Parallax Portrait Matting

Xin Cai[1,3], Jiawen Chen[2], Lars Jebe[2], Tianfan Xue[1,3,4], and Zhoutong Zhang[2†]

[1] Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Adobe NextCam, San Jose, CA, USA
[3] Shanghai AI Laboratory, Shanghai, China
[4] CPII under InnoHK, Hong Kong SAR, China

Abstract. Image matting is highly ill-posed, especially when both the foreground and background are richly textured. While single-image matting methods learn strong priors from data, they often struggle on these challenging cases. Existing approaches improve results by requiring additional signals such as green screens, polarized lighting, or clean background images, but these typically rely on specialized capture setups. We present *Parallax Portrait Matting*, a practical two-frame matting method that uses a second image captured with slight viewpoint change. Such a setting arises naturally in burst photography, where small camera motion induces foreground-background parallax and provides complementary observations for matting. Our pipeline estimates trimaps and foreground/background motion, then constructs aligned views for prediction. To handle imperfect motion estimation, the network uses the background-aligned pair for direct fusion and the foreground-aligned cue through cross-attention for error compensation. Experiments show that our method recovers finer details and more accurate foreground colors than strong single-image matting baselines on challenging portrait cases.

1 Introduction

Image matting has a long history in computer vision and graphics [10, 31]. 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.

Most modern methods resolve this ambiguity either by learning strong priors from annotated data [21, 27, 41, 44, 45] or by leveraging generative models [38], but single-image prediction remains difficult in highly ambiguous cases. Another line of work tackles the ill-posedness of single-image matting by acquiring additional observations. However, existing solutions often rely on specialized capture

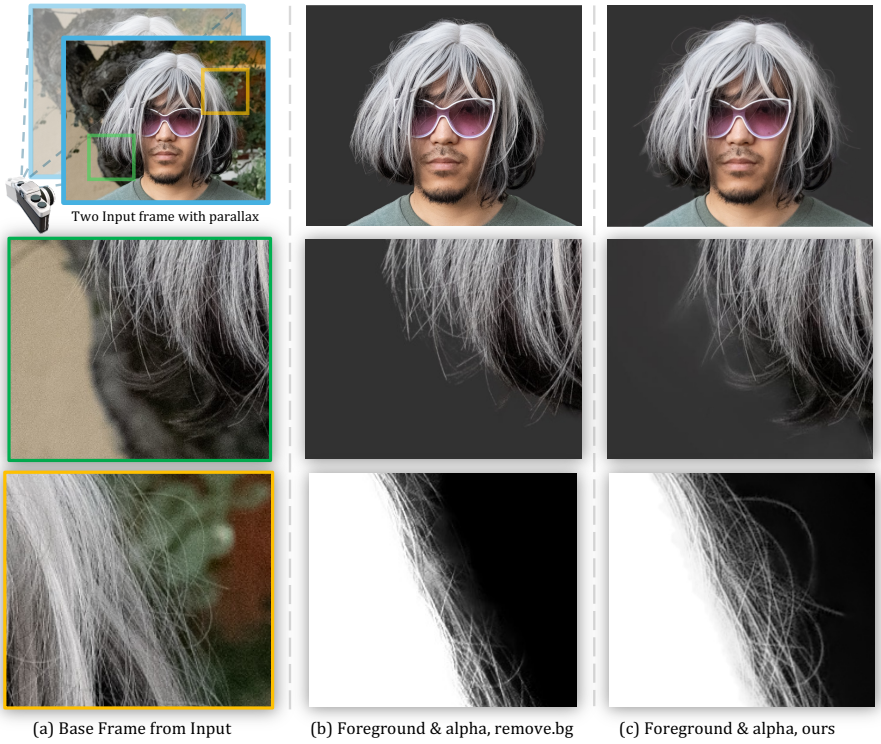

Fig. 1: Our matting method exploits camera-motion-induced parallax between the foreground and the background. It takes two frames as input, each taken with a slightly different camera location, and predicts both a pre-multiplied foreground image and an alpha map. Trained on public datasets, our method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

setups such as green screens, polarization, camera arrays, focal stacks, or clean background images [1,2,9,13,14,29]. These approaches can produce high-quality mattes, especially for challenging regions such as hair and semi-transparent boundaries, but they typically require dedicated hardware or carefully controlled capture procedures, making them impractical for everyday photography.

In this work, we introduce *Parallax Portrait Matting* (Fig. 1), a practical two-frame matting method for portrait scenes. Our key idea is to exploit an additional image captured with slight viewpoint change, which is often already available in casual photography or mobile burst capture. Because the portrait subject is typically closer to the camera than the background, the two layers exhibit different apparent motion, and this parallax provides extra constraints for separating fine foreground structures from complex backgrounds. This makes the additional frame a practical and informative cue for portrait matting without requiring specialized hardware or controlled capture setups.

