SUPPLEMENTARY MATERIAL
CAPTURE STAGE ENVIRONMENTS: A GUIDE TO BETTER MATTING

In this supplementary material, we provide additional details regarding the hardware setup of our capture stage, as well as further training specifics. We also include qualitative results to illustrate our findings more clearly, along with visual outcomes from the downstream NeRF application. Furthermore, we present experimental evaluations of how effectively the student model learns from the teacher model and discuss why a student–teacher framework is preferable to directly training on scribble data. Lastly, we discuss the limitations of our approach.

1. HARDWARE AND IMPLEMENTATION

In the following, we discuss our capture stage setup and provide additional details on our dataset, the applied model, and the training.

1.1. Capture Stage Setup

Our capture stage features 40 24.5 megapixel RGB cameras and 46 video lights affixed to cylindrical scaffolding with a diameter of 5.25 meters. The cameras are capable of capturing 75 frames per second with a sensor depth of 12 bits. Subjects in the stage are lit by all video lights and captured by all cameras simultaneously. The brightness of the lights in the stage can be controlled, and the camera positions are static, therefore the backgrounds of the cameras do not change and can be pre-captured.

1.2. Dataset, Models, and Training Details

Our approach uses three different models. A large offline *teacher* model, a lightweight real-time *student* model and a *supervisior* model for validation. As *teacher*, we developed a modified version of *ViTMatte-S* [1] as the offline teacher model, referred to as *BgViTMatte* in the following. This modification exploits the unique characteristics of capture stages by replacing the trimaps originally required by ViTMatte-S with background images, where we train the model in its new configuration on the Adobe Deep Image Matting dataset [2]. Please note that due to the presence of noise in our data, we augmented the training data with Gaussian noise with a standard deviation of up to 0.1 to improve stability. The impact of this augmentation is demonstrated in Fig. 1.

To address common failure cases in the teacher's predictions, we analysed a set of 41 resulting alpha masks that show recurring errors and annotated the respective regions. Given the annotated failure cases, the teacher model is refined on a hybrid dataset that combines the Adobe Deep Image Matting dataset [2] with the capture stage content along with their corresponding sparse annotations. During training, 80% of each batch is drawn from the former, while the remaining 20% are sampled from the latter data. The proportion of these two datasets has a significant influence on matting performance, as shown in Fig. 2, with metrics computed on the hold-out validation set. Notably, fine-tuning exclusively with failure annotations results in poor predictions. We trained BgViTMatte for 2000 iterations with a batch size of 16, using an initial learning rate of $5 \cdot 10^{-5}$, scheduled to $2.5 \cdot 10^{-5}$ after 600 iterations.

In a second phase, BgViTMatte generates high-quality ground truth masks for a new capture stage dataset comprising 20 scenes with 29 different camera-views each (580 images in

Fig. 1: Influence of noise in the input images on matting. The left image shows a matting result without introducing noise during training, whereas the right image shows the result with noise added to the augmentation phase during training.

