computational load. We additionally address the scenario where there is not adequate room for the cameraperson in the frame by using image carving techniques.

## Related Works

Approaching this as a problem of image composition, a good overview of the problem can be found in Niu et al., 2021 [6], touching on related works in image segmentation/matting [3, 4], object placement [8], image blending [12, 13], image harmonization [7], and shadow generation as the main aspects of creating a realistic-looking composite image. Human-specific segmentation was explored by Chen et al. (2023) [1] and was useful to this project. For the purposes of benchmarking, there are existing composite image quality assessment metrics explored in Golestaneh et al. (2022) [2], Mittal et al. (2012) [5], and Zhu et al. (2015) [9].

## Methods

We outline below the four following steps of our pipeline: segmentation of the cameraperson, placement of the cameraperson in the target image, blending the cameraperson into the target image, and harmonizing the final image.

I. Segmentation

To segment the cameraperson from the background in their image, we investigated both manual and automatic methods for generating a mask of foreground people in images. Ultimately, we selected Semantic-Guided Human Matting (SGHM) [1] as the most robust and efficient option that generated accurate masks to segment humans in images completely automatically. Both the source and the target images are run through the SGHM model to obtain a mask of the cameraperson, as well as a mask of the people in the target for object placement purposes in addition to blending.

II. Cameraperson Placement

Once the cameraperson has been segmented from the background of their image, we determine optimal placement in the target image by locating empty space in the human mask of the target image. For this purpose, we developed an algorithm that takes in the masks of both the cameraperson and the target photo person/people and 1) calculates the width of the cameraperson, 2) calculates the column of the leftmost pixel and the rightmost pixel containing humans in the target image, and the distance from these pixels to the image edges, and 3) returns the central coordinates of the largest space, vertically aligned with the target person or group.

In the case where neither side of the image was large enough, we apply an image carving algorithm [10] to widen the image by the appropriate number of pixels to accommodate the width of the cameraperson before image blending is performed.

III. Image blending

We found that the most effective and efficient technique for blending the cameraperson was traditional Poisson blending, and used OpenCV's seamlessClone for this portion of the pipeline. This library function implements the seamless cloning algorithm from Perez et al. (2003).

IV. Harmonization

Due to the fact that the cameraperson and target images are taken at the same time in roughly the same locations, we found that illumination harmonization was largely unnecessary. In the case where the resulting image has a noticeably unnatural result, we optionally applied Contrast Limited Adaptive Histogram Equalization [11] as a final "smoothing" operation.

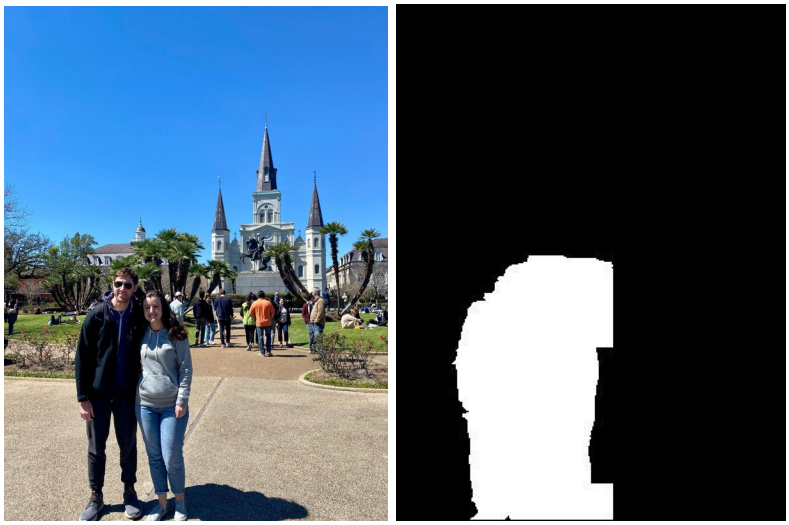
## Experiments and Results

For each step of the pipeline outlined above, we tested at least one traditional computer vision method as well as at least one deep learning based method, and chose the methods with the best performance for our final application. A further discussion of our process with some sample images follows below.