The central challenge is that this cue is only useful if motion can be modeled robustly. Off-the-shelf motion estimators [36, 40] are unreliable precisely where

matting is difficult: pixels that mix foreground and background do not follow a single motion. We therefore approximate the motion in such regions as a combination of two locally smooth fields, one for the foreground and one for the background, extrapolated from nearby certain regions. This approximation is only imperfectly valid in real capture, where wind, hair motion, or slight subject movement may introduce residual motion. Our method is thus designed to use motion conservatively: it exploits parallax when helpful, and degrades gracefully to strong single-image predictions when motion estimation is imperfect.

Our key observation is that foreground and background motion are not equally reliable. In typical burst portrait capture, background motion is easier to estimate: the background is often farther from the camera, its apparent motion is smoother, and it can frequently be aligned accurately enough for direct pixel-level fusion. Foreground motion, in contrast, is much harder to estimate reliably, especially around thin structures such as hair, where local non-rigid motion and residual flow errors are common. This asymmetry motivates our network design. We use the background-aligned pair as the primary pixel-aligned input for matte prediction, while treating the foreground-aligned pair as a noisier auxiliary cue. Rather than fusing it directly in pixel space, we inject it through cross-attention in feature space to compensate for residual foreground alignment errors. This allows the model to use multi-view information when it is helpful, while remaining robust when foreground alignment is inaccurate.

We train the model on local patches and apply image and motion augmentations to improve robustness to real-world degradation and alignment noise. Experiments show that our method outperforms single-image matting models on challenging portrait cases, producing both finer alpha mattes and substantially more accurate foreground colors than existing baselines and closed-source commercial solutions. Accurate foreground color estimation is critical for downstream tasks such as compositing onto new backgrounds, yet most existing matting methods either ignore it entirely or produce colors contaminated by the original background. These results suggest that a casually captured second frame is a practical and complementary cue for portrait matting.

In summary, our contributions are threefold:

- We introduce *Parallax Portrait Matting*, a practical two-frame portrait matting formulation that leverages an additional image with slight viewpoint change to better constrain the decomposition.
- We develop a robust matting framework that uses parallax cues under imperfect motion estimation, with direct fusion from reliable background alignment and feature-level compensation from noisier foreground alignment.
- We show empirically that a captured second frame improves both alpha matte quality and foreground color accuracy over strong single-image baselines on challenging portrait cases, while remaining robust to motion estimation errors.

2 Related Work

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 [21, 42, 45], instance segmentation [12, 32], a trimap [11, 25, 27, 35] or different types of user annotations [17, 24]. More recent learning-based methods take as input a single image without any guidance, either by implicitly incorporating guidance signal prediction [6, 15, 20, 30, 44], or by leveraging strong learned priors, *e.g.* from generative models [38]. Such generative priors have also proven effective for recovering or preserving fine high-frequency details in other ill-posed reconstruction problems, such as lensless image reconstruction [4] and high-compression latent diffusion [5]. 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 methods extend beyond current single-image matting models with temporal feature aggregation designs, such as graph neural networks [37], temporal RNNs [23], transformers [16], deformable convolution [34], or temporal image difference [33]. All these methods aim to learn such feature aggregation end-to-end with video data supervision, making single-image matting consistent over the entire video sequence. The most similar work to ours is [7], 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 methods in that we focus on utilizing motion between frames to improve single-image matting results for challenging cases.

Matting with additional signals. Additional information helps better condition the matting problem. Background matting [22, 29] uses an additional background image as conditioning and is able to produce high quality matting for images and videos. Polarization systems [9] are able to capture ground truth transmittance maps. Attaching a color filter [2] 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 [13]. In addition, focal stacks [14] can be used to provide additional information since foreground and background are blurred differently. Our method differs from those methods in that we do not ask the user to perform additional setups before capturing; we simply 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:

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

The objective of parallax portrait 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 handles this assumption by being tolerant of 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.

Together, these conditions define the operating regime of our method: small-baseline burst portraits with visible foreground-background parallax, a mostly

static subject and scene, and reasonably consistent appearance across the two frames. We emphasize that they are not hard requirements. Because our design uses motion conservatively—trusting the reliable background alignment for direct fusion while treating foreground alignment only as a noisy feature-level cue—the model degrades gracefully toward single-frame behavior when parallax is weak or alignment is unreliable, rather than failing catastrophically.