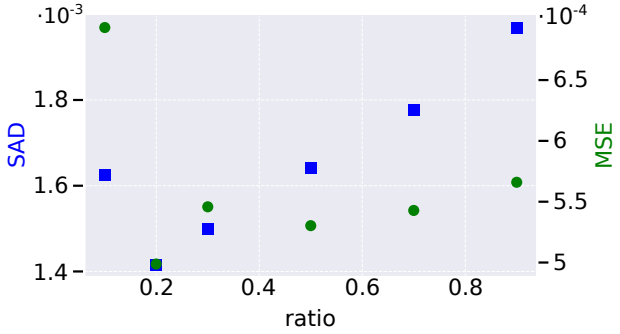
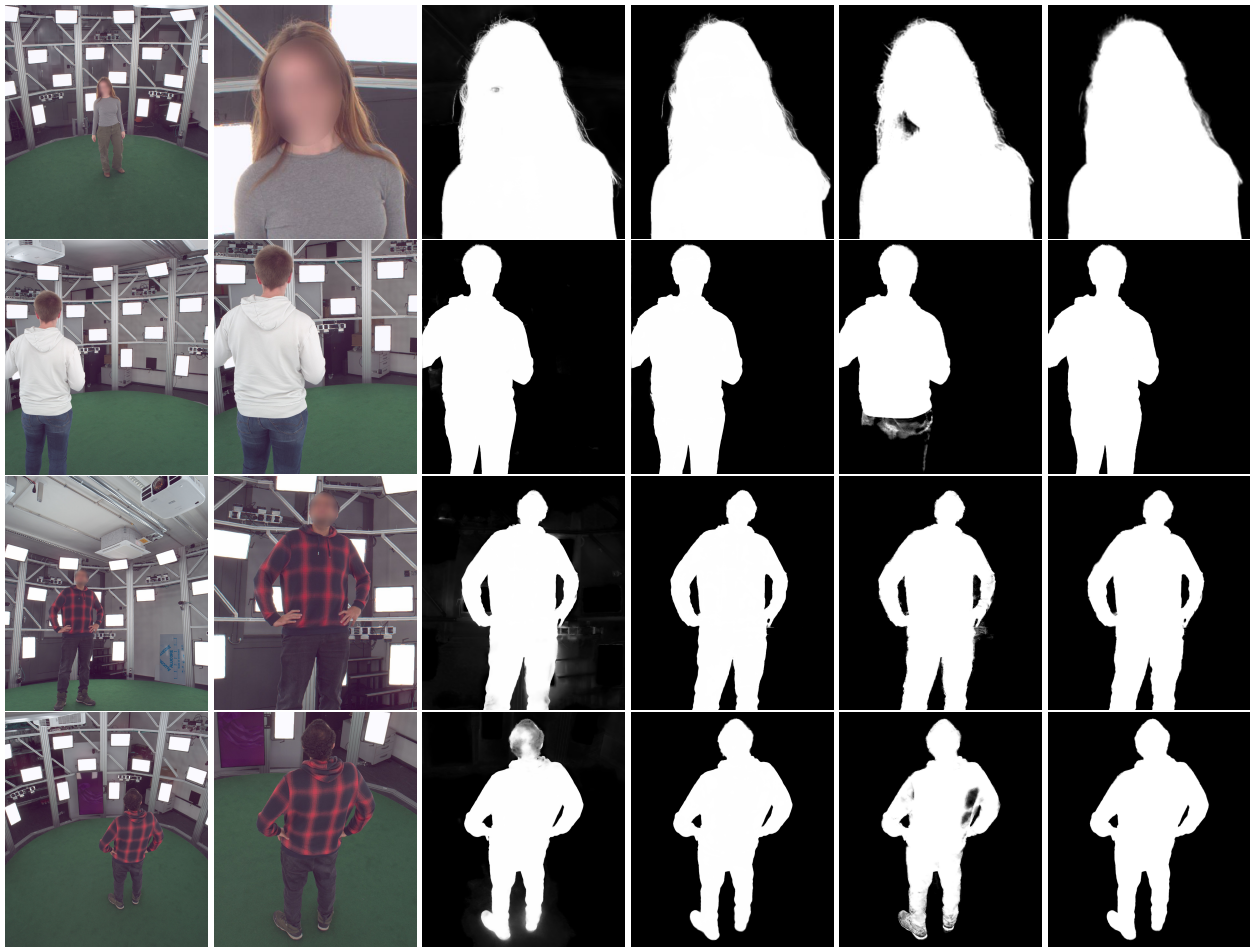


Fig. 2: The ratio of sparsely annotated data during training significantly affects performance. Training exclusively on scribbled data is not included in the plotted range, resulting in an MSE of $78 \cdot 10^{-4}$ and an SAD of $11 \cdot 10^{-3}$.



(a) Input Image (b) Zoomed-In (c) Teacher Base (d) Teacher Refined (e) Student Baseline (f) Student Refined

Fig. 3: Visual comparison of the outputs from the teacher and student models, predicted from the input image in (a). (b) provides a zoomed-in view of a region of interest. (c) and (d) present the outputs of the teacher model ViTMatte-S [1] in its baseline and refined configurations, respectively. Similarly, (e) and (f) show the outputs of the student model BGMV2 [3] in its baseline and fine-tuned versions.

total). These masks were used to train our real-time student model, based on *BackgroundMattingV2* (BGMV2) [3] with a MobileNetV2 backbone. BGMV2 employs a two-stage architecture: a base model that generates a coarse mask from a foreground and background image, and a second model that improves fine details along the mask's borders. We fine-tuned the base model for 5 epochs, and the combined model for an additional 10 epochs using the capture stage dataset and the generated ground truth masks. We have kept all other hyperparameters and losses unchanged, and refer to [3] for more details. As the supervisor model, we used *DiffMatte* [4] to generate a validation set of 14 high-resolution masks based on hand-drawn trimaps from various scenes and camera positions in the capture stage.