I. Segmentation

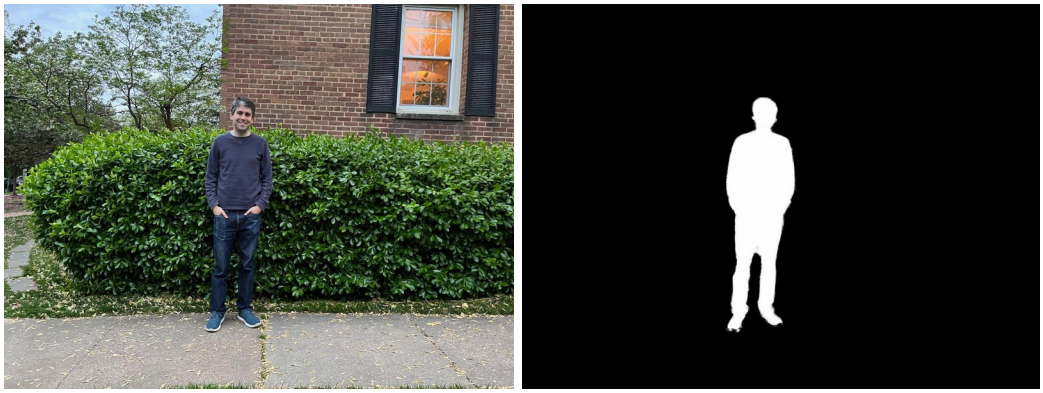
***Traditional method(s)*** We used a two-step process, which required first drawing a bounding box around the cameraman in the source image either through user input or via Histogram of Oriented Gradients human detection algorithm, and then applying Grab-Cut Segmentation (treating the cameraman inside the box as “foreground” and everything outside the box as “background”).

Input Output



***Deep Learning method(s)*** We explored two deep learning approaches: the first was a Deep Image Matting [4] model, which produced more accurate masks automatically, but had a long runtime and had particular difficulty correctly segmenting feet. We switched to a Semantic Guided Human Matting model, with results shown below, which yielded much more consistent results than both the traditional method and the Deep Image Matting approach and with a faster runtime.

Input Output



II. Cameraperson Placement

***Traditional method(s)*** We tried both allowing the user to specify where in the target image to place the cameraman, and creating an algorithm to find “blank” space in the target image based on the mask

returned by step 1. Both methods performed well, but we wanted the application to be able to run with minimal user input, so we opted for the latter approach.

***Deep Learning method(s)*** We investigated several object placement models, including one specifically for placing people in composite images [8]. In the end, we found that in this specific scenario, the traditional method described above was completely adequate and required minimal computational resources.

***Extra method*** We found that sometimes the optimal location for the source would cause it to exceed the bounds of the target image, so we added an optional Seam Carving step to enlarge the target image when this occurred.

Source Target Output

III. Blending

***Traditional method(s)*** Our first attempt used the OpenCV SeamlessClone function implementing traditional Poisson blending, which worked quite well.

***Deep Learning method(s)*** We initially received decent results with a “ghostly” but otherwise realistic cameraperson blended into the target image using a pre-trained Deep Image Blending [13] model, which uses a Poisson loss function to improve texture transfer to the composite image. We likely could have achieved better results with this by adjusting some parameters or the input mask, but eventually discontinued this method due to the excellent results we achieved with traditional Poisson blending.

Source Target Output

IV. Evaluating the image

***Traditional method(s):*** We used a qualitative analysis, which consisted of attempting to objectively decide whether the resulting image was believable. We considered placement, blurry spots, shadows, luminescence, and paid particular attention to the boundary regions and whether any background color made it into the source after blending.

***Deep Learning method(s):*** We tried using a BRISQUE model for No-Reference Image Quality Assessment, but after trialing this on a number of composite and real images, we found the results were inconsistent and decided to focus on qualitative assessment..