4 Method

Given two portrait images captured from slightly different viewpoints, our goal is to predict the foreground and alpha matte of a chosen base frame. Our pipeline (Fig. 4) first estimates a trimap for each frame, then computes foreground and background motion fields, and finally uses the resulting aligned views as input to a matting network. The network takes both background-aligned and foreground-aligned observations together with the trimaps, and predicts the foreground and alpha map of the base frame.

In practice, background alignment is usually more reliable than foreground alignment. We therefore use the background-aligned pair as the primary input for direct fusion, and use the foreground-aligned cue only through cross-attention in feature space to compensate for residual foreground motion errors. This design allows the model to exploit parallax when it is informative, while remaining robust when foreground alignment is imperfect.

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 [47], we generate a binary foreground mask and then erode and dilate it by 100 pixels to produce foreground and background regions, respectively (see Fig. 2). The resulting uncertain band is 200 pixels wide, which is generous enough to cover the error margin of modern 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 [3] 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 [40]. 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.

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

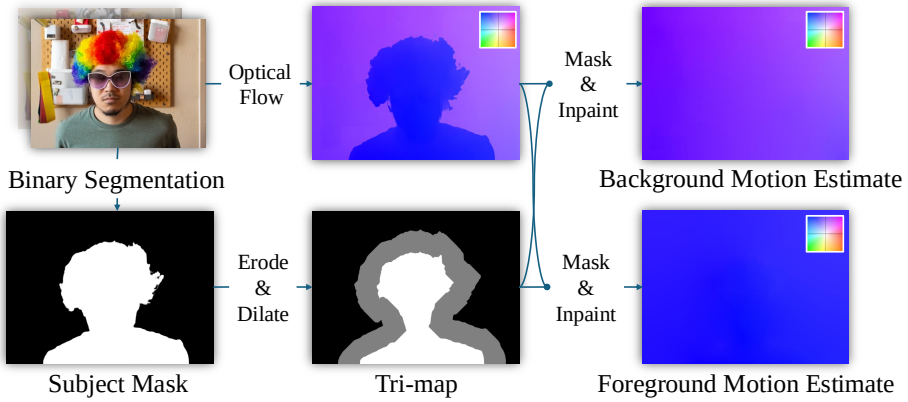

Fig. 2: Our motion estimation pipeline. Given two frames, we first estimate optical flow between the two, together with a trimap generated as described in Sec. 4.1. Since flow estimation in overlapping areas is often incorrect, we inpaint them by their nearest neighbor in non-overlapping regions.

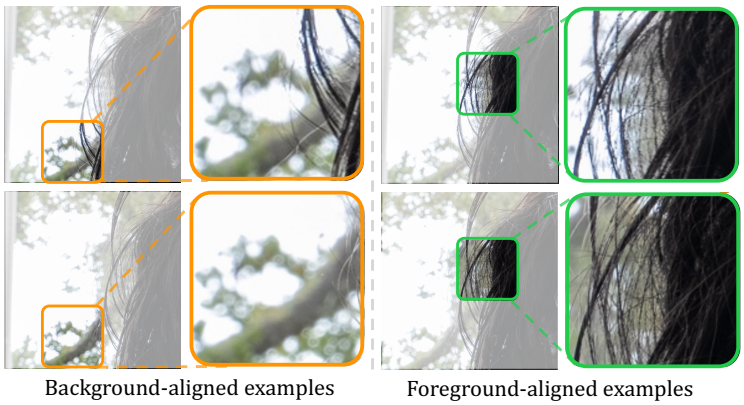

Fig. 3: An example of foreground and background alignment. Background-aligned patches only contain foreground subject motion, and the dis-occluded background in the alternate frame provides more context. Foreground-aligned patches fixate on the subject, which directly helps in separating the foreground.

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

A straightforward design is to warp the alternate frame using both estimated motion fields and directly concatenate all aligned images in pixel space to predict

