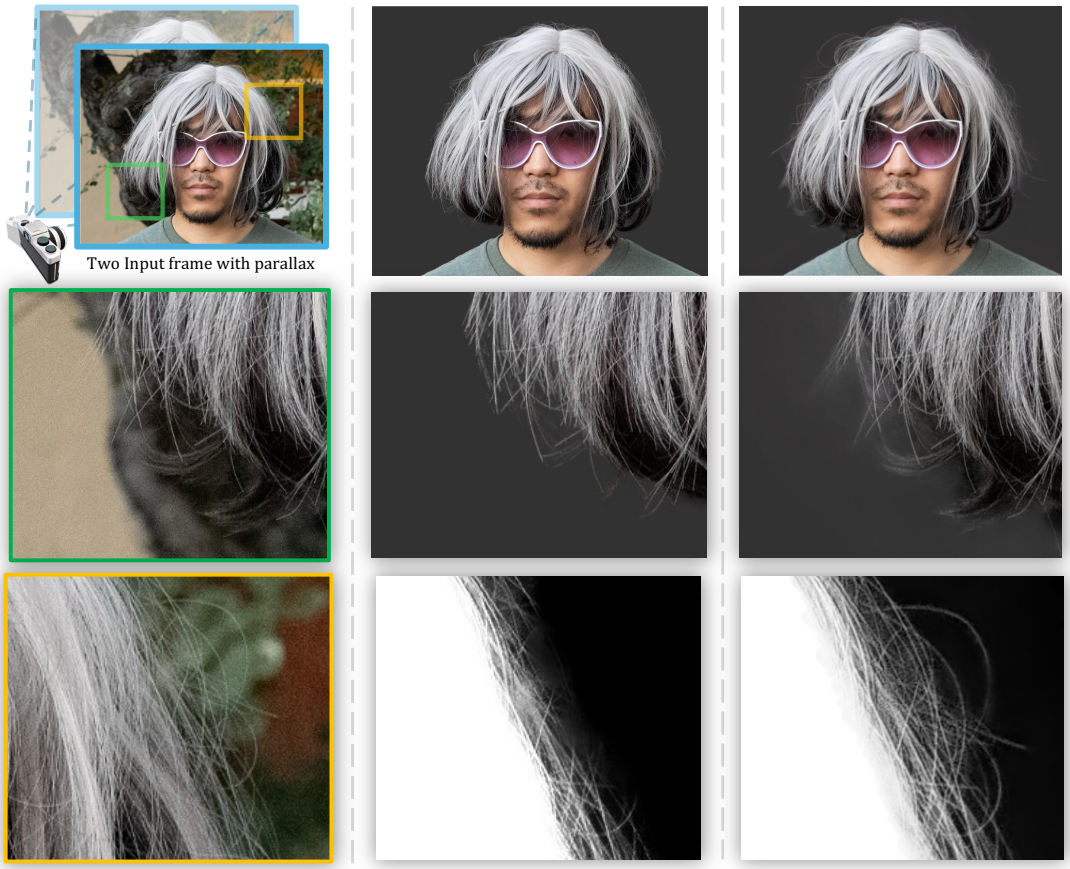
Parallax Portrait Matting

Xin Cai[13], Jiawen Chen[2], Lars Jebe[2], Tianfan Xue[134], Zhoutong Zhang[2]
[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]Adobe NextCam [3]Shanghai AI Laboratory [4]CPII under InnoHK
caixin025@gmail.com, tfxue@ie.cuhk.edu.hk, zhoutongz@adobe.com



(a) Base Frame from Input (b) Foreground & alpha, remove.bg (c) Foreground & alpha, ours

Figure 1. Our matting method exploits camera-motion-induced parallax between the foreground and the background. It takes two frames as input, each taken with a slightly different camera location, and predicts both a pre-multiplied foreground image and an alpha map. Trained on public datasets, our method produces a cleaner foreground with more details than closed-source commercial solutions like remove.bg.

Abstract

Image matting has long served as a critical piece for many editing tasks, such as re-composition, re-lighting, bokeh simulation, etc. Being an ill-posed problem, matting usually involves modeling priors of the foreground object, including both transmittance and color, and the background scene. Single-image matting methods aim to model those priors from data, but usually struggle with hard cases where both the foreground and background are highly textured. To get better results for those challenging cases, previous works usually rely on additional information, such as a green screen, polarized lighting, or an additional image of the scene without the foreground object. Providing accurate result, they usually require costly setup from the user. We therefore propose utilizing another source of signal, motion, to tackle cases that are intrinsically difficult for single-image methods. The motion signal comes naturally from the

user capturing a burst of images, where the camera moves due to hand tremor or the intention to make multiple takes to choose the best shot later, while the foreground subject remains relatively still. Such motion induces a slight parallax between the foreground and background, which gives us different observations of the same foreground object composed on slightly different regions of the background. We use this induced parallax to recover fine details and color of the foreground object, which are otherwise lost for state-of-the-art single-image matting models.

1. Introduction

Image matting has a long history in computer vision and graphics [9, 32]. This decades-old problem aims to decompose an image I into a foreground image F, a background image B, and an opacity (alpha) map α, where they jointly reconstruct the input image through a linear composition process:

$$I = \alpha F + (1 - \alpha)B.$$

Like most inverse problems, the matting problem is known to be ill-posed. To arrive at a solution representing the actual scene, it requires either priors over F, B, and α, or additional information as constraints to the solution space.

Works that utilize priors have made significant progress. With the success of neural networks, it is common to collect datasets with annotated alpha map α, or with an additional foreground layer F, to train neural networks to implicitly learn those priors [20, 28, 43, 45, 46]. However, it is hard to collect high-quality matting annotations at scale, which limits the performance of those learning-based models. To overcome the data limitation, matting by generation [40] resolves the ambiguity using the natural image priors captured by text-to-image diffusion models, but it significantly increases the computational cost.

