

MaGGIe: Masked Guided Gradual Human Instance Matting

Chuong Huynh^{1*} Seoung Wug Oh² Abhinav Shrivastava¹ Joon-Young Lee²
¹University of Maryland, College Park ²Adobe Research
¹{chuonghm, abhinav}@cs.umd.edu ²{seoh, jolee}@adobe.com

Abstract

Human matting is a foundation task in image and video processing where human foreground pixels are extracted from the input. Prior works either improve the accuracy by additional guidance or improve the temporal consistency of a single instance across frames. We propose a new framework **MaGGIe**, **M**asked **G**uided **G**radual **H**uman **I**nstance **M**atting, which predicts alpha mattes progressively for each human instances while maintaining the computational cost, precision, and consistency. Our method leverages modern architectures, including transformer attention and sparse convolution, to output all instance mattes simultaneously without exploding memory and latency. Although keeping constant inference costs in the multiple-instance scenario, our framework achieves robust and versatile performance on our proposed synthesized benchmarks. With the higher quality image and video matting benchmarks, the novel multi-instance synthesis approach from publicly available sources is introduced to increase the generalization of models in real-world scenarios. Our code and datasets are available at <https://maggie-matt.github.io>.

1. Introduction

In image matting, a trivial solution is to predict the pixel transparency - alpha matte $\alpha \in [0, 1]$ for precise background removal. Considering a saliency image I with two main components, foreground F and background B , the image I is expressed as $I = \alpha F + (1 - \alpha)B$. Because of the ambiguity in detecting the foreground region, for example, whether a person's belongings are a part of the human foreground or not, many methods [11, 16, 31, 37] leverage additional guidance, typically trimaps, defining foreground, background, and unknown or transition regions. However, creating trimaps, especially for videos, is resource-intensive. Alternative binary masks [39, 56] are simpler to obtain by human drawings or off-the-shelf segmentation models while offering greater flexibility without hardly con-

*This work was done during Chuong Huynh's internship at Adobe

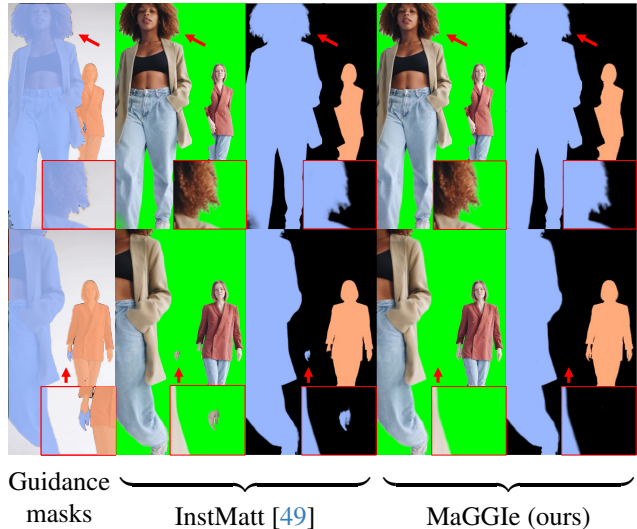


Figure 1. **Our MaGGIe delivers precise and temporally consistent alpha mattes.** It adeptly preserves intricate details and demonstrates robustness against noise in instance guidance masks by effectively utilizing information from adjacent frames. Red arrows highlight the areas of detailed zoom-in. (Optimally viewed in color and digital zoom in).

straint output values of regions as trimaps. Our work focuses but is not limited to human matting because of the higher number of available academic datasets and user demand in many applications [1, 2, 12, 15, 44] compared to other objects.

When working with video input, the problem of creating trimap guidance is often resolved by guidance propagation [17, 45] where the main idea coming from video object segmentation [8, 38]. However, the performance of trimap propagation degrades when video length grows. The failed trimap predictions, which miss some natures like the alignment between foreground-unknown-background regions, lead to incorrect alpha mattes. We observe that using binary masks for each frame gives more robust results. However, the consistency between the frame's output is still important for any video matting approach. For example, holes appearing in a random frame because of wrong guidance should be corrected by consecutive frames. Many

works [17, 32, 34, 45, 53] constrain the temporal consistency at feature maps between frames. Since the alpha matte values are very sensitive, feature-level aggregation is not an absolute guarantee of the problem. Some methods [21, 50] in video segmentation and matting compute the incoherent regions to update values across frames. We propose a temporal consistency module that works in both feature and output spaces to produce consistent alpha mattes.

Instance matting [49] is an extension of the matting problem where there exists multiple $\alpha_i, i \in 0..N$, and each belongs to one foreground instance. This problem creates another constraint for each spatial location (x, y) value such that $\sum_i \alpha_i(x, y) = 1$. The main prior work InstMatt [49] handles the multi-instance images by predicting each alpha matte separately from binary guided masks before the instance refinement at the end. Although this approach produces impressive results in both synthesized and natural image benchmarks, the efficiency and accuracy of this model are unexplored in video processing. The separated prediction for each instance yields inefficiency in the architecture, which makes it costly to adapt to video input. Another concurrent work [30] with ours extends the InstMatt to process video input, but the complexity and efficiency of the network are unexplored. Fig. 1 illustrates the comparison between our MaGGIE and InstMatt when working with video. Our work improves not only the accuracy but also the consistency between frames when errors occur in guidance.

Besides the temporal consistency, when extending the instance matting to videos containing a large number of frames and instances, the careful network design to prevent the explosion in the computational cost is also a key challenge. In this work, we propose several adjustments to the popular mask-guided progressive refinement architecture [56]. Firstly, by using the mask guidance embedding inspired by AOT [55], the input size reduces to a constant number of channels. Secondly, with the advancement of transformer attention in various vision tasks [40–42], we inherit the query-based instance segmentation [7, 19, 23] to predict instance mattes in one forward pass instead of separated estimation. It also replaces the complex refinement in previous work with the interaction between instances by attention mechanism. To save the high cost of transformer attention, we only perform multi-instance prediction at the coarse level and adapt the progressive refinement at multiple scales [18, 56]. However, using full convolution for the refinement as previous works are inefficient as less than 10% of values are updated at each scale, which is also mentioned in [50]. The replacement of sparse convolution [36] saves the inference cost significantly, keeping the constant complexity of the algorithm since only interested locations are refined. Nevertheless, the lack of information at a larger scale when using sparse convolution can cause a dominance problem, which leads to the higher-scale prediction copying

the lower outputs without adding fine-grained details. We propose an instance guidance method to help the coarser prediction guide but not contribute to the finer alpha matte.