2. QUALITATIVE EVALUATION

To further analyse the impact of the task-specific refinements, we performed a qualitative comparison of both the teacher and student models before and after these refinements. From Fig. 3 we can observe that the teacher model tends to produce false positives in its base configuration, which is mitigated through the refinement. Meanwhile, the student model rather struggles to capture all foreground regions accurately in its base setup. Fine-tuning on the teacher-generated data, however, significantly improves its performance by more precisely outlining foreground elements and reducing errors.

3. DOWNSTREAM APPLICATION

A qualitative comparison of the reconstructed capture stage content by InstantNGP, utilizing alpha masks generated by

Fig. 4: Qualitative comparison of novel views from Instant-NGP using the original alpha masks generated by BGMV2 [3] (left) and our improved masks (right). The erroneous matting of the baseline not only cuts out part of the content but also results in an unfaithful color representation to compensate for the wrong inputs.

our proposed matting pipeline, is presented alongside results from the baseline model in Fig. 4.

4. LEARNING CAPACITY STUDENT MODEL

To assess the student model's learning capability, we additionally compared its predicted masks to those generated by the teacher model on a validation dataset of 519 images taken in the capture stage. Specifically, we evaluated the baseline student model BGMV2 [3] with its fine-tuned counterpart in terms of the sum of absolute difference (SAD), mean squared error (MSE), and spatial-gradient metric (Grad). Tab. 1 demonstrate the improved alignment between the student and teacher models after fine-tuning.

5. SCRIBBLE-BASED TRAINING OF THE STUDENT

Fine-tuning the real-time student model, BGMV2 [3], directly on scribble data is challenging, as the second stage of this two-stage model (designed for mask refinement) is inherently unsuited for training on sparse annotations. To address this, we limited fine-tuning on scribble data to the model's first stage. Initial experiments using a hybrid dataset composed of

	MSE↓ $\cdot 10^{-4}$	**SAD↓** $\cdot 10^{-3}$	**Grad↓** $\cdot 10^{-5}$
without fine-tuning	17.892	3.209	6.068
with fine-tuning	9.098	2.126	5.126

Table 1: Quantitative comparison of the student model's performance against the teacher model on validation data. The results highlight the alignment between the student and teacher models in terms of predicted mask quality.

Fig. 5: Potential failure cases of the matting prediction that can occur despite fine-tuning.

scribble annotations and the Adobe Image Matting Dataset, while keeping the second stage fix, resulted in an MSE of 89.2×10^{-4} and SAD of 10.4×10^{-3}. Subsequent fine-tuning of the mask refinement network using the Adobe Image Matting Dataset over several epochs improved performance, with an MSE of 27.8×10^{-4} and an SAD of 57.5×10^{-4}. However, this mask refinement process alone failed to fully exploit the potential of our student-teacher approach.

By integrating the student-teacher framework, the real-time model can train on highly detailed masks from the teacher model and can such produce high-quality outputs while maintaining real-time performance.

6. LIMITATIONS

While our method has shown improvements in matting quality, it does not always guarantee accurate alpha mask predictions due to the inherent limitations of learning-based approaches. See Fig. 5 for example failure cases of the fine-tuned BGMV2 model, showing mispredictions at object borders and an incorrect assignment of a foreground region appearing as a "hole" in the foreground.

7. REFERENCES

[1] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang, "Vitmatte: Boosting image matting with pre-trained plain vision transformers," *Information Fusion*, vol. 103, pp. 102091, 2024. 1, 2

[2] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang, "Deep image matting," in *CVPR*, 2017, pp. 2970–2979. 1

[3] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *CVPR*, 2021, pp. 8762–8771. 2, 3

[4] Yihan Hu, Yiheng Lin, Wei Wang, Yao Zhao, Yunchao Wei, and Humphrey Shi, "Diffusion for natural image matting," in *ECCV*. Springer, 2024, pp. 181–199. 2