ONE-PASS AUTOMATIC IMAGE MATTING: ALPHA, DEPTH, AND HARMONIZATION

Dillon Gustafson, John Swenson, Max Hansen
Team s■n
{gust0785,swen0481,hans8439}@umn.edu

ABSTRACT

We propose an image matting project that explores a one-pass automatic result involving three aspects: converting a trimap-based transformer (MatteFormer) model into a trimap-free variant for ease of use, providing predicted depth output for occlusion-aware pasting, and including image harmonization of the final matte result against a target background for automatic recoloring. Our central hypothesis is that predicting depth and color together reduces common post-matting artifacts such as haloing. To achieve this, we have synthesized custom harmonization and matting models into a pipeline that considers predicted depths to create a coherent composite image. We were also able to compare performance of individual sections of the pipeline using the standard fMSE (foreground mean squared error), MSE (mean squared error of composite), PSNR (peak signal-to-noise ratio), and SSIM (structural similarity index measure) metrics. Our novelty lies in a single process that predicts alpha, depth, and a harmonized foreground while requiring no manual input from the user aside from optional parameters to tweak the output composite.

1 INTRODUCTION

Image matting is a technique in image processing that builds upon image segmentation (which divides an image into distinct regions or objects by classifying each pixel with a label corresponding to the object it makes up). Specifically, image matting determines the degree to which each pixel belongs to a single foreground object. This is in contrast to image segmentation, which assigns binary labels that declare a pixel part of either one object or another, with no blending. Image matting data is represented by an alpha map called a "matte", which, when multiplied or "masked" with each pixel value, ensures smoother anti-aliased contours for compositing applications, retaining wispy details like hair or fur at a level of quality surpassing that of traditional image segmentation tasks. This task is critical for photo editing and AR (augmented reality), allowing objects to be plausibly composited onto new backgrounds.

However, algorithms remain challenging because local color information alone is often ambiguous for properly detecting smoothed alpha matte edges. Deep models significantly improved accuracy using CNNs (Convolutional Neural Networks) and transformers (Xu et al., 2017; Lu et al., 2019; Park et al., 2022), but typical methods still assume extra inputs from the user called "trimaps" composed of three components: known foreground pixels, known background pixels, and remaining unknown pixels to be filled in by the model. Such data is usually presented as an extra manual step in processing an image's matte, e.g. through mouse-clicks specifying foreground regions, which is often cumbersome to supply, especially in video matting applications. Additionally, such methods often ignore future requirements of the matte result when placed onto new backgrounds, such as proper occlusion by nearer objects or appearance matching. In recent years, so-called automatic image matting methods have been proposed, which remove the trimap requirement (Ke et al., 2022; Li et al., 2021), however, these models tend to yield lower-quality mattes with more frequent connectivity issues and artifacting. Separately, various image harmonization and depth-detection models exist that could allow for a single-pass automatic solution when combined with image matting (Cong et al., 2020; Guo et al., 2021).

Our solution is to combine matting with a second task that ensures final composites look convincing without any manual inputs aside from optional parameters like foreground depth shifts. As such, we accomplished three sub-projects: (A) Removed the trimap dependence from a pre-existing transformer matting model MatteFormer; (B) predicted alpha mattes along with a depth pass stored in the final result so that pasted foregrounds could be correctly occluded by background scene objects; and (C) predicted a harmonized and recolored foreground alongside the alpha and depth ensuring lighting is consistent with new backgrounds.

2 Related Works

Modern exploration of improvements to image matting has focused on two primary pillars, namely, trimap-based matting and automatic trimap-free methods. For the former problem, deep learning models have shown promise for improved image matting implementations. Xu et al. (2017) tackle this by dilating ground truth alpha mattes to generate trimaps for training. This training set is fed into an encoder-decoder deep learning network that trains against the combined alpha-prediction and compositional loss. Noticeably, adding a convolutional post-processing network demonstrated sizable improvements in the resulting image quality, especially around fine details.