In addition to the framework design, we propose a new training video dataset and benchmarks for instance-awareness matting. Besides the new large-scale high-quality synthesized image instance matting, an extension of the current instance image matting benchmark adds more robustness with different guidance quality. For video input, our synthesized training and benchmark are constructed from various public instance-agnostic datasets with three levels of difficulty.

In summary, our contributions include:

- A highly efficient instance matting framework with mask guidance that has all instances interacting and processed in a single forward pass.
- A novel approach that considers feature-matte levels to maintain matte temporal consistency in videos.
- Diverse training datasets and robust benchmarks for image and video instance matting that bridge the gap between synthesized and natural cases.

2. Related Works

There are many ways to categorize matting methods, here we revise previous works based on their primary input types. The brief comparison of others and our MaGGIE is shown in Table 1.

Image Matting. Traditional matting methods [4, 24, 25] rely on color sampling to estimate foreground and background, often resulting in noisy outcomes due to limited high-level object features. Advanced deep learning-based methods [9, 11, 31, 37, 46, 47, 54] have significantly improved results by integrating image and trimap inputs or focusing on high-level and detailed feature learning. However, these methods often struggle with trimap inaccuracies and assume single-object scenarios. Recent approaches [5, 6, 22] require only image inputs but face challenges with multiple salient objects. MGM [56] and its extension MGM-in-the-wild [39] introduce binary mask-based matting, addressing multi-salient object issues and reducing trimap dependency. InstMatt [49] further customizes this approach for multi-instance scenarios with a complex refinement algorithm. Our work extends these developments, focusing on efficient, end-to-end instance matting with binary mask guidance. Image matting also benefits from diverse datasets [22, 26, 27, 29, 33, 50, 54], supplemented by background augmentation from sources like BG20K [29] or COCO [35]. Our work also leverages currently available datasets to concretize a robust benchmark for human-masked guided instance matting.

Video Matting. Temporal consistency is a key challenge in video matting. Trimap-propagation methods [17,

Table 1. **Comparing MaGGIe with previous works in image and video matting.** Our work is the first instance-aware framework producing alpha matte from a binary mask with both feature and output temporal consistency in constant processing time.

Method	Avenue	Guidance	Instance -awareness	Temp. aggre.		Time complexity
				Feat.	Matte.	
MGM [39, 56]	CVPR21+23	Mask				$O(n)$
InstMatt [49]	CVPR22	Mask	✓			$O(n)$
TCVOM [57]	MM21	-	-	✓		-
OTVM [45]	ECCV22	1st trimap		✓		$O(n)$
FTP-VM [17]	CVPR23	1st trimap		✓		$O(n)$
SparseMatt [50]	CVPR23	No			✓	$O(n)$
MaGGIe	-	Mask	✓	✓	✓	$\approx O(1)$

45, 48] and background knowledge-based approaches like BGMv2 [33] aim to reduce trimap dependency. Recent techniques [28, 32, 34, 53, 57] incorporate ConvGRU, attention memory matching, or transformer-based architectures for temporal feature aggregation. SparseMat [50] uniquely focuses on fusing outputs for consistency. Our approach builds on these foundations, combining feature and output fusion for enhanced temporal consistency in alpha maps. There is a lack of video matting datasets due to the difficulty in data collecting. VideoMatte240K [33] and VM108 [57] focus on composited videos, while CRGNN [52] is the only offering natural videos for human matting. To address the gap in instance-aware video matting datasets, we propose adapting existing public datasets for training and evaluation, particularly for human subjects.

3. MaGGIe

We introduce our efficient instance matting framework guided by instance binary masks, structured into two parts. The first Sec. 3.1 details our novel architecture to maintain accuracy and efficiency. The second Sec. 3.2 describes our approach for ensuring temporal consistency across frames in video processing.

3.1. Efficient Masked Guided Instance Matting

Our framework, depicted in Fig. 2, processes images or video frames $\mathbf{I} \in [0, 255]^{T \times 3 \times H \times W}$ with corresponding binary instance guidance masks $\mathbf{M} \in \{0, 1\}^{T \times N \times H \times W}$, and then predicts alpha mattes $\mathbf{A} \in [0, 1]^{T \times N \times H \times W}$ for each instance per frame. Here, T, N, H, W represent the number of frames, instances, and input resolution, respectively. Each spatial-temporal location (x, y, t) in \mathbf{M} is a one-hot vector $\{0, 1\}^N$ highlighting the instance it belongs to. The pipeline comprises five stages: (1) Input construction; (2) Image features extraction; (3) Coarse instance alpha mattes prediction; (4) Progressive detail refinement; and (5) Coarse-to-fine fusion.

Input Construction. The input $\mathbf{I}' \in \mathbb{R}^{T \times (3+C_e) \times H \times W}$ to our model is the concatenation of input image \mathbf{I} and guidance embedding $\mathbf{E} \in \mathbb{R}^{T \times C_e \times H \times W}$ constructed from \mathbf{M} by ID Embedding layer [55]. More details about transforming \mathbf{M} to \mathbf{E} are in the supplementary material.

Image Features Extraction. We extract features map $\mathbf{F}_s \in \mathbb{R}^{T \times C_s \times H/s \times W/s}$ from \mathbf{I}' by feature-pyramid networks. As shown in the left part of Fig. 2, there are four scales $s = 1, 2, 4, 8$ for our coarse-to-fine matting pipeline.

Coarse instance alpha mattes prediction. Our MaGGIe adopts transformer-style attention to predict instance mattes at the coarsest features \mathbf{F}_8 . We revisit the scaled dot-product attention mechanism in Transformers [51]. Given queries $\mathbf{Q} \in \mathbb{R}^{L \times C}$, keys $\mathbf{K} \in \mathbb{R}^{S \times C}$, and values $\mathbf{V} \in \mathbb{R}^{S \times C}$, the scaled dot-product attention is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}}\right)\mathbf{V}. \quad (1)$$

In cross-attention (CA), \mathbf{Q} and (\mathbf{K}, \mathbf{V}) originate from different sources, whereas in self-attention (SA), they share similar information.