To overcome the challenges of single-image matting, particularly in recovering fine details as shown in Fig. 1, researchers design different specialized setups to capture more information. The additional information may come for the usage of green screens [2], polarized light and lens filter [8], color filter array [3], camera arrays [12], focal stack [13], or the background image of the scene [30]. Although these methods achieve high-quality matting under those specialized setups, it often require dedicated workflow, making them infeasible for daily photo capturing.

In this work, we propose a novel parallax matting method feasible for daily capture, which only requires users to capture another image under a slightly different view. This naturally occurs in cases where the user is already capturing multiple photos, which is typical for intentional photographers, or for mobile users where the camera already takes multiple images under the hood [21]. This additional information incurs little overhead for capturing but provides extra conditioning information for the matting problem, especially for challenging cases where fine details are overlaid over a complex background. Intuitively, assuming we can reliably model the foreground and background motion, the matting equation solutions then need to explain two instead of one image, making the problem better conditioned.

However, robustly estimating foreground and background motion is not trivial. Off-the-shelf motion estimation methods, such as optical flow estimators [37, 42], can not predict reliable per-pixel motion for regions where the foreground and background are mixed. To tackle that, we assume the apparent motion in those regions is a mixture of two smooth motion fields, each of which we can reliably estimate using their neighbors that are not part of the mixture. This is equivalent to assuming the scene is almost static, which only holds true approximately. Though the capture time between two frames is just a few seconds, the scene is not completely static due to wind or subject movement. We therefore design our model to be robust against motion estimation errors: it only utilizes motion when it is helpful and falls back gracefully to single image estimates when motion estimation fails.

To achieve the robust matting, our model takes local patches from both images and their warped versions as input. The model, designed symmetrically, utilizes cross-attention to pick up helpful correspondences, inspired by recent works in 3D computer vision [38]. We train the model with image and motion augmentations to make it robust to image and motion degradations. We show that such a design and training strategy can achieve better matting and color separation results than all single-image matting models, and with better details than closed-source commercial single-image matting solutions, especially on challenging cases. This demonstrates that our method is a complementary matting solution to single-image matting methods under those scenarios, if an additional frame is available.

In summary, our contributions are threefold:

- We propose utilizing parallax information between foreground and background for challenging portrait matting scenarios, where foreground details are mixed with complex backgrounds.
- We developed a framework that robustly uses parallax for joint alpha and foreground estimation under those cases.
- We showed empirical results in which the use of two frames can result in better results than single-image models for challenging cases, and its robustness towards motion estimation errors.

2. Related Works

Single Image Matting. Most existing work aims to predict alpha and/or foreground color from a single input image. Many methods use additional guidance signals such as semantic segmentation [20, 44, 46], instance segmentation [11, 34] a trimap [10, 26, 28, 36] or different types

of user annotations [18, 25]. More recent learning-based methods take as input a single image without any guidance, either by implicitly incorporating guidance signal prediction [5, 14, 19, 31, 45], or by leveraging strong learned priors, e.g., from generative models [40]. Since the matting problem is intrinsically ill-posed, those methods usually fail to recover fine details if the background is highly textured.

Video matting. Video matting methods aim to predict alpha and/or decontaminated color for an entire sequence. Most learning-based method extend beyond current single image matting model with temporal feature aggregation designs, such as graph neural networks [39], temporal RNNs [23, 24], transformers [17], deformable convolution [33] or temporal image difference [35]. All those method aim to learn such feature aggregation end-to-end with video data supervision, and aim to make single image matting consistent over the entire video sequence. Most similar to ours is [6], where the authors use optical flow to correlate frames in the video to better estimate the background, then optimize for per-frame alpha maps. Our method differs from video matting method where we focus on utilizing motion between frames to improve single-image matting results over challenging cases.

Matting with additional signals. Additional information helps better condition the matting problem. Background matting [22, 30] uses an additional background image as conditioning and is able to produce high quality matting for images and videos. Polarization systems [8] are able to capture ground truth transmittance maps. Attaching a color filter [3] to a camera lens allows one to simultaneously capture multiple views from a single exposure, where stereo algorithms can provide richer details than traditional matting models. One can also achieve the same effect with a camera array [12]. In addition, focal stacks [13] can be used to provide additional information since foreground and background are blurred differently. Our method differs from those methods where we do not ask the user to perform additional setups before capturing; we simple require an additional image.

3. Formulation & Assumptions

We first describe the problem formulation and examine how the parallax between the foreground and background can better condition the matting equation. We then introduce the assumptions made by our formulation and discuss how realistic they are. Finally, we discuss the building blocks of our matting pipeline which includes foreground and background motion estimation, trimap generation, and the design of our matting prediction network.

For a single image $\mathbf{I}_1$, the matting problem tries to decompose it into a foreground image $\mathbf{F}_1$, a background image $\mathbf{B}_1$ and an alpha map α_1 that encodes the opacity of the foreground image, where:

$$\mathbf{I}_1 = \alpha_1 \mathbf{F}_1 + (1 - \alpha_1)\mathbf{B}_1.$$

This linear system is underdetermined, with only one constraint but three unknowns (in grayscale; in color it has three constraints and seven unknowns). Unless we impose priors over each of the unknowns, there are an infinite number of solutions that all satisfy the equation.