Another area of focus in image matting are trimap-free methods. MODNet, introduced by Ke et al. (2022), approaches this problem by splitting the matte network into three parts: semantic prediction, detail prediction, and semantic-detail fusion. This modern method generally performs the best among trimap-free methods, but still falls short compared to matting methods that use trimaps. Another trimap-free method in the matting field is HAttMatting++ (Qiao et al., 2022), an attention-driven encoder–decoder. By selectively fusing multiscale features, it preserves fine structures (e.g., hairlike strands, holes, semi-transparent areas) and mitigates background leakage, followed by progressive refinement of the alpha matte. Compared to DIM (Xu et al., 2017), IndexNet (Lu et al., 2019), and GCA (Li & Lu, 2020), it improves edge quality by explicitly guiding features with attention.

3 Methodology

(a) Original Foreground (b) Synthetic Composite (c) Composite Depth map

Figure 1: An initial attempt at creating a depth-aware dataset using Distinctions-646 as a base. Since this dataset is synthetic, the generated depth maps (such as 1c, created from 1b) tended to exhibit more artifacts, especially with foregrounds that had so-called "meticulous" structure including lots of fine edges, like the branches in 1a. Depths in these depth maps were often inaccurate due to a lack of contextual information.

To construct our combined matting, depth, and harmonization pipeline, we first needed a dataset with combined ground-truth alpha mattes, depth map, and harmonization supervision. However, to the best of our knowledge, no such joint dataset exists. Combined alpha and depth datasets are available, such as JXNU-RGBD (Li et al., 2024) and HDM-2K (Chen et al., 2023), but they only include human foregrounds and are usually specialized for Background Matting (BGM), where image/video matting is performed against a fixed background – such datasets are better tailored for use-cases like automatic matting in video-conferencing scenarios, as opposed to the general-purpose image editing applications we target.

In creating our own dataset, we can create a script to apply stronger "supervisor" models for monocular depth prediction and image harmonization on an existing image matting dataset. This way, we

(a) Original Image (b) Depth Map (Full) (c) Alpha Matte (d) Depth Map + Matte

Figure 2: Depth map data acquisition. Our script produces a full depth map 2b from an original image 2a, before applying the ground-truth alpha matte 2c to produce a matted foreground depth map. 2d shows the matted depth map after depth inversion for clarity. Note that our script produces this data as raw `.npy` files, and the depth maps above are normalized for visualization.

can generate depth map and harmonization data that approximates ground truths, which is satisfactory for general use. For these depth maps, we used DepthPro (Bochkovskii et al., 2024) to generate pseudo-depths – the predicted depths are "metric", meaning that they are absolute rather than relative to their surroundings, which is useful for compositing images into depth-aware backgrounds.

At first, we used Distinctions-646 (Qiao et al., 2022), a synthetic dataset which pastes foregrounds onto random backgrounds using the ground-truth foreground mattes. However, this led to depth maps with considerable artifacts as can be seen in Figure 1c. The problem laid in the fact that DepthPro was originally trained on real-world data, as opposed to artificial composites, and thus generalized poorly to these cases. Therefore, we switched to AIM-500's real-world matting dataset (Li et al., 2021), applying depth maps to original images with their backgrounds intact to get contextual information necessary for proper depth predictions, and then masking these images with their ground-truth alpha mattes. Figure2 shows this process taken by the script used to generate our dataset.

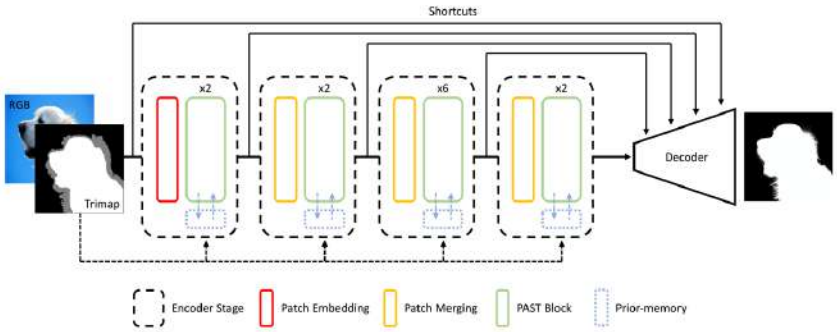
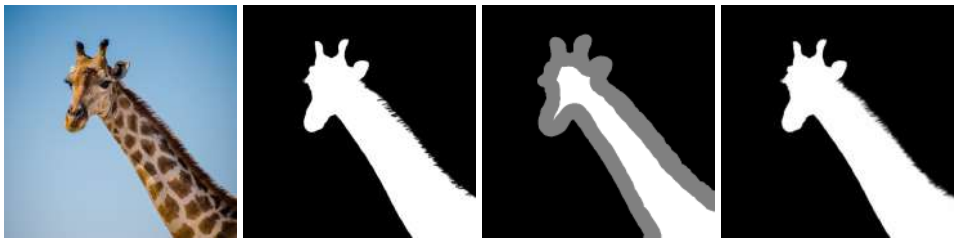