In our Instance Matte Decoder, the organization of CA and SA blocks inspired by SAM [23] is depicted in the bottom right of Fig. 2. The downscaled guidance masks \mathbf{M}_8 also participate as the additional embedding for image features in attention procedures. The coarse alpha matte \mathbf{A}_8 is computed as the dot product between instance tokens $\mathbf{T} = \{\mathbf{T}_i | 1 \leq i \leq N\} \in \mathbb{R}^{N \times C_s}$ and enriched feature map $\bar{\mathbf{F}}_8$ with a sigmoid activation applied. Those components are used in the following steps of matte detail refinement.

Progressive Detail Refinement. From the coarse instance alpha matte, we leverage the Progressive Refinement [56] to improve the details at uncertain locations $\mathbf{U} = \{u_p = (x, y, t, i) | 0 < \mathbf{A}_8(u_p) < 1\} \in \mathbb{N}^{P \times 4}$ with some highly efficient modifications. It is mandatory to transform enriched dense features $\bar{\mathbf{F}}_8$ to instance-specific features \mathbf{X}_8 for the instance-wise refinement. However, to save memory and computational costs, only transformed features at uncertainty \mathbf{U} are computed as:

$$\mathbf{X}_8(x, y, t, i) = \text{MLP}(\bar{\mathbf{F}}_8(x, y, t) \times \mathbf{T}_i). \quad (2)$$

To combine the coarser instance-specific sparse features \mathbf{X}_8 with the finer image features \mathbf{F}_4 , we propose the Instance Guidance (IG) module. As described in the top right of Fig. 2, this module firstly increases the spatial scale of \mathbf{X}_8 to have \mathbf{X}'_4 by an inverse sparse convolution. For each entry p , we compute a guidance score $\mathbf{G} \in [0, 1]^{C_4}$, which is then channel-wise multiplied with \mathbf{F}_4 to produce detailed sparse instance-specific features \mathbf{X}_4 :

$$\mathbf{X}_4(p) = \mathcal{G}(\{\mathbf{X}'_4(p); \mathbf{F}_4(p)\}) \times \mathbf{F}_4(p), \quad (3)$$

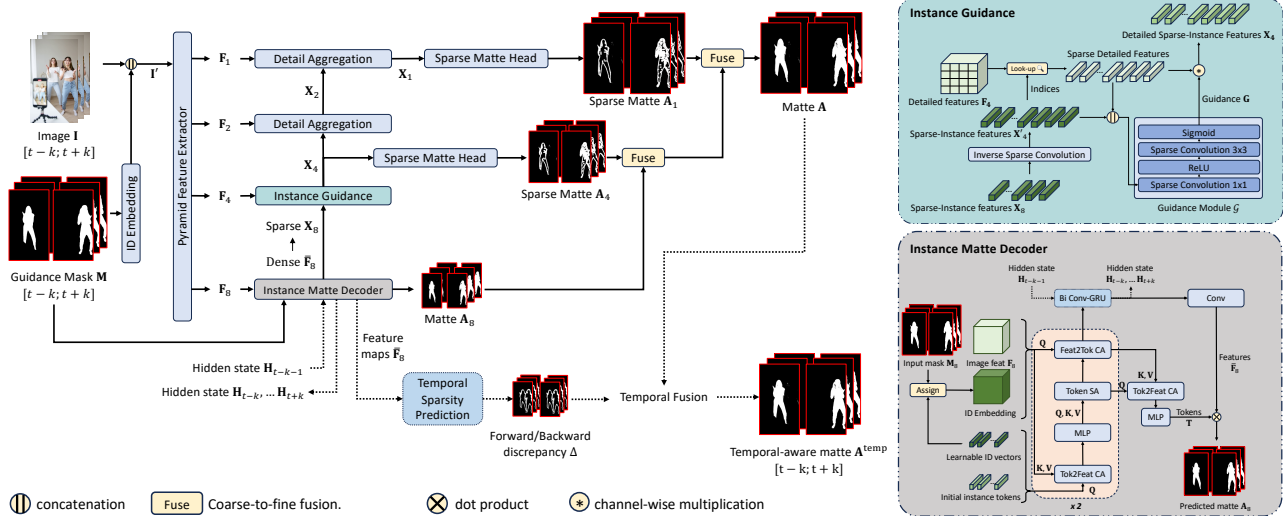


Figure 2. **Overall pipeline of MaGgIE.** This framework processes frame sequences I and instance masks M to generate per-instance alpha mattes A' for each frame. It employs progressive refinement and sparse convolutions for accurate mattes in multi-instance scenarios, optimizing computational efficiency. The subfigures on the right illustrate the Instance Matte Decoder and the Instance Guidance, where we use mask guidance to predict coarse instance mattes and guide detail refinement by deep features, respectively. (Optimal in color and zoomed view).

where $\{;\}$ denotes concatenation along the feature dimension, and \mathcal{G} is a series of sparse convolutions with sigmoid activation.

The sparse features X_4 is then aggregated with other dense features F_2, F_1 respectively at corresponding indices to have X_2, X_1 . At each scale, we predict alpha matte A_4, A_1 with gradual detail improvement. You can find more aggregation and sparse matting head details in the supplementary material.

Coarse-to-fine fusion. This stage is to combine alpha mattes of different scales in a progressive way (PRM): $A_8 \rightarrow A_4 \rightarrow A_1$ to obtain A . At each step, only values at uncertain locations and belonging to unknown masks are refined.

Training Losses. In addition to standard losses (\mathcal{L}_1 for reconstruction, Laplacian \mathcal{L}_{lap} for detail, Gradient $\mathcal{L}_{\text{grad}}$ for smoothness), we supervise the affinity score matrix Aff between instance tokens T (as Q) and image feature maps F (as K, V) by the attention loss \mathcal{L}_{att} . Additionally, our network’s progressive refinement process necessitates accurate coarse-level predictions to determine U accurately. We assign customized weights W_8 for losses at scale $s = 8$ to prioritize uncertain locations. More details about \mathcal{L}_{att} and W_8 is in the supplementary material.