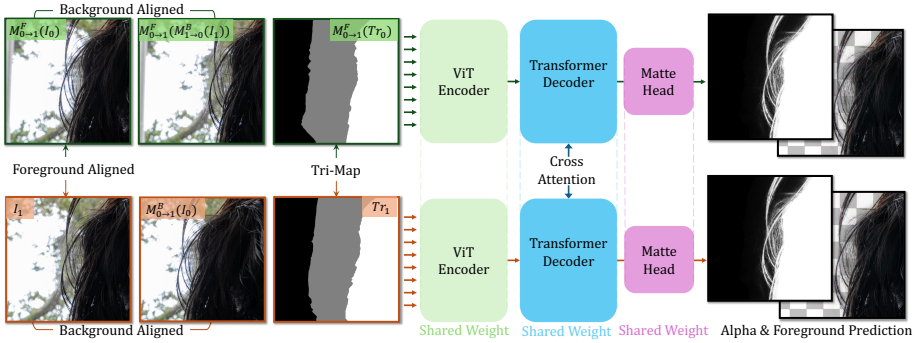

Fig. 4: Overview of our framework. The bottom branch is the primary prediction path and uses the base frame together with a background-aligned alternate frame, which provides reliable pixel-aligned evidence for matting. The top branch uses a foreground-aligned view as an auxiliary cue. Because foreground alignment is often imperfect, information from this branch is incorporated through cross-attention rather than direct pixel-space fusion, allowing the network to compensate for foreground motion errors while suppressing unreliable correspondences.

the alpha matte and foreground color. In practice, however, this strategy is brittle because it implicitly assumes that foreground and background alignments are equally reliable. This assumption does not hold in real captures. Background motion is usually easier to estimate and often remains smooth across the image, so the background-aligned image $\mathbf{I}^B_{0\to1} = M^B_{0\to1}(\mathbf{I}_0)$ typically provides relatively trustworthy pixel correspondences with the base image $\mathbf{I}_1$. Foreground motion, by contrast, is much harder to estimate accurately due to local non-rigid motion, especially around thin structures such as hair. As a result, directly concatenating a foreground-aligned image $\mathbf{I}^F_{0\to1} = M^F_{0\to1}(\mathbf{I}_0)$ may inject incorrect pixel-level evidence and confuse the prediction network.

To address this issue, we explicitly decouple *reliable fusion* from *error compensation*. Our model uses the background-aligned pair as the primary input for matte prediction, since $\mathbf{I}_1$ and $\mathbf{I}^B_{0\to1}$ are already approximately aligned in pixel space and can therefore be fused directly. Concretely, one branch of our framework takes the base image $\mathbf{I}_1$, the background-aligned image $\mathbf{I}^B_{0\to1}$, and the trimap $\mathbf{Tr}_1$ as input, and predicts the foreground image F_1 and alpha map α_1 of the base frame. This branch serves as the main prediction path and is responsible for learning a strong and stable matting solution.

We then introduce a second branch based on foreground alignment. Instead of treating the foreground-aligned image as another source of hard pixel-aligned evidence, we use it as a noisy auxiliary cue. Specifically, the second branch takes the foreground-aligned image $\mathbf{I}^F_{0\to1} = M^F_{0\to1}(\mathbf{I}_0)$, a companion frame $M^F_{0\to1}(M^B_{1\to0}(\mathbf{I}_1))$, and the warped trimap $M^F_{0\to1}(\mathbf{Tr}_0)$ as input. This branch predicts the alpha and foreground of the foreground-aligned view. During decoding, the two branches interact through cross-attention layers, which allow the model to establish soft correspondences in feature space rather than relying on exact pixel alignment. In this way, the network can selectively borrow use-

ful information from the foreground-aligned branch where the correspondence is reliable, while suppressing regions affected by motion estimation errors.

Our framework is therefore symmetric in architecture but asymmetric in function: the background-aligned stream provides reliable pixel-level evidence for direct fusion, whereas the foreground-aligned stream provides feature-level correction for residual alignment errors. The cross-attention mechanism is particularly important near ambiguous regions such as thin structures and semi-transparent boundaries, where foreground motion is most difficult to estimate accurately. Moreover, because the main branch already forms a strong prediction path using the base image and the more reliable background-aligned observation, our model naturally remains robust when parallax is weak or foreground alignment is inaccurate. The shared weights between the two branches further encourage a unified representation, allowing the model to benefit from both single-image matting data and multi-frame parallax cues. As a result, the network can exploit foreground parallax when it is informative, while gracefully falling back to stable single-frame behavior when the auxiliary motion cue is noisy.

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 the set of pixels that are "soft", meaning they have a ground truth alpha value between 0 and 1, and H denotes the set of pixels that are "hard" and have a ground truth alpha value of exactly 0 or 1. Following prior work [15, 44], 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, $\alpha_i F_i$ 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 high-resolution inputs, we train and test our model on selected local patches that contain mixtures of foreground and background. Specifically, we first prepare full-resolution inputs, 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 $\mathbf{Tr}_1$ of the base frame to extract 448×448 patches covering all uncertain regions. To ensure smooth transitions when fusing patches, we leave an overlap of 224 pixels between neighbors. After predicting the alpha map and foreground color for each patch, we merge overlapping patches with a Gaussian window function to produce the final result.