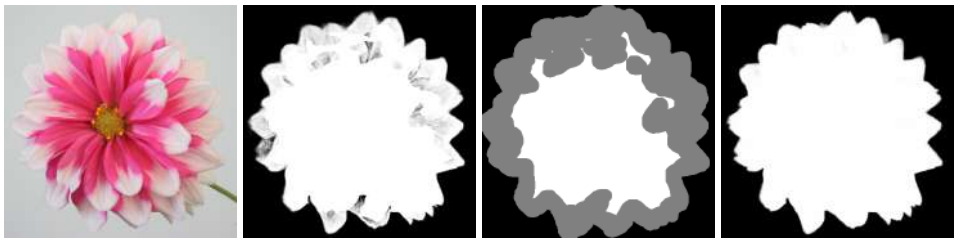
Figure 3: Overview of MatteFormer network (Park et al., 2022)

In order to attain comparable accuracy to trimap-based methods while also removing the difficulties of requiring a trimap input, we first trained an automatic trimap-free matting network based on Deep Automatic Natural Image Matting (AIM) (Li et al., 2021) in order to use it as a base with which to generate trimaps. This network, which we call MatteNet, predicts alpha mattes directly from RGB images. The model architecture uses a residual network (He et al., 2016) encoder-decoder architecture, where a backbone extracts multi-scale features which are then fused into unified representations via a Pyramid Pooling Module (PPM) coupled with semantic-enhancement (SE) blocks. These representations helped guide the matting head in predicting a trimap-free alpha matte directly from an RGB image. However, as can be seen in Figure 4, the initial alpha often exhibited gaps and connectivity issues, which were perfect candidates for further processing.

To refine the generated alpha mattes and create higher quality composites, we used the base prediction as input to an automatic trimap generation script, so that we could benefit from the higher performance found in trimap-based methods such as the MatteFormer transformer model (Park et al., 2022). Since the initial matte usually gives higher accuracy within inner pixels of the foreground and only suffers when dealing with fine detail present in the contours of the object, we generate trimaps



(a) Foreground input (b) Initial matte alpha (c) Generated trimap (d) MatteFormer alpha (refined)



(e) Foreground input (f) Initial matte alpha (g) Generated trimap (h) MatteFormer alpha (refined)

Figure 4: MatteFormer refinement of the initial matte alpha. The initial result is fed into a script that creates an "unknown" region (visualized in gray above) by dilating contours, allowing MatteFormer to further refine this area. Compared to the raw matte prediction, MatteFormer produces a cleaner alpha by reducing jagged edges (e.g. the hairs on the giraffe in 4b are significantly smoothed in 4d) and filling interior holes (e.g. the base flower matte in 4f shows spurious transparency where there is none in the original; this is fixed in 4h). The refined alpha is then used as the foreground mask passed into the compositing step.

Table 1: Quantitative comparison on our Distinction 646 trained MatteFormer model vs the pre-trained Composition-1k modelPark et al. (2022). Both are tested on Composition-1k's dataset.

Method	Gradient	Connectivity	MSE	SAD
Ours (Distinction-646 trained model)	43.03	70.60	0.021	58.73
Paper's (Composition-1k trained model)	31.00	50.08	0.014	50.40

that span these contours (dilating them by a set amount) to indicate 'low-confidence' regions that the MatteFormer should refine. The MatteFormer would then selectively improve these regions using its attention architecture.