3.2. Feature-Matte Temporal Consistency

We propose to enhance temporal consistency at both feature and alpha matte levels.

Feature Temporal Consistency. Utilizing Conv-GRU [3] for video inputs, we ensure bidirectional consistency among feature maps of adjacent frames. With a temporal window size k , bidirectional Conv-GRU processes frames $\{t -$

$k, \dots, t + k\}$, as shown in Fig. 2. For simplicity, we set $k = 1$ with an overlap of 2 frames. The initial hidden state H_0 is zeroed, and H_{t-k-1} from the previous window aids the current one. This module fuses the feature map at time t with two consecutive frames, averaging forward and backward aggregations. The resultant temporal features are used to predict the coarse alpha matte A_8 .

Alpha Matte Temporal Consistency. We propose fusing frame mattes by predicting their temporal sparsity. Unlike the previous method [50] using image processing kernels, we leverage deep features for this prediction. A shallow convolutional network with a sigmoid activation processes stacked feature maps \bar{F}_8 at $t - 1$ and t , outputting alpha matte discrepancy between two frames $\Delta(t) \in \{0, 1\}^{H \times W}$. For each frame t , with $\Delta(t)$ and $\Delta(t + 1)$, we compute the forward propagation A^f and backward propagation A^b to reject the propagation at misalignment regions and obtain temporal aware output A^{temp} . The supplementary material provides more details about the implementation.

Training Losses. Besides the dtSSD loss for temporal consistency, we introduce an L1 loss for the alpha matte discrepancy. The loss compares predicted $\Delta(t)$ with the ground truth $\Delta^{gt}(t) = \max_i (|A^{gt}(t - 1, i) - A^{gt}(t, i)| > \beta)$, where $\beta = 0.001$ to simplify the problem to binary pixel classification.

4. Instance Matting Datasets

This section outlines the datasets used in our experiments. With the lack of public datasets for the instance matting task, we synthesized training data from existing public instance-agnostic sources. Our evaluation combines syn-



Figure 3. **Variations of Masks for the Same Image in M-HIM2K Dataset.** Masks generated using R50-C4-3x, R50-FPN-3x, R101-FPN-400e MaskRCNN models trained on COCO. (Optimal in color).

thetic and natural sets to assess the model’s robustness and generalization.

4.1. Image Instance Matting

We derived the **Image Human Instance Matting 50K (I-HIM50K)** training dataset from HHM50K [50], featuring multiple human subjects. This dataset includes 49,737 synthesized images with 2-5 instances each, created by compositing human foregrounds with random backgrounds and modifying alpha mattes for guidance binary masks. For benchmarking, we used HIM2K [49] and created the **Mask HIM2K (M-HIM2K)** set to test robustness against varying mask qualities from available instance segmentation models (as shown in Fig. 3). Details on the generation process are available in the supplementary material.

4.2. Video Instance Matting

Our video instance matte dataset, synthesized from VM108 [57], VideoMatte240K [33], and CRGNN [52], includes subsets **V-HIM2K5** for training and **V-HIM60** for testing. We categorized the dataset into three difficulty levels based on instance overlap. Table 2 shows some details of the synthesized datasets. Masks in training involved dilation and erosion on binarized alpha mattes. For testing, masks are generated using XMem [8]. Further details on dataset synthesis and difficulty levels are provided in the supplementary material.

5. Experiments

We developed our model using PyTorch [20] and the Sparse convolution library Spconv [10]. Our codebase is built upon the publicly available implementations of MGM [56] and

Table 2. **Details of Video Instance Matting Training and Testing Sets.** V-HIM2K5 for training and V-HIM60 for model evaluation. Each video contains 30 frames.

Name	Sources			# videos			# instance/video		
	[57]	[33]	[52]	Easy	Med.	Hard	Easy	Med.	Hard
V-HIM2K5	33	410	0	500	1,294	667	2.67	2.65	3.21
V-HIM60	3	8	18	20	20	20	2.35	2.15	2.70

Table 3. **Superiority of Mask Embedding Over Stacking in HIM2K+M-HIM2K.** Our mask embedding technique demonstrates enhanced performance compared to traditional stacking methods.

Mask input	Composition			Natural		
	MAD	Grad	Conn	MAD	Grad	Conn
Stacked	27.01	16.80	15.72	39.29	16.44	23.26
Embedded($C_e = 1$)	19.18	13.00	11.16	33.60	13.44	19.18
Embedded($C_e = 2$)	21.74	14.39	12.69	35.16	14.51	20.40
Embedded($C_e = 3$)	17.75	12.52	10.32	33.06	13.11	17.30
Embedded($C_e = 5$)	24.79	16.19	14.58	34.25	15.66	19.70

OTVM [45]. In the first Sec. 5.1, we discuss the results when pre-training on the image matting dataset. The performance on the video dataset is shown in the Sec. 5.2. All training settings are reported in the supplementary material.

5.1. Pre-training on image data

Metrics. Our evaluation metrics included Mean Absolute Differences (MAD), Mean Squared Error (MSE), Gradient (Grad), and Connectivity (Conn). We also separately computed these metrics for the foreground and unknown regions, denoted as MAD_f and MAD_u , by estimating the trimap on the ground truth. Since our images contain multiple instances, metrics were calculated for each instance individually and then averaged. We did not use the IMQ from InstMatt, as our focus is not on instance detection.

Ablation studies. Each ablation study setting was trained for 10,000 iterations with a batch size 96. We first assessed the performance of the embedding layer versus stacked masks and image inputs in Table 3. The mean results on M-HIM2K are reported, with full results in the supplementary material. The embedding layer showed improved performance, particularly effective with $C_e = 3$. We also evaluated the impact of using \mathcal{L}_{att} and \mathbf{W}_8 in training in Table 4. \mathcal{L}_{att} significantly enhanced model performance, while \mathbf{W}_8 provided a slight boost.