Shown below are several results of full runs of our final pipeline:

Source Target Output

Source Target Output

Source Target Output

Source Target Output

## Conclusion

In conclusion, we found that for the purposes of this application, traditional methods were extremely useful for most stages of our pipeline. Most machine learning approaches would require fine-tuning and excessive computational resources to perform well, and many traditional methods can still produce excellent results. The primary area that benefited from a deep learning approach was creating the masks of the source and target images, where Semantic-Guided Human Matting outperformed both a more general deep image matting model and the traditional GrabCut approach. With a good enough mask of the targets computed automatically via Semantic-Guided Human Matting, we were able to automate placing blending the photographer into the final image and produce a realistic composite.

Further research is needed to handle cases where there are illumination or quality differences between the two images. In addition, while we tested several methods of quantitative assessment of results, we did not find a way to accurately measure result image quality and had to focus on qualitative assessment, which was sufficient for these experiments.

## References


1. Chen, X. et al. (2023). Robust Human Matting via Semantic Guidance. In: Wang, L., Gall, J., Chin, TJ., Sato, I., Chellappa, R. (eds) Computer Vision – ACCV 2022. ACCV 2022. Lecture Notes in Computer Science, vol 13842. Springer, Cham. https://doi.org/10.1007/978-3-031-26284-5_37
2. Golestaneh, S. A., Dadsetan, S. and Kitani, K. M. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 3989-3999, doi: 10.1109/WACV51458.2022.00404.
3. Khattab, Dina & Ebied, Hala & Hussein, Ashraf & Tolba, Mohamed. (2015). Automatic GrabCut for Bi-label Image Segmentation Using SOFM. Advances in Intelligent Systems and Computing. 323. 579-592.
4. Li, J., Zhang, J., Maybank, S.J. et al. Bridging Composite and Real: Towards End-to-End Deep Image Matting. Int J Comput Vis 130, 246–266 (2022).
5. Mittal, A., Moorthy, A. K., and Bovik, A. C. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.
6. Niu, L., Cong, W., Liu, L., Hong, Y., Zhang, B., Liang, J., & Zhang, L. (2021). Making Images Real Again: A Comprehensive Survey on Deep Image Composition. ArXiv, abs/2106.14490.
7. Song, S., Zhong, F., Qin, X., Tu, C. (2020). Illumination Harmonization with Gray Mean Scale. In: , et al. Advances in Computer Graphics. CGI 2020. Lecture Notes in Computer Science(), vol 12221. Springer, Cham. https://doi.org/10.1007/978-3-030-61864-3_17
8. Tan, F., Bernier, C., Cohen, B., Ordonez, V., and Barnes, C. Where and Who? Automatic Semantic-Aware Person Composition. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018, pp. 1519-1528, doi: 10.1109/WACV.2018.00170.
9. Zhu, Jun-Yan & Krähenbühl, Philipp & Shechtman, Eli & Efros, Alexei. (2015). Learning a Discriminative Model for the Perception of Realism in Composite Images. 10.1109/ICCV.2015.449.
10. Avidan, Shai & Shamir, Ariel. (2007). Seam Carving for Content-Aware Image Resizing. SIGGRAPH ‘07: ACM SIGGRAPH 2007 papers. doi: 10.1145/1275808.1276390.
11. K. Zuiderveld: Contrast Limited Adaptive Histogram Equalization. Graphics Gems IV, Academic Press 1994. doi: 10.1016/B978-0-12-336156-1.50061-6
12. Pérez, P., Gangnet, M., and Blake, A. 2003. Poisson image editing. ACM Trans. Graph. 22, 3 (July 2003), 313–318. https://doi.org/10.1145/882262.882269
13. Zhang, L., Wen, T. and Shi, J. "Deep Image Blending," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 2020 pp. 231-240. doi: 10.1109/WACV45572.2020.9093632