MatteFormer uses an encoder-decoder network centered around the SWIN transformer (Liu et al., 2021). As depicted in figure 3, each stage includes a Prior-Attentive Swin Transformer (PAST), as well as prior-memory, which stores prior tokens. These prior tokens are used in conjunction with spatial-tokens in the PAST block, the structure of which is shown in figure 5. Combined with the network's post processing, a high retention of details can be achieved and some holes and fidelity issues in our MatteNet generated alphas can be fixed as seen in figure 4. This allows us to achieve greater accuracy while avoiding a trimap input from the user in our pipeline.

When training MatteFormer from scratch, the Distinctions-646 dataset demonstrated some issues with output as previously mentioned; although performance was acceptable, it struggled to resolve fine details compared to the pretrained model on Composition-1k. Comparing metrics using the Composition-1k test set also demonstrates comparable but slightly worse performance for the Distinction-646 model1. Thus, while both are usable in our final build, we decided to collect full pipeline data using the Composition-1k model.

To handle occlusions between the foreground cutout and the target background, we additionally estimated a depth map for background images using Depth Pro (Bochkovskii et al., 2024). After an

optional foreground depth shift Δ supplied by the user (positive pushes the foreground farther), we normalized depth maps per-image, performing the following soft visibility gate:

$$g = \sigma\big((\hat{D}_{\text{bg}} - (\hat{D}_{\text{fg}} + \Delta)) \cdot s\big) \tag{1}$$

where s controls boundary sharpness and smaller depth values correspond to closer geometry, and $\hat{D}_{\text{bg}}$ $\hat{D}_{\text{fg}}$ are the respective background and foreground depths. The effective foreground alpha would then become $\alpha_{\text{eff}} = \alpha \cdot g$, and the final depth-aware composite is produced with standard alpha blending using α_{eff}.

For harmonization, we intentionally avoided using the original matte α, since parts of the predicted foreground may be occluded after factoring in background depth. Instead, we passed a custom-trained harmonizer the composite image and the altered effective alpha α_{eff}, ensuring the model would only apply appearance corrections to the visible foreground regions. Our harmonization module is built from the DCCF-style pipeline (Xue et al., 2022), which predicts foreground appearance adjustments from masked composites. Using α_{eff} prevents the network from "harmonizing" pixels that do not actually contribute to the final image (i.e., pixels where the background is closer), improving stability and reducing color bleeding into occluded areas.

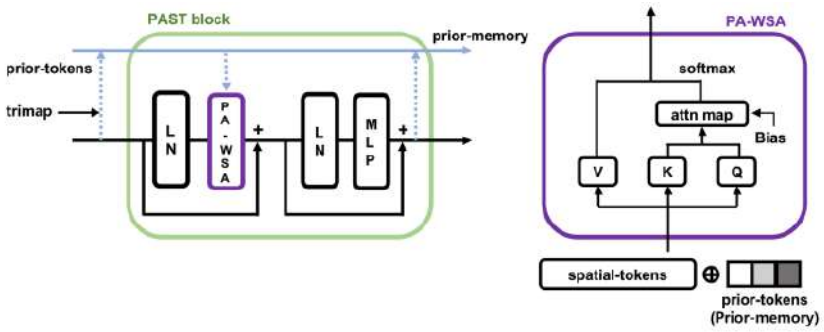


Figure 5: Overview of PAST (Prior-Attentive Swin Transformer) used in MatteFormer Park et al. (2022) (Park et al., 2022)