Now assume that we capture another frame $\mathbf{I}_0$ which differs from $\mathbf{I}_1$ due to parallax between the foreground object and the background. Let us describe this parallax with two motion fields $M^F_{1\to0}(\cdot)$ and $M^B_{1\to0}(\cdot)$, where $M^F_{1\to0}(\cdot)$ is the warping function for the foreground motion and $M^B_{1\to0}(\cdot)$ is the background motion. The matting equation for $\mathbf{I}_0$ is:

$$\mathbf{I}_0 = \alpha_0 \mathbf{F}_0 + (1 - \alpha_0)\mathbf{B}_0,$$

we can correlate α_0, $\mathbf{F}_0$ and $\mathbf{B}_0$ with α_1, $\mathbf{F}_1$ and $\mathbf{B}_1$ by:

$$\alpha_0 = M^F_{1\to0}(\alpha_1), \mathbf{F}_0 = M^F_{1\to0}(\mathbf{F}_1), \mathbf{B}_0 = M^B_{1\to0}(\mathbf{B}_1).$$

Substituting into the equation above, we have:

$$\mathbf{I}_0 = M^F_{1\to0}(\alpha_1 \mathbf{F}_1) + M^F_{1\to0}(1 - \alpha_1) M^B_{1\to0}(\mathbf{B}_1).$$

The objective of burst matting is to solve for the foreground $\mathbf{F}_1$ and alpha map α_1 of the base frame, given the additional information provided by the alternate frame $\mathbf{I}_0$,

The key idea for parallax matting is that, given a correct motion estimate ($M^F_{1\to0}$ and $M^B_{1\to0}$), the extra frame we observe serves as another constraint on the *same* unknowns we want to estimate. Note that the constraint is only useful if the parallax between foreground and background exists. At pixels where $M^F_{1\to0}$ is the same as $M^B_{1\to0}$, the equations are linearly dependent. This assumption in turn imposes some mild conditions on the scene and capture process.

Parallax between foreground and background. We aim to extract the foreground, which is by definition closer to the camera than the background. In other words, there should be sufficient motion parallax between the two layers.

Mostly static scene. The second assumption is that the scene is almost static, so that we can model parallax simply through two warping fields, one for the foreground and one for the background. Our method robustly handle this assumption by being robust to motion estimation errors.

Consistent camera settings. Finally, we assume that both frames are captured using the same settings (exposure, white balance, color, tone, etc). Most cameras feature an auto-exposure-and-lock function, which we use when capturing our dataset. In our experiments, we capture raw images and render them with the same parameters in Adobe Lightroom to ensure maximum consistency.

4. Method

Our method starts with a trimap estimation stage for each input image, which partitions an image into three regions: foreground, background, and unknown. We then estimate the background motion $M^B_{1\to0}$ and foreground motion

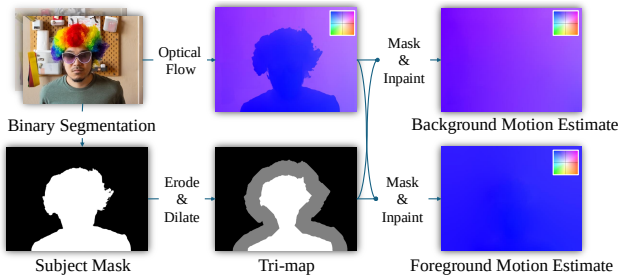


Figure 2. Our motion estimation pipeline. Given two frames, we first estimate optical flow between the two, together with a trimap generated as described in Sec. 4.1. Since flow estimation in overlapping areas is often incorrect, we inpaint them by their nearest neighbor in non-overlapping regions.

$M^F_{1\to0}$ using the partitions from the trimap. Then we input the background and foreground-aligned image pairs to predict the alpha and foreground colors. In our experiments, we find that estimating background motion $M^B_{1\to0}$ is easier and more robust than estimating foreground motion $M^F_{1\to0}$. $M^B_{1\to0}$ is usually smooth everywhere, but $M^F_{1\to0}$ often is not, especially for small details like strands of hair moving due to wind or small motion of the subject. We specifically design our prediction network to deal with such situations by adopting a symmetric design and utilizing $M^F_{0\to1}$ through a cross-attention mechanism in the feature space.

4.1. Trimap Estimation

Following most prior work in image matting, we first create trimaps. Starting with a state-of-the-art dichotomous segmentation network, BiRefNet [48] that generates a binary foreground mask, we erode and dilate it 100 times to produce foreground and background masks, respectively (see Fig. 2). This is based on the assumption that the predicted binary segmentation mask has an error margin of 200 pixels, which is reasonable given the high accuracy of the latest segmentation models.

4.2. Motion Estimation

Given the trimap, we proceed to estimate the motion fields $M^B_{0\to1}$ and $M^F_{0\to1}$. Due to complex occlusions between foreground and background in the uncertain region, separately estimating motion for foreground and background is very hard. To circumvent this challenge, we follow the common assumption [4] that motion for both the foreground and the background is locally smooth, and therefore motion in the uncertain region can be estimated by extrapolating motion estimates from certain regions. Specifically, we first estimate optical flow between two images with an off-the-shelf method such as GMFlow[42]. To handle occlusion, we simply replace the estimated motion in uncertain regions with the values of its nearest neighbor inside the certain region, both for the foreground and background. Fig. 2 shows an example of this motion estimation process.

Figure 3. An example of foreground and background alignment. Background-aligned patches only contain foreground subject motion, and the dis-occluded background in the alternate frame provides more context. Foreground-aligned patches fixate on the subject, which directly helps in separating the foreground.

With these motion estimates, we can more intuitively see how they are helpful for the matting problem. For example, if we warp the image $\mathbf{I}_0$ with the background motion $M^B_{0\to1}$, we get an image $\mathbf{I}^B_{0\to1}$ that differs only from $\mathbf{I}_1$ in foreground regions. Regions that are originally occluded might become dis-occluded and therefore provide a strong hint on what the occluded background is. If we warp $\mathbf{I}_0$ using $M^F_{0\to1}$, then we get an image $\mathbf{I}^F_{0\to1}$ where the foreground stays put and the background has shifted, providing a strong signal on what the foreground object is. Fig. 3 shows an illustration of this intuitive result. Therefore, we would like the network to utilize such motion information by looking at warped frames using both the foreground flow and the background flow. However, reliably and robustly doing so requires a specialized design.