5 Experiments

5.1 Training and Implementation Details

Due to the unsatisfactory ground-truth quality in real data [38], 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 [18] and HHM-2K [33]—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 [46]. The background images are drawn from BG-20K [19], 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 increase training difficulty, we also apply histogram equalization to 50% of the foreground images, aligning their color distribution with that of the background.

To simulate motion, we apply random affine transformations to both the foreground subject and background image before composition, following [23]. Because our network processes warped images, we add random noise (approximately 10 pixels) to these transformations so 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 matting network uses two ViT-Small models [8] for the encoder and decoder, and a ViTMatte head [44], totaling 38.6M parameters, which is comparable to or fewer than MatteFormer (44.7M) and MatAnyone (35.2M). The network is initialized with weights from the CroCo pre-trained model [39], whose cross-

view completion pre-training is a natural fit for our correspondence-aware two-frame prediction; this is an initialization choice rather than a required component of our formulation, and stronger cross-view or 3D foundation models could be substituted in future work. The full pipeline additionally uses BiRefNet [47] (200M) for trimap estimation and GMFlow [40] (7.4M) for motion estimation, both of which are off-the-shelf components and can be replaced with more efficient alternatives. 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 5×10^{-5}. We train for 50 epochs, each consisting of 100K synthetic patch pairs.

5.2 Evaluation on Synthetic Data

Datasets. Our test set is constructed similarly to our training dataset, but we use a different real-world portrait matting dataset for testing. Specifically, we use P3M-500-NP [18], PPM-100 [15], and RWP-636 [45] for foreground, and we use BG-20K's test set as background.

Metrics. We use common evaluation metrics [28] 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 and the ground truth. For methods that do not predict (pre-multiplied) foreground color, we follow the protocol of [23]: we multiply the input frame by the predicted alpha matte and treat the result 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 [15] and ViTAE-S [26]. For trimap-based methods, we compare against MG-Matting [45] and MatteFormer [27]. We further evaluate our method against video matting methods MaGGIe [12] and MatAnyone [43]. Among the baselines, ViTAE-S, MODNet, and MatteFormer only predict alpha; for these methods we use the input image masked by the predicted alpha as their foreground color. For a fair comparison, all trimap-based methods (MG-Matting, MatteFormer, and ours) use the *same* BiRefNet-generated trimap described in Sec. 4.1; no manual or method-specific trimaps are used, so the comparison does not depend on trimap quality. The trimap-free methods are included only as single-image context. The video matting baselines are designed for full-sequence temporal aggregation rather than two-frame parallax; we therefore treat them as adjacent-frame references, and their gap to our method reflects the benefit of explicit parallax reasoning rather than mere access to additional frames.

Quantitative Results. As shown in Tab. 1, our approach consistently outperforms all baselines across every metric, validating that motion is a strong complementary signal for matting. The gain is especially pronounced in foreground color accuracy (MSE of αF), where our method reduces error by 35–45% over the best baseline—critical for downstream compositing.

Table 1: Quantitative comparison on synthetic test set.

		PPM-100					P3M-NP-500					RWP-636				
Method	Input	SAD	MSE	Conn	Grad	MSE(αF)	SAD	MSE	Conn	Grad	MSE(αF)	SAD	MSE	Conn	Grad	MSE(αF)
MODNet [15]	TF	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 [26]	TF	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 [45]	TB	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 [27]	TB	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
MaGGIe [12]	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 [43]	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	2-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**

TF = trimap-free, TB = trimap-based.

Fig. 5: Qualitative results on synthetic test sets.

Qualitative Results. Fig. 5 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 cannot reliably distinguish the foreground subject from the background using a single frame. Motion effectively separates the two, and our method can recover fine details in such cases.