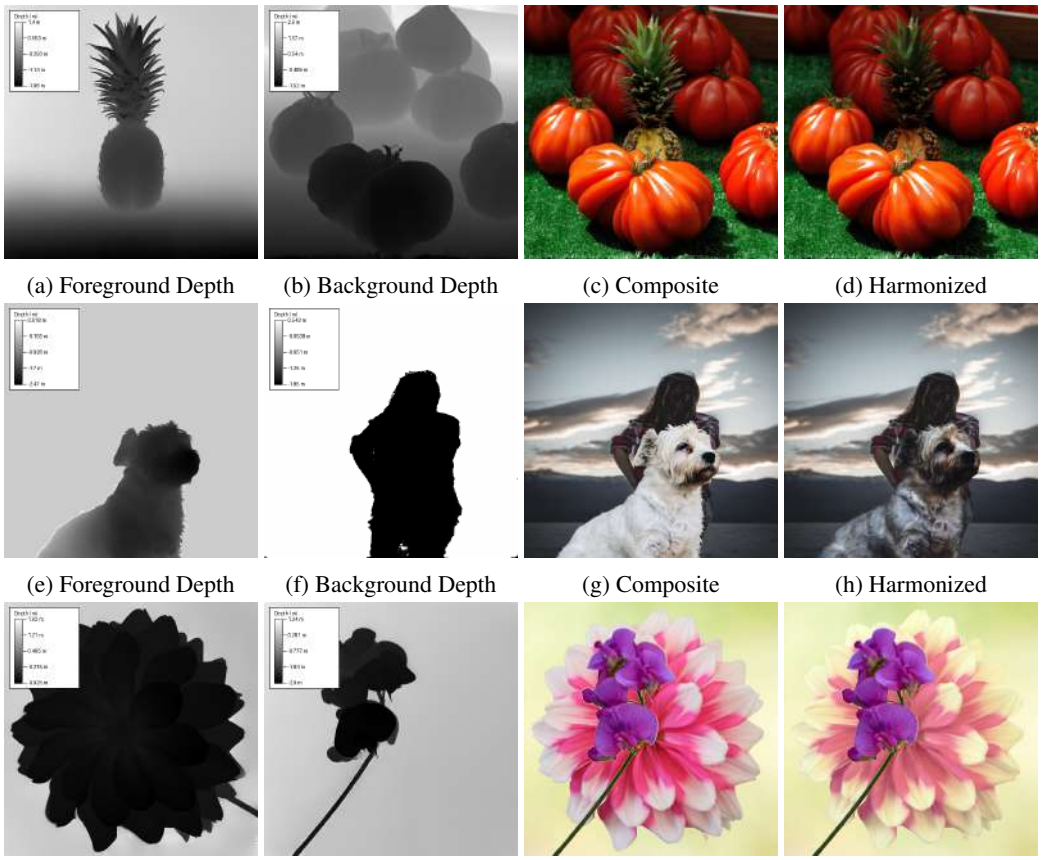
4 Results

A sampling of example final composites as well as intermediate steps in the pipeline may be seen in Figure 6. These examples showcase only the latter part of the whole pipeline, wherein depthmaps are inferred for foregrounds and backgrounds, and are used in a final composite and harmonization result. In Figure 6d the pineapple is convincingly placed behind the front pumpkin, but still rests in front of the other pumpkins in the back. In 6h, the program determines that the dog is closer to the camera than the girl, and applies harmonization that implies a twilight ambiance. In 7, we are also able to place the hollow ball above the sea, giving an effect akin to a large planetary body on the horizon.

To evaluate our results, we performed an ablation study which served to isolate the contributions of the MatteFormer-based matte refinement, DepthPro-based occlusion reasoning during compositing, and our occlusion-aware mask for harmonization. We evaluated on the iHarmony4 HCOCO test split at 512×512 resolution using standard harmonization metrics computed over the whole image (MSE, PSNR, SSIM) and over the foreground region only (fMSE). These results may be seen in Table 2. As can be seen, each extra step in our pipeline served to improve overall metrics, lending credence to their inclusion in the process.

5 Discussion

Notably, there is nothing stopping our program from creating a composite that could not exist in the real world (e.g. rigid objects that occupy the same position in space simultaneously). For future work, we could try to resolve these issues by including a step in the pipeline where a model could



(a) Foreground Depth (b) Background Depth (c) Composite (d) Harmonized

(e) Foreground Depth (f) Background Depth (g) Composite (h) Harmonized

(i) Foreground Depth (j) Background Depth (k) Composite (l) Harmonized

Figure 6: Depth and composite steps of the pipeline. After mattes are generated using the models and process in Figure 4, accompanying depthmaps are created for the foreground and background using DepthPro. Using the combined depth and mattes, a composite is created, and the alpha matte of the foreground is updated to exclude areas where the background occludes it. Finally, this updated alpha matte is used as the region to apply the custom harmonization model. Depth shifts used per row: 1.8, 0.0, 1.0.