Quantitative results. We evaluated our model against previous baselines after retraining them on our I-HIM50K dataset. Besides original works, we modified SparseMat’s

Table 4. **Optimal Performance with \mathcal{L}_{att} and \mathbf{W}_8 on HIM2K+M-HIM2K.** Utilizing both \mathcal{L}_{att} and \mathbf{W}_8 leads to superior results.

\mathcal{L}_{att}	\mathbf{W}_8	Composition			Natural		
		MAD	Grad	Conn	MAD	Grad	Conn
		31.77	16.58	18.27	46.68	15.68	30.64
	✓	25.41	14.53	14.75	46.30	15.84	29.26
✓		17.56	12.34	10.22	32.95	13.29	17.06
✓	✓	17.55	12.34	10.19	32.03	13.16	17.43

Table 5. **Comparative Performance on HIM2K+M-HIM2K.** Our method outperforms baselines, with average results (large numbers) and standard deviations (small numbers) on the benchmark. The upper group represents methods predicting each instance separately, while the lower models utilize instance information. Gray rows denote public weights trained on external data, not retrained on I-HIM50K. MGM[†] denotes the MGM-in-the-wild. MGM* refers to MGM with all masks stacked with the input image. Models are tested on images with a short side of 576px. **Bold** and underline highlight the best and second-best models per metric, respectively.

Method	Composition set						Natural set					
	MAD	MSE	Grad	Conn	MAD _f	MAD _u	MAD	MSE	Grad	Conn	MAD _f	MAD _u
<i>Instance-agnostic</i>												
MGM [†] [39]	23.15 (1.5)	14.76 (1.3)	12.75 (0.5)	13.30 (0.9)	64.39 (4.5)	309.38 (12.0)	32.52 (6.7)	18.80 (6.0)	12.52 (1.2)	18.51 (18.5)	65.20 (15.9)	179.76 (23.9)
MGM [56]	15.32 (0.6)	9.13 (0.5)	<u>9.94 (0.2)</u>	8.83 (0.3)	<u>33.54 (1.9)</u>	261.43 (4.0)	30.23 (3.6)	17.40 (3.3)	<u>10.53 (0.5)</u>	15.70 (1.9)	63.16 (13.0)	167.35 (12.1)
SparseMat [50]	21.05 (1.2)	14.55 (1.0)	14.64 (0.5)	12.26 (0.7)	45.19 (2.9)	352.95 (14.2)	35.03 (5.1)	21.79 (4.7)	15.85 (1.2)	18.50 (3.1)	67.82 (15.2)	212.63 (20.8)
<i>Instance-aware</i>												
InstMatt [49]	12.85 (0.2)	5.71 (0.2)	9.41 (0.1)	7.19 (0.1)	22.24 (1.3)	255.61 (2.0)	26.76 (2.5)	12.52 (2.0)	10.20 (0.3)	13.81 (1.1)	48.63 (6.8)	161.52 (6.9)
InstMatt [49]	16.99 (0.7)	9.70 (0.5)	10.93 (0.3)	9.74 (0.5)	53.76 (3.0)	286.90 (7.0)	<u>28.16 (4.5)</u>	14.30 (3.7)	10.98 (0.7)	<u>14.63 (2.0)</u>	57.83 (12.1)	168.74 (15.5)
MGM*	<u>14.31 (0.4)</u>	<u>7.89 (0.4)</u>	10.12 (0.2)	<u>8.01 (0.2)</u>	41.94 (3.1)	<u>251.08 (3.6)</u>	31.38 (3.3)	18.38 (3.1)	10.97 (0.4)	14.75 (1.4)	<u>53.89 (9.6)</u>	<u>165.13 (10.6)</u>
MaGGIe (ours)	12.93 (0.3)	7.26 (0.3)	8.91 (0.1)	7.37 (0.2)	19.54 (1.0)	235.95 (3.4)	27.17 (3.3)	<u>16.09 (3.2)</u>	9.94 (0.6)	13.42 (1.4)	49.52 (8.0)	146.71 (11.6)

first layer to accept a single mask input. Additionally, we expanded MGM to handle up to 10 instances, denoted as MGM*. We also include the public weights of InstMatt [49] and MGM-in-the-wild [39]. The performance with different masks M-HIM2K are reported in Table 5. The public InstMatt showed the best performance, but this comparison may not be entirely fair as it was trained on private external data. Our model demonstrated comparable results on composite and natural sets, achieving the lowest error in most metrics. MGM* also performed well, suggesting that processing multiple masks simultaneously can facilitate instance interaction, although this approach slightly impacted the Grad metric, which reflects the output’s detail.

We also measure the memory and speed of models on M-HIM2K natural set in Fig. 4. While InstMatt, MGM, and SparseMat have the inference time increasing linearly to the number of instances, MGM* and ours keep steady performance in both memory and speed.

Qualitative results. MaGGIe’s ability to capture fine details and effectively separate instances is showcased in Fig. 5. At the exact resolution, our model not only achieves highly detailed outcomes comparable to running MGM separately for each instance but also surpasses both

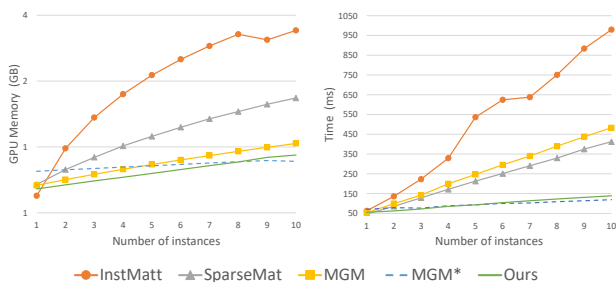


Figure 4. **Our model keeps steady memory and time complexity when the number of instance increases.** InstMatt’s complexity increases linearly with the number of instances.

the public and retrained versions of InstMatt. A key strength of our approach is its proficiency in distinguishing between different instances. This is particularly evident when compared to MGM, where we observed overlapping instances, and MGM*, which has noise issues caused by processing multiple masks simultaneously. Our model’s refined instance separation capabilities highlight its effectiveness in handling complex matting scenarios.