4.3. Foreground and Alpha Estimation

As shown in the previous section, warping the image using foreground and background motion reveals the foreground and background parallax. It is then tempting to directly train a neural network that takes the tuple $(\mathbf{I}_1, M^F_{0\to1}(\mathbf{I}_0), M^B_{0\to1}(\mathbf{I}_0))$ as input and predict the alpha map α_1 and the foreground color F_1 directly. However, we empirically find that this design fails to deliver good results. The reason is that this naïve design cannot effectively learn to compensate for motion estimation errors, especially for the foreground. For real-world images, foreground motion estimates are often inaccurate, while background motion estimation is more reliable. The background is usually far away from the camera and camera motion is small, so the background motion can usually be modeled by a simple smooth field such as a homography. Foreground motion, however, is hard to approximate parametrically, due to the locally inconsistent movement of the small features like hair. To improve the accuracy of foreground motion, one possible solution is to train on synthetic data with ground truth motion fields, but this often generalizes poorly to real

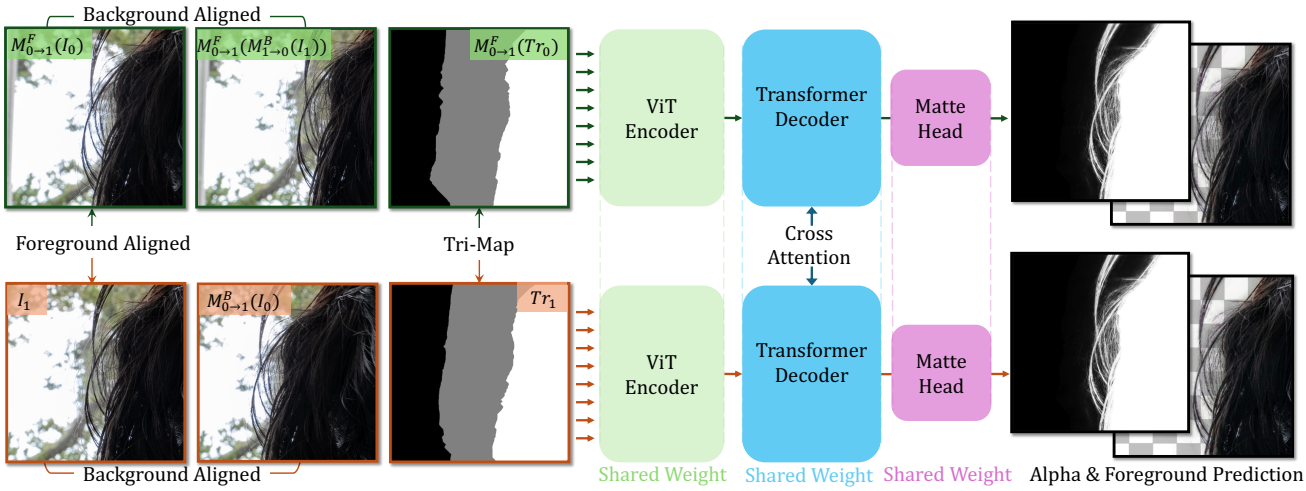


Figure 4. Our model framework. We process background-aligned image patches and trimaps through two symmetric branches (top and bottom). Each branch independently estimates foreground information using background-aligned pairs. Cross-attention between branches corrects the foreground information from another foreground-aligned branch by exchanging foreground features.

world inputs with different distributions.

The difficulty of motion estimation motivates us to design a network that can implicitly compensate for foreground motion errors. We learns multi-view correspondences through a cross-attention mechanism between two streams, and use cross attention to compensate for foreground motion errors. Our framework employs a symmetric dual-branch architecture with weight-shared encoder-decoder networks. Each branch aims to predict both the foreground color and the alpha map, taking two background-aligned images and the corresponding trimap as input. One branch (the bottom branch of Fig. 4) takes the base image I_1, a background-aligned image from another frame $\mathbf{I}^B_{0\to1} = M^B_{0\to1}(\mathbf{I}_0)$, and the trimap corresponding to the base $\mathbf{Tr}_1$ as input. This branch aims to predict the foreground image F_1 and alpha map α_1 of the base frame. The other branch (the top branch) takes a similar input that is warpped by the estimated foreground motion:$\mathbf{I}^F_{0\to1} = M^F_{0\to1}(\mathbf{I}_0)$, a frame where background is aligned with $\mathbf{I}^F_{0\to1}$, which is $M^F_{0\to1}(M^B_{1\to0}(\mathbf{I}_1))$, and a warped trimap $M^F_{0\to1}(\mathbf{Tr}_0)$. This branch predicts the corresponding alpha and the foreground of $\mathbf{I}^B_{0\to1}$. During decoding, cross-attention layers facilitate feature exchange between branches and adaptively select the most relevant features to mitigate matting ambiguities. This mechanism can compensate for simple motion estimation errors by learning reliable inter-frame correspondences. Moreover, our design is crafted to primarily learn robust single-frame matting through a dedicated branch while allowing the cross-attention mechanism to flexibly integrate complementary features from a second frame. The shared weights between branches further enforce a unified feature representation between single-frame and multi-frame inputs, making us learn the model with single image matting datasets. As a result, even when parallax is absent or significant motion errors occur, our model effectively defaults to stable single-frame matting, ensuring overall system robustness.

4.4. Training Objectives

Our training loss consists of an alpha loss and a pre-multiplied foreground color loss, described below.