Method Variant	fMSE ↓	MSE ↓	PSNR ↑	SSIM ↑
(A) Base: MatteNet only + naive alpha composite + harmonize w/ α	58.9	20.1	25.1	0.918
(B) A + extra MatteFormer refinement step	54.3	18.1	25.5	0.927
(C) B using soft alpha composite with sigmoid instead of naive	50.7	18.8	25.9	0.923
(D) C using harmonize w/ α_{eff} instead of α	**47.9**	**17.4**	**26.2**	**0.931**

Table 2: Ablation study on iHarmony4 HCOCO dataset. α_{eff} denotes the occlusion-aware foreground mask that ignores areas where the background occludes the foreground, as opposed to α which is not depth-aware. fMSE is computed over the foreground region only.

produce a determination as to whether or not a composite adheres to a real-world environment and adjust the depth shift amount of the foreground accordingly until it finds a space in which the foreground can exist in a convincing manner.

However, for creative applications it can be desirable to allow these "impossible" composites. When the depth ordering is contradictory, letting the foreground partially blend through the background can produce a merging of the images, which can look believable for foliage backgrounds as seen in Figure 9, or allow for humorous scenarios, such as the dog popping out from underneath sea coral in Figure 8. At any rate, if objects intersect slightly, we would prefer that their composite is still smooth – at first, we implemented a naive composite algorithm that performed strict occlusion which would

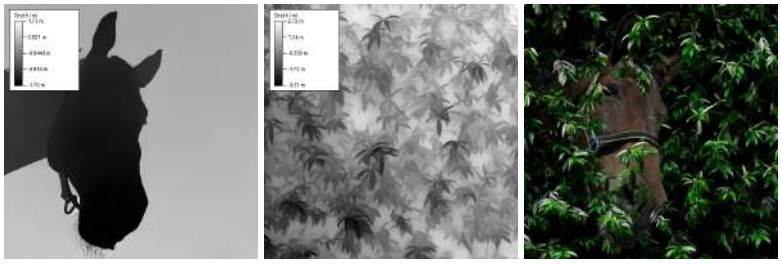
(a) Hollow Ball Foreground (b) Fisherman Background (c) Composite result

Figure 7: Full pipeline results including harmonization. Note the positioning of the ball on the horizon behind the sea in 7c, with parts of the sky visible through the holes of the ball. The algorithm shows some difficulty with maintaining transparency in backgrounds, as can be seen in the man's hair.

(a) Foreground Depth (b) Background Depth (c) Strict Occlusion (d) Smooth Composite

Figure 8: Comparison of strict vs. relaxed occlusion algorithms for "impossible" composites. Above, the model determines that the dog's body is at the same depth as the coral, but in the strict occlusion case, this leads to harsh boundaries at the bottom of the dog's body. Sigmoid smoothing means that the dog merges into the coral. A depth shift of 1.5 was used.

simply choose the closer object for every pixel in the image, seen in Figure 8c. However, this ended up leading to noticeable jagged boundaries. Our final compositing equation (as seen in Equation 1) used a sigmoid curve, which smoothed the composite as in 8d to get a better result.



(a) Foreground Depth (b) Background Depth (c) Smooth Composite

Figure 9: Smooth compositing algorithm can give a believable result for certain backgrounds such as foliage. Depth shift of 1.0 was used.

Difficulties in setting up and running the matting pipeline (based on AIM) have been logistical rather than conceptual: accessing MSI GPUs introduced queuing delays that slowed iteration, especially when testing multiple configurations. Our alternate solution available was to use a powerful local machine that one of our group members owns. In addition, we did not have access to the Composition-1K training set used in the original AIM training pipeline, preventing a clean reproduction of the full training regime. As such, we retrained the base matte model using AIM-500

as a substitute for the unavailable Composition-1k training set. As illustrated in 4, MatteFormer improves the initial matte alpha by filling interior holes and producing a more spatially consistent matte while better preserving delicate boundaries. Using the refined alpha rather than the raw MatteNet prediction makes the subsequent harmonization and final compositing steps substantially more stable and visually coherent. Matching the ratio of iterations to images in the set with the paper's iteration count resulted in 140,000 iterations, which took 32 hours to train running on Dillon's RTX 3090.