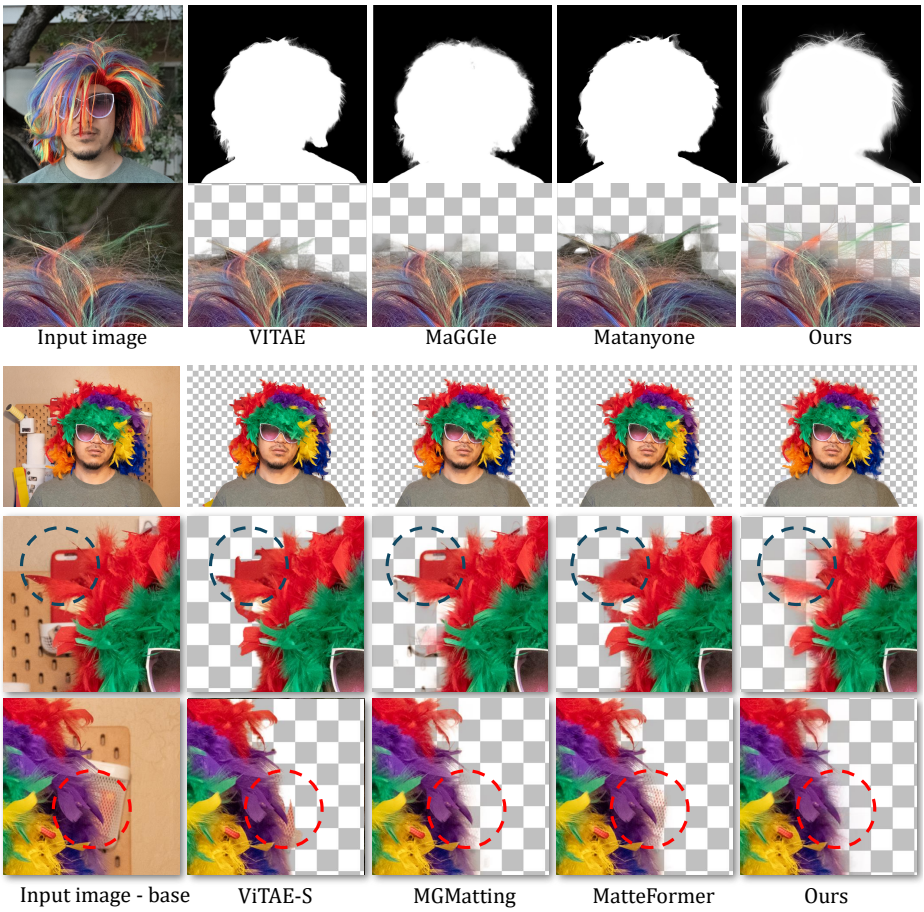

Fig. 6: Qualitative results on real-world images. Our method produces higher-quality alpha mattes, more accurate foreground colors, and finer structural details compared to existing methods.

5.3 Evaluation on Real-World Data

We further assess the robustness of our method in real-world scenarios. For each real-world case, we capture RAW image pairs with fixed camera settings and render them consistently in Adobe Lightroom. Each scene contains slight parallax between the portrait subject and the background. Notably, our model is trained using only synthetic data and does not use any real-world matting pairs for training. Nevertheless, as shown in Fig. 6, our method consistently produces higher-quality alpha mattes and more accurate foreground colors than existing approaches. Figure 7 further compares our method against closed-source commercial solutions (Adobe Photoshop and Remove.bg), showing that our approach recovers more detail in fine structures such as hair strands. More real-world examples and qualitative results are provided in the **supplementary material**.

Fig. 7: Comparison against closed-source commercial solutions (Adobe Photoshop and Remove.bg) on real-world images. Despite competing with proprietary systems, our method recovers significantly more detail in fine structures such as hair strands.

5.4 Ablation Study

Table 2: Ablation study on PPM-100. We evaluate five variants of our method: (1) without using the background-aligned frame; (2) without using the foreground-aligned frame; (3) single-image version (single branch); (4) same frame input 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* bg-aligned frame	7.52	3.78	1.57
- (2) *w/o* fg-aligned frame	6.31	3.12	1.04
- (3) *w/o* 2nd frame (single branch)	9.53	4.75	1.80
- (4) *w/o* 2nd frame (same input)	9.77	4.88	1.86
+ (5) motion noise	7.34	3.46	1.23

We further perform ablation studies to justify our design choices. Removing the second branch together with both aligned inputs reduces the model to a single-image baseline, while ablating the background-aligned and foreground-aligned cues individually reveals their distinct roles. As shown in Tab. 2, using both cues yields the best overall performance, indicating that they provide complementary rather than redundant information. Qualitatively, removing the background-aligned cue mainly degrades fine-detail recovery, whereas removing the foreground-aligned cue increases foreground-background confusion, as illustrated in Fig. 8. We also test the degenerate case where the same frame is fed into both branches; the result is close to that of the single-branch model, con-