Alpha Loss. To supervise the alpha prediction, we use an L_1 loss. However, in a given image, only a small set of pixels have an alpha value between 0 and 1, which biases the training towards modeling pixels that have an alpha value of 0 and 1. To mitigate this, we adaptively weight those pixels by normalizing them separately. Specifically, we define

$$\mathcal{L}_{\text{sep}} = \frac{1}{|S|}\sum_{x\in S}\left|\alpha[x] - \alpha^{\text{gt}}[x]\right| + \frac{1}{|H|}\sum_{x\in H}\left|\alpha[x] - \alpha^{\text{gt}}[x]\right|,$$

where $|S|$ denotes a set of pixels that are "soft", meaning they have a ground truth alpha value between 0 and 1, and $|H|$ denotes a set of pixels that are "hard" and have a ground truth alpha value of exactly 0 or 1. Following prior work [14, 45], we also use a Laplacian loss $\mathcal{L}_{\text{laplacian}}$ which calculates the L_1 loss after applying a Laplacian filter, and a gradient penalty loss $\mathcal{L}_{\text{grad}}$ which calculates an L_1 loss over spatial gradients of the alpha map.

Foreground Color Loss. To improve robustness, we ask our network to predict a pre-multiplied foreground image $(\alpha F)_i$ of the image $\mathbf{I}_i$ along with the alpha map. We supervise the pre-multiplied foreground prediction with a composition loss. That is, we recompose our predicted pre-multiplied foreground color back to the original image using the ground truth alpha and the background image. Formally, the composition loss can be written as:

$$\mathcal{L}_{\text{composition}} = \left|I_i - (\alpha_i F_i + (1-\alpha_i^{gt})B_i^{gt})\right|.$$

where I_i is the original pixel value, αF is the predicted pre-multiplied foreground color, and $(1-\alpha_i^{gt})B_i^{gt}$ is the ground truth pre-multiplied background color.

Combining all the components, the total loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sep}} + \mathcal{L}_{\text{laplacian}} + \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{composition}}.$$

4.5. Patch-based training and inference

Since matting usually involves a high-resolution image input, we train and test our models only over selected local patches that contain mixtures of foreground and background. Specifically, we first prepare full resolution input images, alignment results ($\mathbf{I}_1$, $M^B_{0\to1}(\mathbf{I}_0)$, $M^F_{0\to1}(\mathbf{I}_0)$, and $M^F_{0\to1}(M^B_{1\to0}(\mathbf{I}_1))$), and trimaps ($\mathbf{Tr}_1$ and $M^F_{0\to1}(\mathbf{Tr}_0)$). We then use the trimap $\mathbf{Tr}_1$ from the base frame to extract 448×448 patches that cover all uncertain regions. To ensure smooth transitions when fusing these patches, we leave an overlap of 224 pixels between neighboring patches. After predicting the alpha map and foreground color for each patch, we merge the overlapping patches with a Gaussian window function to produce the final results.

5. Experiments

5.1. Training and Implementation Details

Due to the unsatisfactory ground-truth quality in real data [40], our model is trained only on synthetic data. We create synthetic training samples by randomly compositing foreground subjects onto background images while simulating camera-induced parallax. The foreground subjects are sourced from two real-world portrait datasets—P3M-10K [15] and HHM-2K [35]—which contain 9,421 and 2,000 high-resolution images, respectively. Although these datasets provide imperfect alpha mattes and lack foreground color annotations, we generate pseudo-foreground color annotations using the layer-diffusion strategy [47]. The background images are drawn from BG-20K [16], which supplies 15,000 images for training. During composition, we further augment the alpha matte annotations through random gamma transformations. To reduce the gap between real and synthetic data and complicate the training, we also apply histogram equalization to the 50% foreground, aligning its color distribution with that of the background.

To simulate motion, random affine transformations are applied to both the foreground subject and the background image prior to composition, following the approach in [24]. Because our network processes warped images, we introduce additional random noise (approximately 10 pixels) to the affine transformations to ensure that the network never encounters perfectly aligned patches during training.

Given that our method relies on a rough trimap estimate as input, we first binarize the ground-truth alpha mask during training and then generate a trimap by applying random dilation/erosion operations (typically conducted for 60 to 120 iterations). At inference time, we generate the trimap using the strategy described in Sec. 4.1.

Our network architecture uses two ViT models [7] for the encoder and decoder, and a ViTMatte head [45]. The network is initialized with weights from the CroCo pre-trained model [41]. Training is conducted on 8 NVIDIA RTX 4090 GPUs with a batch size of 2 per GPU using the AdamW optimizer with a learning rate of 5e-5. We train our model with 50×100K synthetic patch pairs.

5.2. Evaluation on Synthetic Datasets

Datasets. Our test set is constructed similarly to our training data set, but we use a different real-world portrait matting dataset for testing. Specifically, we use P3M-500-NP [15], PPM-100 [14], and RWP-636 [46] for foreground, and we use BG-20K's test set as background.

Metrics. We use common evaluation metrics [29] to measure the accuracy of our predicted alpha maps which include the Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Connectivity (Conn), and the spatial gradient (Grad) metric. We follow the common practice to scale up the SAD and MSE numbers by 10^3 for better readability. To measure how accurate our foreground color estimation is, we calculate the MSE between the estimated pre-multiplied foreground colors αF v.s. the ground-truth. For methods that do not predict (pre-multiplied) foreground color, we follow the protocol of [24], where we use the input frame as its foreground prediction and apply the alpha matte on the input and treat it as the pre-multiplied foreground color.

Baselines. We compare against several state-of-the-art single-image matting and video matting methods. For trimap-free methods, we compare against MODNet [14] and ViTAE-S [27]. For trimap-based methods, we compare against MG-Mat [46] and MatteFormer [28]. We further evaluate our method against video matting methods, RVM [24], MaGGIe [11] and MatAnyone [1]. Among the baselines, ViTAE-S, MODNet, and MatteFormer only predict alpha. While MG-Mat can predict pre-multiplied foreground color, there is no public checkpoint. MODNet has an official demo that predicts pre-multiplied foreground color, which is a closed-source variant of its original version. We use their demo for qualitative results and open-source checkpoint for quantitative evaluation.