6 Conclusion

We presented a one-pass automatic matting pipeline that combines trimap-free alpha estimation, depth prediction for occlusion-aware compositing, and foreground harmonization. By using an AIM-based matting network to generate an initial alpha (and a derived trimap for refinement), we reduced user input while still benefiting from MatteFormer's strong edge and fine-structure reconstruction. We also showed that adding depth enables more realistic placement of foreground objects within depth-aware backgrounds. Our results demonstrate that the prediction of alpha and depth is feasible. Qualitative comparisons indicate that MatteFormer refinement substantially stabilizes downstream harmonization and reduces artifacts that would otherwise be amplified in the final composite. Overall, the system moves toward an end-to-end matting workflow that is more automated and compositing-aware than standard alpha-only pipelines. Future work could include a step in the composition process that accounts for foreshortening, since currently the foreground and background images are not scaled at all in the process, and so the foreground may appear much larger or smaller than it should in the final result, leading to unconvincing environments in some cases.

Work Done by Each Member + Things Learned

John: Set up overall pipeline, trained the base matting network, added depth map generation, soft-compositing step, and trained the image harmonizer. Learned how to train deep networks with PyTorch from scratch.
Dillon: Implemented MatteFormer with both pretrained model and self trained mode. Ran local training and tests on RTX-3090 to avoid time and resource limitations of MSI access. Modified MatteFormer to create a trimap from AIM to use for alpha matte generation. Integrated MatteFormer into the overall pipeline. Learned methodology for training transformer networks and details important to image matting which was really cool!. Also learned natural vs synthetic training sets can make a noticable difference.
Max: Implemented and trained AIM on MSI GPU nodes using multiple datasets to supplement Comp-1K. Generated alpha mattes, assembled composite test images for evaluation, and compared AIM results against MatteFormer. Learned that synthetic composites generalize poorly compared to natural-image training (missing context), and that simple architecture tweaks can reduce artifacting.

Acknowledgments

Utilities for the image harmonization model use some code from DCCF (Xue et al., 2022) which is present in the 'iharm/' folder on GitHub.

Some MatteFormer utilities present in 'utils/' and 'networks/' on GitHub come from the original MatteFormer code (Park et al., 2022).

References

Alexey Bochkovskii et al. Sharp monocular metric depth in less than a second (depth pro). *arXiv:2410.02073*, 2024. URL https://arxiv.org/abs/2410.02073.

Xi Chen, Jiayi Song, Yabin Chen, Dongbo Min, Qing Lin, and Zhiqiang Sun. HDM-2K: A high-resolution RGB-D human matting dataset. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3238580.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *CVPR*, 2020.

Zongyu Guo et al. Image harmonization with transformer. In *ICCV*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W. H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition, 2022. URL https://arxiv.org/abs/2011.11961.

Haotian Li, Guobiao Jiang, Yang Li, Yuming Lu, Binren Li, Zhongqi Zhang, and Qi Wang. Depth-enhanced accurate and real-time background matting. *arXiv preprint arXiv:2402.15820*, 2024.

Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting, 2021. URL https://arxiv.org/abs/2107.07235.

Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention, 2020. URL https://arxiv.org/abs/2001.04069.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting, 2019. URL https://arxiv.org/abs/1908.00672.

GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens, 2022. URL https://arxiv.org/abs/2203.15662.

Yu Qiao, Yuhao Liu, Ziqi Wei, Yuxin Wang, Qiang Cai, Guofeng Zhang, and Xin Yang. Hierarchical and progressive image matting, 2022. URL https://arxiv.org/abs/2210.06906.

Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting, 2017. URL https://arxiv.org/abs/1703.03872.

Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, volume 13667 of *Lecture Notes in Computer Science*, pp. 300–316. Springer, 2022. doi: 10.1007/978-3-031-20071-7_18. URL https://link.springer.com/chapter/10.1007/978-3-031-20071-7_18.