5.2. Training on video data

Temporal consistency metrics. Following previous works [45, 48, 57], we extended our evaluation metrics to include dtSSD and MESSDdt to assess the temporal consistency of instance matting across frames.

Ablation studies. Our tests, detailed in Table 6, show that each temporal module significantly impacts performance. Omitting these modules increased errors in all subsets. Single-direction Conv-GRU use improved outcomes, with further gains from adding backward pass fusion. Forward fusion alone was less effective, possibly due to error propagation. The optimal setup involved combining backward propagation to reduce errors, yielding the best results.

Performance evaluation. Our model was benchmarked

Table 6. **Superiority of Temporal Consistency in Feature and Prediction Levels.** Our MaGGIe, integrating temporal consistency at both feature and matte levels, outperforms non-temporal methods and those with only feature level.

Conv-GRU	Fusion	Easy		Medium		Hard			
		MAD	dtSSD	MAD	dtSSD	MAD	dtSSD		
Single	Bi	\hat{A}^f	\hat{A}^b	10.26	16.57	13.88	23.67	21.62	30.50
✓				10.15	16.42	13.84	23.66	21.26	29.95
	✓			10.14	16.41	13.83	23.66	21.25	29.92
	✓	✓		11.32	16.51	15.33	24.08	24.97	30.66
	✓	✓	✓	10.12	16.40	13.85	23.63	21.23	29.90

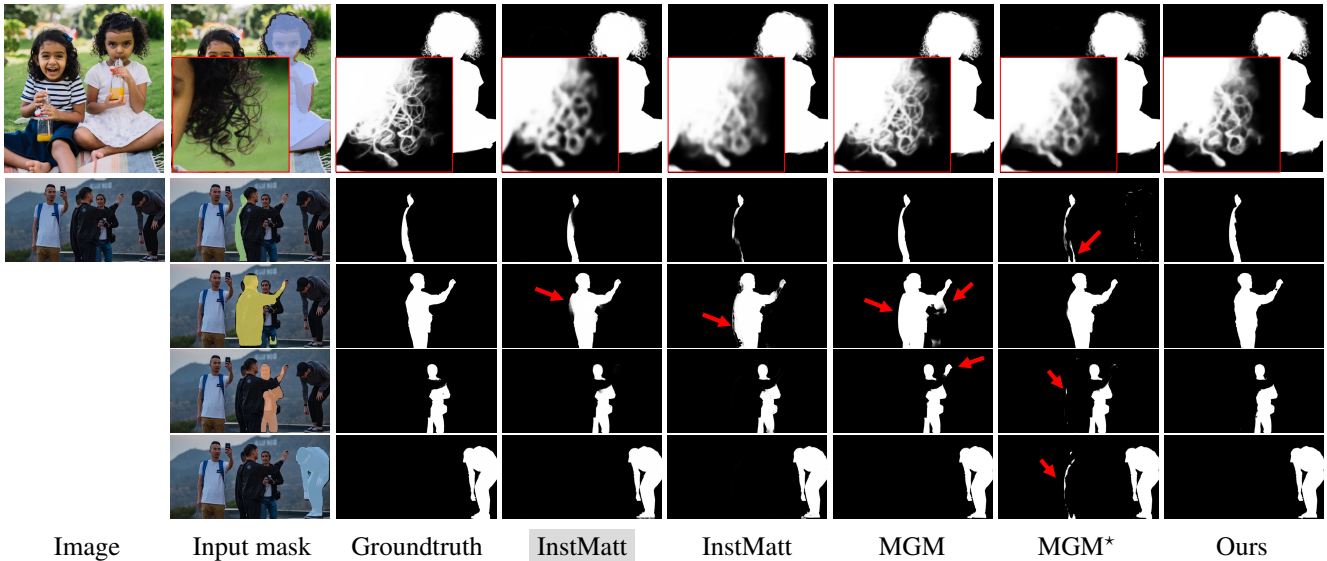


Figure 5. **Enhanced Detail and Instance Separation by MaGGie.** Our model excels in rendering detailed outputs and effectively separating instances, as highlighted by **red squares** (detail focus) and **red arrows** (errors in other methods).

Table 7. **Comparative Analysis of Video Matting Methods on V-HIM60.** This table categorizes methods into two groups: those utilizing first-frame trimaps (upper group) and mask-guided approaches (lower group). Gray rows denotes models with public weights not retrained on I-HIM50K and V-HIM50K. MGM*-TCVOM represents MGM with stacked guidance masks and the TCVOM temporal module. **Bold** and underline highlight the top and second-best performing models in each metric, respectively.

Method	Easy					Medium					Hard				
	MAD	Grad	Conn	dtSSD	MESSDdt	MAD	Grad	Conn	dtSSD	MESSDdt	MAD	Grad	Conn	dtSSD	MESSDdt
First-frame trimap															
OTVM [45]	204.59	15.25	76.36	46.58	397.59	247.97	21.02	97.74	66.09	587.47	412.41	29.97	146.11	90.15	764.36
OTVM [45]	36.56	6.62	14.01	24.86	69.26	48.59	10.19	17.03	36.06	80.38	140.96	17.60	47.84	59.66	298.46
FTP-VM [17]	12.69	6.03	4.27	19.83	18.77	40.46	12.18	15.13	32.96	125.73	46.77	14.40	15.82	45.04	76.48
FTP-VM [17]	13.69	6.69	4.78	20.51	22.54	26.86	12.39	9.95	32.64	126.14	48.11	14.87	16.12	45.29	78.66
Frame-by-frame binary mask															
MGM-TCVOM [45]	11.36	4.57	3.83	17.02	19.69	14.76	7.17	5.41	23.39	<u>39.22</u>	<u>22.16</u>	7.91	<u>7.27</u>	<u>31.00</u>	47.82
MGM*-TCVOM [45]	<u>10.97</u>	4.19	<u>3.70</u>	<u>16.86</u>	15.63	13.76	<u>6.47</u>	5.02	23.99	42.71	22.59	<u>7.86</u>	7.32	32.75	37.83
InstMatt [49]	13.77	4.95	3.98	17.86	18.22	19.34	7.21	6.02	24.98	54.27	27.24	7.88	8.02	31.89	47.19
SparseMat [50]	12.02	4.49	4.11	19.86	24.75	18.20	8.03	6.87	30.19	85.79	24.83	8.47	8.19	36.92	55.98
MaGGie (ours)	10.12	4.08	3.43	16.40	<u>16.41</u>	<u>13.85</u>	6.31	<u>5.11</u>	<u>23.63</u>	38.12	21.23	7.08	6.89	29.90	<u>42.98</u>