Quantitative Results. As shown in Tab. 1, our approach consistently outperforms existing state-of-the-art methods across all metrics. This result first validates our motivation, where motion is a strong signal to improve matting beyond just a single frame. It also proves that our design is able to utilize such information to recover high-quality alpha maps and foreground color.

Qualitative Results. Fig. 6 shows qualitative results from all baselines, our method, and the ground-truth annotation. Note that single-image matting struggles when the background is cluttered, where one can not reliably tell the fore-

Table 1. Quantitative comparison on synthetic test set.

Methods	Input Type	PPM-100					P3M-NP-500					RWP-636				
		SAD	MSE	Conn	Grad	MSE(αF)	SAD	MSE	Conn	Grad	MSE(αF)	SAD	MSE	Conn	Grad	MSE(αF)
MODNet [14]	Trimap-free	28.02	36.81	11.27	16.80	7.45	34.29	68.64	14.11	22.17	14.96	60.46	119.25	36.28	60.36	28.97
ViTAE-S [27]	Trimap-free	18.24	25.02	8.35	15.26	6.16	14.82	26.02	7.81	13.14	6.95	24.30	37.70	17.94	39.92	10.26
MG-Matting [46]	Trimap-based	49.99	73.76	31.95	22.34	15.97	35.64	53.01	22.41	15.93	12.86	49.42	85.92	29.27	32.67	20.81
MatteFormer [28]	Trimap-based	5.53	2.72	3.54	3.46	0.96	4.90	2.99	2.93	3.40	1.08	8.43	7.14	5.82	7.46	2.11
RVM [24]	Video	198.2	335.5	89.73	93.83	44.16	161.0	331.2	80.49	83.333	152.50	181.3	313.8	93.56	129.7	54.40
MaGGIe [11]	Video	27.41	7.86	5.51	9.72	2.04	18.77	7.02	6.04	6.35	2.47	21.33	15.48	10.64	14.92	4.79
MatAnyone [1]	Video	26.88	8.07	5.64	7.83	2.13	18.54	6.94	5.81	6.12	2.33	17.65	13.40	8.79	12.42	4.63
Ours	Two-view	**4.13**	**2.28**	**3.26**	**2.98**	**0.55**	**3.45**	**1.75**	**2.13**	**2.48**	**0.59**	**5.57**	**3.51**	**4.72**	**6.74**	**1.37**

Table 2. Ablation study. We implement four variants of our method and conduct the ablation study on PPM-100: (1) Our model without using the background-aligned frame; (2) Our model without using the foreground-aligned frame; (3) The single image version of our model. (4) Input the same frame to the dual-branch model. (5) Adding motion noise during inference.

	SAD	MSE	MSE (αF)
Ours	**4.13**	**2.28**	**0.55**
- (1) *w/o* background-aligned frame	7.52	3.78	1.57
- (2) *w/o* foreground-aligned frame	6.31	3.12	1.04
- (3) *w/o* another frame (only single branch)	9.53	4.75	1.80
- (4) *w/o* another frame (two same branch)	9.77	4.88	1.86
+ (5) motion noise	7.34	3.46	1.23

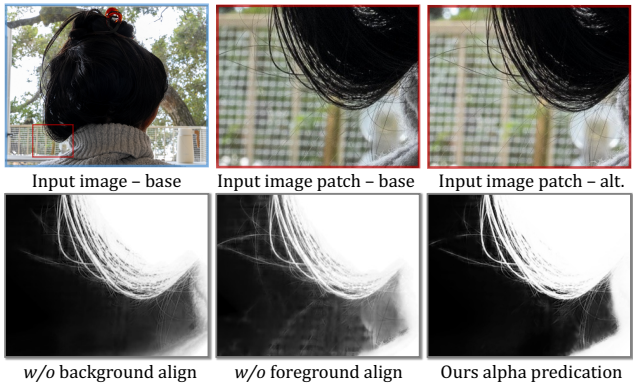


Figure 5. Ablation study of the background-aligned and foreground-aligned information.

ground subject from the background using a single frame. Motion effectively distinguishes the two, and our method can reliably recover fine details in the case.

5.3. Evaluation on real-world images

We further assess the robustness of our method in real-world scenarios as shown in Fig. 7. For each real-world case, we use the same camera settings to capture RAW images and render them with Adobe Lightroom using the same settings. For each scene, we capture a pair of images where there is some parallax between the subject and the background. For real images, we also compare with state-of-the-art commercial single-image matting solutions: Adobe Photoshop and Remove.bg. Our method consistently provides more details and better color separation than any baseline. In cases where the foreground color is close to the background, our method can still reliably predict the alpha mask and color using motion cues. More evaluation on real-world images can be found in supplementary material.

5.4. Ablation Study

We further perform ablation studies to justify our design choices for our prediction framework. Specifically, we remove the second branch of our network along with the foreground-aligned frame and the input of the background-aligned frame. This is equivalent to a single-image matting model. We also ablate over foreground motion and background motion estimates, where we either remove just one branch of our pipeline or the background-aligned frame from the input for each branch. Tab. 2 shows the quantitative result of our synthetic test dataset. Note that our method with both motion cues performs the best, which justifies our design choice. To further demonstrate robustness, we ran two additional experiments. In one case, we fed the same frame into both branches. The resulting output closely resembled that of the single-branch model, confirming that in the absence of camera motion the network effectively defaults to robust single-image matting. In another experiment, we introduced motion noise (averaging 4 pixels) to the estimated foreground flow. Despite the added noise, the model maintained its robustness and outperformed the single-image baseline, further underscoring its resilience to motion perturbations.

Fig. 5 illustrates the impact of ablating different components on real-world burst cases. Firstly, without a background-aligned burst, the model fails to capture fine details when the foreground objects are very small, resulting in less accurate alpha mattes. Secondly, without the foreground-aligned burst, the model makes errors when distinguishing between foreground and background. However, our method using full information, even with imperfect alignment, maintains high-quality matting results.