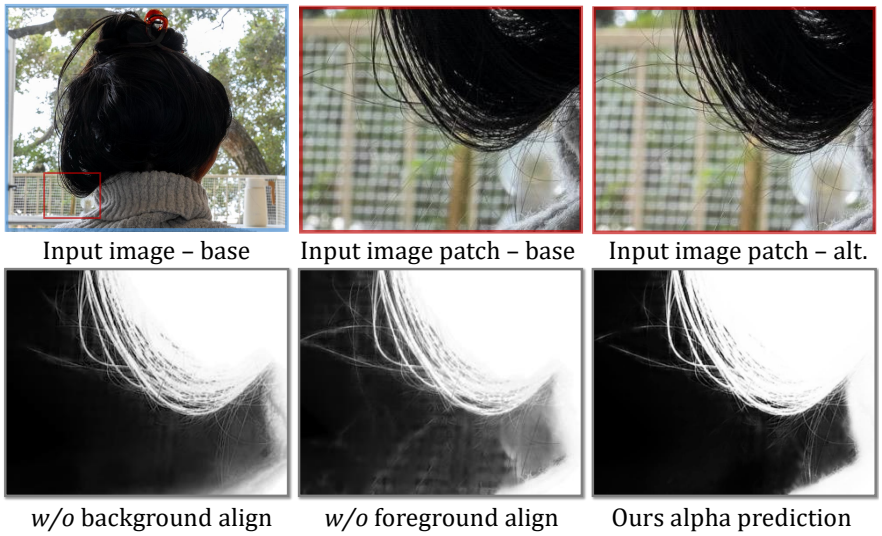

Fig. 8: Ablation on real-world cases. Removing background-aligned burst loses fine details; removing foreground-aligned burst causes foreground-background confusion.

firming graceful degradation when parallax is absent. Finally, we perturb the estimated foreground flow by introducing additional motion noise. Even under this imperfect alignment, the model still outperforms the single-image baseline. This supports our central design choice: foreground alignment should not be treated as hard pixel-level evidence, but as a noisy auxiliary cue whose value is best exploited through feature-level interaction. Taken together, these ablations indicate that the two aligned views play complementary and non-interchangeable roles, and that the consistent gains stem from our asymmetric use of the two cues rather than from the mere availability of a second frame.

6 Conclusion & Limitations

We presented *Parallax Portrait Matting*, showing that foreground-background parallax from slight camera motion provides a practical cue for portrait matting. Our method uses this cue robustly by treating background alignment as reliable evidence and foreground alignment as a noisier auxiliary signal.

Our method has limitations that may point to interesting future work. First, motion estimation is detached from training and cannot be jointly optimized. Second, the model's performance may degrade under larger or more complex motion between frames. Finally, parallax is not sufficient in extremely ambiguous cases, such as low-light scenes or when the subject and background remain nearly indistinguishable throughout the burst.

Acknowledgements

This work was supported by the Research Grants Council (RGC) of Hong Kong under the Early Career Scheme (ECS) No. 24209224.

References

1. Aksoy, Y., Aydin, T.O., Pollefeys, M., Smolić, A.: Interactive high-quality green-screen keying via color unmixing. ACM Transactions on Graphics (TOG) **35**(5), 152:1–152:12 (2016)
2. Bando, Y., Chen, B.Y., Nishita, T.: Extracting depth and matte using a color-filtered aperture. In: ACM SIGGRAPH Asia 2008 Papers. SIGGRAPH Asia '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1457515.1409087, https://doi.org/10.1145/1457515.1409087
3. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Deep burst super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9209–9218 (2021)
4. Cai, X., You, Z., Zhang, H., Liu, W., Gu, J., Xue, T.: PhoCoLens: Photorealistic and consistent reconstruction in lensless imaging. In: NeurIPS (2024)
5. Cai, X., You, Z., Zhang, Z., Xue, T.: DA-VAE: Plug-in latent compression for diffusion via detail alignment. In: CVPR (2026)
6. Chen, Q., Ge, T., Xu, Y., Zhang, Z., Yang, X., Gai, K.: Semantic human matting. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 618–626 (2018)
7. Chuang, Y.Y., Agarwala, A., Curless, B., Salesin, D.H., Szeliski, R.: Video matting of complex scenes. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 243–248 (2002)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Enomoto, K., Rhodes, T., Price, B., Miller, G.: 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. pp. 3901–3909 (2024)
10. Fry, R., Fourzon, P.: The saga of special effects. Englewood Cliffs, NJ: Prentice-Hall (1977)
11. He, K., Rhemann, C., Rother, C., Tang, X., Sun, J.: A global sampling method for alpha matting. In: CVPR (2011)
12. Huynh, C., Oh, S.W., Shrivastava, A., Lee, J.Y.: Maggie: Masked guided gradual human instance matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3870–3879 (2024)
13. Joshi, N., Matusik, W., Avidan, S.: Natural video matting using camera arrays. ACM Transactions on Graphics (TOG) **25**(3), 779–786 (2006)
14. Joshi, N., Matusik, W., Avidan, S., Pfister, H., Freeman, W.T.: Exploring defocus matting: Nonparametric acceleration, super-resolution, and off-center matting. IEEE Computer Graphics and Applications **27**(2), 43–52 (2007)
15. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: MODNet: Real-time trimap-free portrait matting via objective decomposition. In: AAAI (2022)
16. Li, J., Goel, V., Ohanyan, M., Navasardyan, S., Wei, Y., Shi, H.: Vmformer: End-to-end video matting with transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6678–6687 (2024)
17. Li, J., Jain, J., Shi, H.: Matting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1775–1785 (2024)
18. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3501–3509 (2021)

19. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. International Journal of Computer Vision **130**(2), 246–266 (2022)
20. Li, Y., Lu, H.: Natural image matting via guided contextual attention. In: AAAI (2020)
21. Li, Y., Zhang, J., Zhao, W., Jiang, W., Lu, H.: Inductive guided filter: Real-time deep matting with weakly annotated masks on mobile devices. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
22. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: CVPR (2021)
23. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: WACV (2022)
24. Liu, J., Yao, Y., Hou, W., Cui, M., Xie, X., Zhang, C., Hua, X.s.: Boosting semantic human matting with coarse annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8563–8572 (2020)
25. Lu, H., Dai, Y., Shen, C., Xu, S.: Context-aware image matting for simultaneous foreground and alpha estimation. In: ICCV (2019)
26. Ma, S., Li, J., Zhang, J., Zhang, H., Tao, D.: Rethinking portrait matting with privacy preserving. IJCV **131**(8), 2172–2197 (2023)
27. Park, G., Son, S., Yoo, J., Kim, S., Kwak, N.: Matteformer: Transformer-based image matting via prior-tokens. In: CVPR (2022)
28. Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., Rott, P.: A perceptually motivated online benchmark for image matting. In: CVPR (2009)
29. Sengupta, S., Jayaram, V., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Background matting: The world is your green screen. In: CVPR (2020)
30. Shen, X., Tao, X., Gao, H., Zhou, C., Jia, J.: Deep automatic portrait matting. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 92–107. Springer (2016)
31. Smith, A.R., Blinn, J.F.: Blue screen matting. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. p. 259–268. SIGGRAPH '96, Association for Computing Machinery, New York, NY, USA (1996). https://doi.org/10.1145/237170.237263, https://doi.org/10.1145/237170.237263
32. Sun, Y., Tang, C.K., Tai, Y.W.: Human instance matting via mutual guidance and multi-instance refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2647–2656 (2022)
33. Sun, Y., Tang, C.K., Tai, Y.W.: Ultrahigh resolution image/video matting with spatio-temporal sparsity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14112–14121 (2023)
34. Sun, Y., Wang, G., Gu, Q., Tang, C.K., Tai, Y.W.: Deep video matting via spatio-temporal alignment and aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6975–6984 (2021)
35. Tang, J., Aksoy, Y., Oztireli, C., Gross, M., Aydin, T.O.: Learning-based sampling for natural image matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3055–3063 (2019)
36. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020)
37. Wang, T., Liu, S., Tian, Y., Li, K., Yang, M.H.: Video matting via consistency-regularized graph neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4902–4911 (2021)

38. Wang, Z., Li, B., Wang, J., Liu, Y.L., Gu, J., Chuang, Y.Y., Satoh, S.: Matting by generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
39. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: 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. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
40. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8121–8130 (2022)
41. Xu, N., Price, B., Cohen, S., Huang, T.: Designing effective inter-pixel information flow for natural image matting. In: CVPR (2017)
42. Yaman, D., Ekenel, H.K., Waibel, A.: Alpha matte generation from single input for portrait matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 696–705 (2022)
43. Yang, P., Zhou, S., Zhao, J., Tao, Q., Loy, C.C.: MatAnyone: Stable video matting with consistent memory propagation. In: CVPR (2025)
44. Yao, J., Wang, X., Yang, S., Wang, B.: Vitmatte: Boosting image matting with pre-trained plain vision transformers. Information Fusion **103**, 102091 (2024)
45. Yu, Q., Zhang, J., Zhang, H., Wang, Y., Lin, Z., Xu, N., Bai, Y., Yuille, A.: Mask guided matting via progressive refinement network. In: CVPR (2021)
46. Zhang, L., Agrawala, M.: Transparent image layer diffusion using latent transparency. arXiv preprint arXiv:2402.17113 (2024)
47. Zheng, P., Gao, D., Fan, D.P., Liu, L., Laaksonen, J., Ouyang, W., Sebe, N.: Bilateral reference for high-resolution dichotomous image segmentation. CAAI Artificial Intelligence Research **3**, 9150038 (2024)