against leading methods in trimap video matting, mask-guided matting, and instance matting. For trimap video matting, we chose OTVM [45] and FTP-VM [17], fine-tuning them on our V-HIM2K5 dataset. In masked guided video matting, we compared our model with InstMatt [49], SparseMat [50], and MGM [56] which is combined with the TCVOM [57] module for temporal consistency. InstMatt, after being fine-tuned on I-HIM50K and subsequently on V-HIM2K5, processed each frame in the test set independently, without temporal awareness. SparseMat, featuring a temporal sparsity fusion module, was fine-tuned under the same conditions as our model. MGM and its variant, integrated with the TCVOM module, emerged as strong com-

petitors in our experiments, demonstrating their robustness in maintaining temporal consistency across frames.

The comprehensive results of our model across three test sets, using masks from XMem, are detailed in Table 7. All trimap propagation methods are underperform the mask-guided solutions. When benchmarked against other masked guided matting methods, our approach consistently reduces error across most metrics. Notably, it excels in temporal consistency, evidenced by its top performance in dtSSD for both easy and hard test sets, and in MESSDdt for the medium set. Additionally, our model shows superior performance in capturing fine details, as indicated by its leading scores in the Grad metric across all test sets. These re-

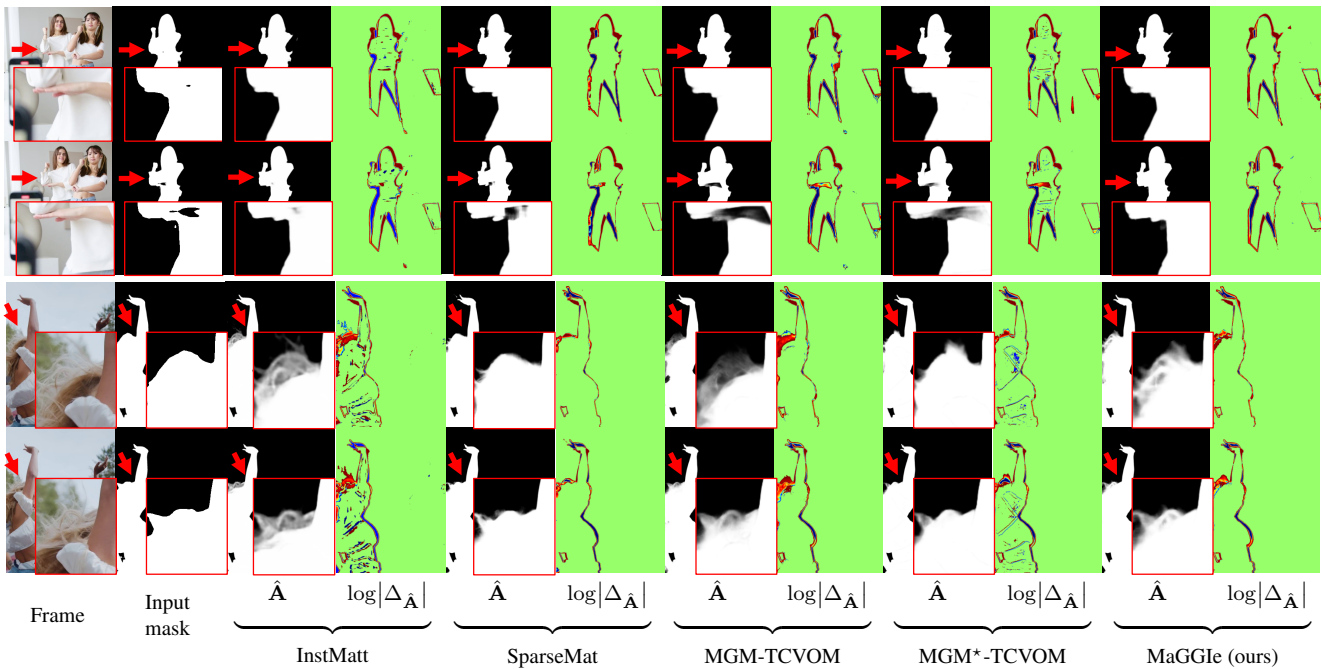



Figure 6. **Detail and Consistency in Frame-to-Frame Predictions.** This figure demonstrates the precision and temporal consistency of our model’s alpha matte predictions, highlighting robustness against noise from input masks. The color-coded map (min-max range) to illustrate differences between consecutive frames is .

sults underscore our model’s effectiveness in video instance matting, particularly in challenging scenarios requiring high temporal consistency and detail preservation.

Temporal consistency and detail preservation. Our model’s effectiveness in video instance matting is evident in Fig. 6 with natural videos. Key highlights include:

- *Handling of Random Noises:* Our method effectively handles random noise in mask inputs, outperforming others that struggle with inconsistent input mask quality.
- *Foreground/Background Region Consistency:* We maintain consistent, accurate foreground predictions across frames, surpassing InstMatt and MGM*-TCVOM.
- *Detail Preservation:* Our model retains intricate details, matching InstMatt’s quality and outperforming MGM variants in video inputs.

These aspects underscore MaGGie’s robustness and effectiveness in video instance matting, particularly in maintaining temporal consistency and preserving fine details across frames.

6. Discussion

Limitation and Future work. Our MaGGie demonstrates effective performance in human video instance matting with binary mask guidance, yet it also presents opportunities for further research and development. One notable limitation is the reliance on one-hot vector representation for each location in the guidance mask, necessitating that each pixel is distinctly associated with a single instance. This require-

ment can pose challenges, particularly when integrating instance masks from varied sources, potentially leading to misalignments in certain regions. Additionally, the use of composite training datasets may constrain the model’s ability to generalize effectively to natural, real-world scenarios. While the creation of a comprehensive natural dataset remains a valuable goal, we propose an interim solution: the utilization of segmentation datasets combined with self-supervised or weakly-supervised learning techniques. This approach could enhance the model’s adaptability and performance in more diverse and realistic settings, paving the way for future advancements in the field.