6. Conclusion & Limitations

In summary, we showed that parallax between foreground and background from camera motion are powerful signals

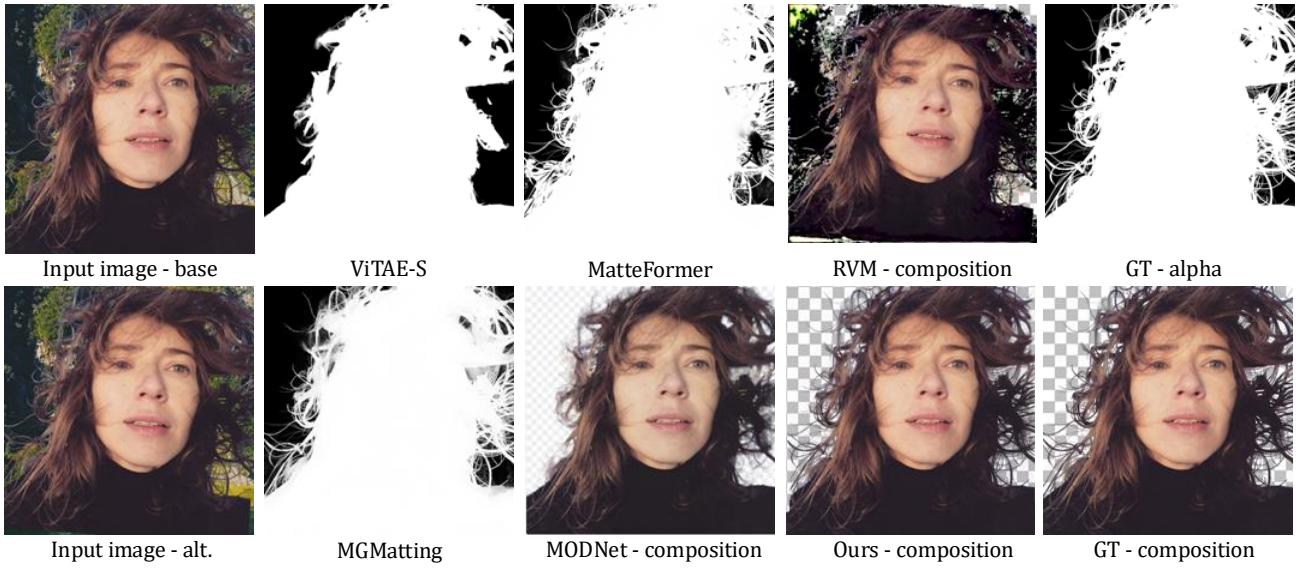


Figure 6. Qualitative results of our method and all the baselines on our synthetic test sets.

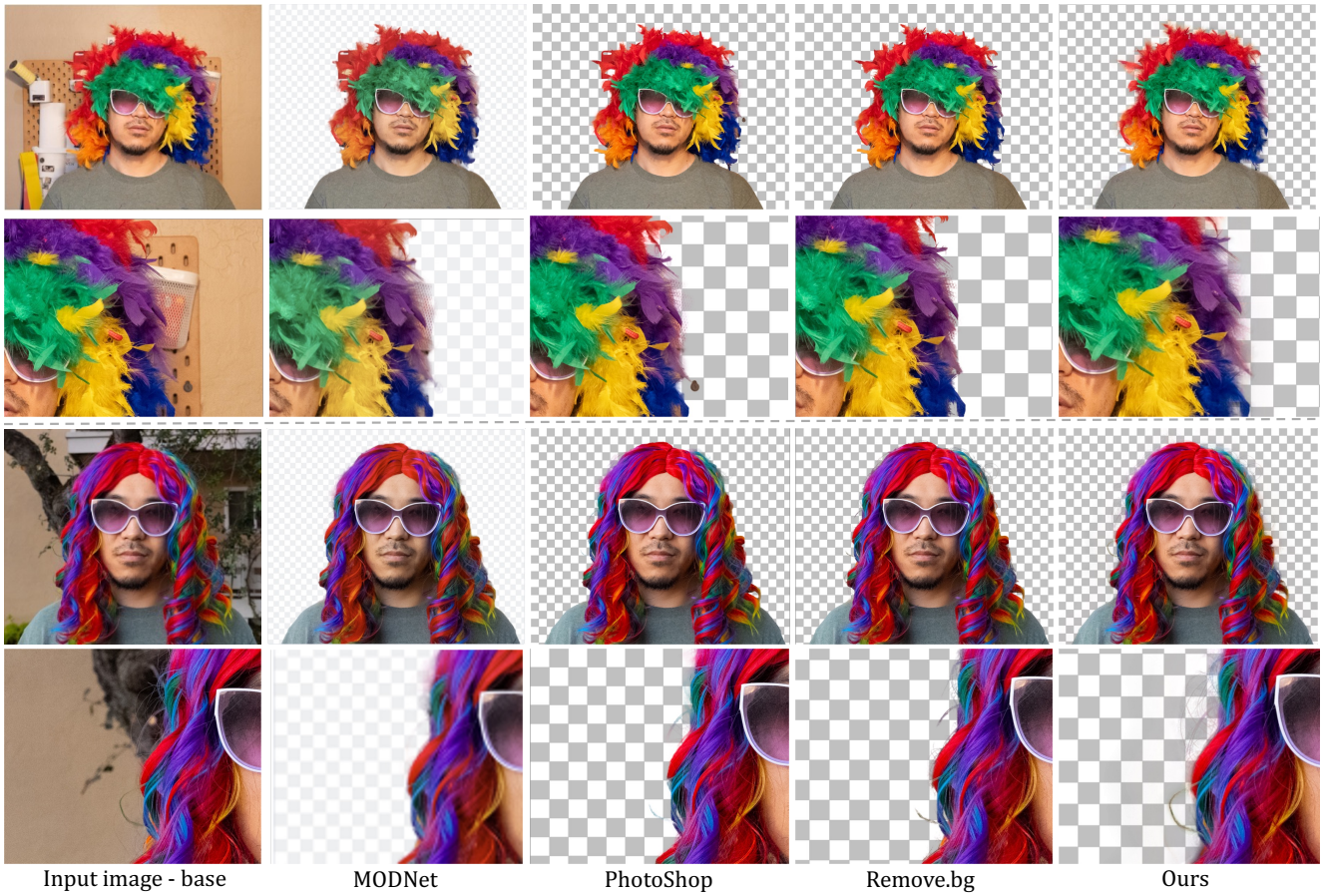


Figure 7. Qualitative results on real-world images.

to separate the two. In addition, we proposed a learning-based solution designed to implicitly handle errors in motion estimates, particularly for the foreground.

Our method has limitations that may point to interesting future work. Firstly, the motion estimation is detached from the training pipeline and cannot be jointly optimized. Secondly, our model's performance degrades if more versatile motions are introduced between the input frames. Finally,

motion may not be the silver bullet for matting in highly ambiguous cases, such as in low-light scenes, where the noise floor washes out details, or when the subject color is almost identical to the background over the entire burst.

References

[1] Matanyone: Stable video matting with consistent memory propagation. CVPR 2025. 6, 7

[2] Yağiz Aksoy, Tunç Ozan Aydin, Marc Pollefeys, and Aljoša Smolić. Interactive high-quality green-screen keying via color unmixing. *ACM Transactions on Graphics (TOG)*, 36 (4):1, 2016. 2

[3] Yosuke Bando, Bing-Yu Chen, and Tomoyuki Nishita. Extracting depth and matte using a color-filtered aperture. In *ACM SIGGRAPH Asia 2008 Papers*, New York, NY, USA, 2008. Association for Computing Machinery. 2, 3

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9209–9218, 2021. 4

[5] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018. 3

[6] Yung-Yu Chuang, Aseem Agarwala, Brian Curless, David H Salesin, and Richard Szeliski. Video matting of complex scenes. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 243–248, 2002. 3

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[8] Kenji Enomoto, TJ Rhodes, Brian Price, and Gavin Miller. Polarmatte: Fully computational ground-truth-quality alpha matte extraction for images and video using polarized screen matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3901–3909, 2024. 2, 3

[9] Ron Fry and Pamela Fourzon. *The saga of special effects*. Englewood Cliffs, NJ: Prentice-Hall, 1977. 2

[10] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR*, 2011. 2

[11] Chuong Huynh, Seoung Wug Oh, Abhinav Shrivastava, and Joon-Young Lee. Maggie: Masked guided gradual human instance matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3870–3879, 2024. 2, 6, 7

[12] Neel Joshi, Wojciech Matusik, and Shai Avidan. Natural video matting using camera arrays. *ACM Transactions on Graphics (TOG)*, 25(3):779–786, 2006. 2, 3

[13] Neel Joshi, Wojciech Matusik, Shai Avidan, Hanspeter Pfister, and William T Freeman. Exploring defocus matting: Nonparametric acceleration, super-resolution, and off-center matting. *IEEE Computer Graphics and Applications*, 27(2): 43–52, 2007. 2, 3

[14] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. MODNet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 3, 5, 6, 7

[15] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021. 6

[16] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 6

[17] Jiachen Li, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, Yunchao Wei, and Humphrey Shi. Vmformer: End-to-end video matting with transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6678–6687, 2024. 3

[18] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1785, 2024. 3

[19] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI*, 2020. 3

[20] Yaoyi Li, Jianfu Zhang, Weijie Zhao, Weihao Jiang, and Hongtao Lu. Inductive guided filter: Real-time deep matting with weakly annotated masks on mobile devices. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[21] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *ACM Trans. Graph.*, 38(6):164–1, 2019. 2

[22] 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. 3

[23] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *WACV*, 2022. 3

[24] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *WACV*, 2022. 3, 6, 7

[25] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 3

[26] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 2019. 2

[27] Sihan Ma, Jizhizi Li, Jing Zhang, He Zhang, and Dacheng Tao. Rethinking portrait matting with pirvacy preserving. *IJCV*, 131(8):2172–2197, 2023. 6, 7

[28] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *CVPR*, 2022. 2, 6, 7

[29] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually

motivated online benchmark for image matting. In *CVPR*, 2009. 6

[30] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 2, 3

[31] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 92–107. Springer, 2016. 3

[32] Alvy Ray Smith and James F. Blinn. Blue screen matting. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, page 259–268, New York, NY, USA, 1996. Association for Computing Machinery. 2

[33] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6975–6984, 2021. 3

[34] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2647–2656, 2022. 2

[35] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Ultrahigh resolution image/video matting with spatio-temporal sparsity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14112–14121, 2023. 3, 6

[36] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3055–3063, 2019. 2

[37] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2

[38] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[39] Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, and Ming-Hsuan Yang. Video matting via consistency-regularized graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4902–4911, 2021. 3

[40] Zhixiang Wang, Baiang Li, Jian Wang, Yu-Lun Liu, Jinwei Gu, Yung-Yu Chuang, and Shin'ichi Satoh. Matting by generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 6

[41] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: self-supervised pre-training for 3d vision tasks by cross-view completion. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 6

[42] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 2, 4

[43] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Designing effective inter-pixel information flow for natural image matting. In *CVPR*, 2017. 2

[44] Dogucan Yaman, Hazim Kemal Ekenel, and Alexander Waibel. Alpha matte generation from single input for portrait matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 696–705, 2022. 2

[45] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 2, 3, 5, 6

[46] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *CVPR*, 2021. 2, 6, 7

[47] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024. 6

[48] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI AIR*, 2024. 4