Conclusion. Our study contributes to the evolving field of instance matting, with a focus that extends beyond human subjects. By integrating advanced techniques like transformer attention and sparse convolution, MaGGie shows promising improvements over previous methods in detailed accuracy, temporal consistency, and computational efficiency for both image and video inputs. Additionally, our approach in synthesizing training data and developing a comprehensive benchmarking schema offers a new way to evaluate the robustness and effectiveness of models in instance matting tasks. This work represents a step forward in video instance matting and provides a foundation for future research in this area.

Acknowledgement. We sincerely appreciate Markus Woodson for the invaluable initial discussions. Additionally, I am deeply thankful to my wife, Quynh Phung, for her meticulous proofreading and feedback.

References

- [1] Adobe. Adobe premiere. <https://www.adobe.com/products/premiere.html>, 2023. 1
- [2] Apple. Cutouts object ios 16. <https://support.apple.com/en-hk/102460>, 2023. 1
- [3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015. 4
- [4] Arie Berman, Arpag Dadourian, and Paul Vlahos. Method for removing from an image the background surrounding a selected object, 2000. US Patent 6,134,346. 2
- [5] Guowei Chen, Yi Liu, Jian Wang, Juncai Peng, Yuying Hao, Lutao Chu, Shiyu Tang, Zewu Wu, Zeyu Chen, Zhiliang Yu, et al. Pp-matting: high-accuracy natural image matting. *arXiv preprint arXiv:2204.09433*, 2022. 2
- [6] Xiangguang Chen, Ye Zhu, Yu Li, Bingtao Fu, Lei Sun, Ying Shan, and Shan Liu. Robust human matting via semantic guidance. In *ACCV*, 2022. 2
- [7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2
- [8] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 1, 5
- [9] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *ECCV*, 2016. 2
- [10] Spconv Contributors. Spconv: Spatially sparse convolution library. <https://github.com/traveller59/spconv>, 2022. 5
- [11] Marco Forte and François Pitié. f , b , alpha matting. *arXiv preprint arXiv:2003.07711*, 2020. 1, 2
- [12] Google. Magic editor in google pixel 8. https://pixel.withgoogle.com/Pixel_8_Pro/use-magic-editor, 2023. 1
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 11
- [14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 13
- [15] Anna Katharina Hebborn, Nils Höhner, and Stefan Müller. Occlusion matting: realistic occlusion handling for augmented reality applications. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2017. 1
- [16] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 2019. 1
- [17] Wei-Lun Huang and Ming-Sui Lee. End-to-end video matting with trimap propagation. In *CVPR*, 2023. 1, 2, 3, 7, 23
- [18] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *CVPR*, 2021. 2
- [19] Chuong Huynh, Yuqian Zhou, Zhe Lin, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Abhinav Shrivastava. Simpson: Simplifying photo cleanup with single-click distracting object segmentation network. In *CVPR*, 2023. 2
- [20] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. Pytorch. *Programming with TensorFlow: Solution for Edge Computing Applications*, 2021. 5
- [21] Lei Ke, Henghui Ding, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Video mask transfiner for high-quality video instance segmentation. In *ECCV*, 2022. 2
- [22] 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. 2
- [23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 2, 3
- [24] Philip Lee and Ying Wu. Nonlocal matting. In *CVPR*, 2011. 2
- [25] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE TPAMI*, 30(2), 2007. 2
- [26] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACM MM*, 2021. 2
- [27] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In *IJCAI*, 2021. 2
- [28] Jiachen Li, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, Yunchao Wei, and Humphrey Shi. Vmformer: End-to-end video matting with transformer. *arXiv preprint arXiv:2208.12801*, 2022. 3
- [29] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *IJCV*, 2022. 2, 13
- [30] Jiachen Li, Roberto Henschel, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, and Humphrey Shi. Video instance matting. In *WACV*, 2024. 2
- [31] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI*, 2020. 1, 2
- [32] Chung-Ching Lin, Jiang Wang, Kun Luo, Kevin Lin, Linjie Li, Lijuan Wang, and Zicheng Liu. Adaptive human matting for dynamic videos. In *CVPR*, 2023. 2, 3
- [33] 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. 2, 3, 5
- [34] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *WACV*, 2022. 2, 3
- [35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2
- [36] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *CVPR*, 2015. 2
- [37] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *CVPR*, 2019. 1, 2

- [38] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1
- [39] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Mask-guided matting in the wild. In *CVPR*, 2023. 1, 2, 3, 6, 19
- [40] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Improving closed and open-vocabulary attribute prediction using transformers. In *ECCV*, 2022. 2
- [41] Khoi Pham, Chuong Huynh, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *CVPR*, 2024.
- [42] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *CVPR*, 2024. 2
- [43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 13
- [44] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 1
- [45] Hongje Seong, Seoung Wug Oh, Brian Price, Euntai Kim, and Joon-Young Lee. One-trimap video matting. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7, 23
- [46] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, 2016. 2
- [47] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *CVPR*, 2021. 2
- [48] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *CVPR*, 2021. 3, 6
- [49] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 11, 13, 14, 16, 17, 18, 20
- [50] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Ultrahigh resolution image/video matting with spatio-temporal sparsity. In *CVPR*, 2023. 2, 3, 4, 5, 6, 7, 12, 16, 17, 18, 20
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3
- [52] Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, and Ming-Hsuan Yang. Video matting via consistency-regularized graph neural networks. In *ICCV*, 2021. 3, 5
- [53] Yumeng Wang, Bo Xu, Ziwen Li, Han Huang, Cheng Lu, and Yandong Guo. Video object matting via hierarchical space-time semantic guidance. In *WACV*, 2023. 2, 3
- [54] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017. 2
- [55] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *NeurIPS*, 2021. 2, 3, 11
- [56] 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. 1, 2, 3, 5, 6, 7, 11, 13, 16, 17, 18, 19
- [57] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuan-song Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *ACM MM*, 2021. 3, 5, 6, 7