Video Object Matting via Hierarchical Space-Time Semantic Guidance

Yumeng Wang[1,2,*], Bo Xu[1,*], Ziwen Li[1], Han Huang[1], Cheng Lu[3], and Yandong Guo[1†]
[1]OPPO Research Institute, [2]Northwestern Polytechnical University, [3]Xmotors
yandong.guo@live.com

Abstract

Different from most existing approaches that require trimap generation for each frame, we reformulate video object matting (VOM) by introducing improved semantic guidance propagation. The proposed approach can achieve a higher degree of temporal coherence between frames with only a single coarse mask as a reference. In this paper, we adapt the hierarchical memory matching mechanism into the space-time baseline to build an efficient and robust framework for semantic guidance propagation and alpha prediction. To enhance the temporal smoothness, we also propose a cross-frame attention refinement (CFAR) module that can refine the feature representations across multiple adjacent frames (both historical and current frames) based on the spatio-temporal correlation among the cross-frame pixels. Extensive experiments demonstrate the effectiveness of hierarchical spatio-temporal semantic guidance and the cross-video-frame attention refinement module, and our model outperforms the state-of-the-art VOM methods. We also analyze the significance of different components in our model.

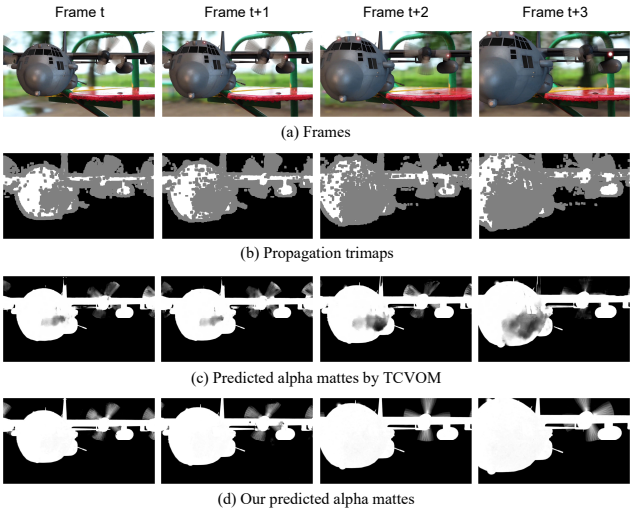


Figure 1. Visual comparisons between one state-of-the-art trimap-based method TCVOM [34] and ours. The trimap propagation network fails to find the ideal three-partition region distribution at some certain scenes, which may cause semantic information propagation failures in matting process. However, our hierarchical space-time semantic guidance VOM method can effectively maintain the integrity of semantic propagation.

1. Introduction

Video object matting (VOM) aims to identify and predict alpha mattes of one or multiple target foreground objects from consecutive video frames. This technology has been successfully applied in many areas where background replacement is needed, for example, live video creation, entertainment video creation, and special-effect film-making. Currently, matting is generally formulated as an image composite problem. It aims to solve the 7 unknown variables per pixel from only 3 known values, $I_i = \alpha_i F_i + (1 - \alpha_i) B_i$, where 3 dimensional RGB color I_i of pixel i, while foreground RGB color F_i, background RGB color B_i, and matte estimation α_i are unknown. Compared to image matting, one of the core challenges in video matting is to maintain the spatio-temporal coherence in alpha prediction. And for the video object matting (VOM), the target foreground of interest should be specified in advance before running the matting models.

The recently proposed algorithms [28, 34] utilize trimaps (a draft marking foreground, background, and unknown areas) as constraint information to locate the target area. This approach divides the matting process into two stages: trimap generation and trimap-based alpha prediction. It first generates trimap frame by frame by propagating one or several user-annotated trimaps to other target frames. Then the networks take the video frames and corresponding trimaps as inputs for alpha prediction. Although such trimap-based video object matting methods make the problem more tractable, there still remain two big challenges. First, given only one or several user-annotated reference frames, current trimap propagation networks struggle to find the ideal three-partition region distribution, which may cause semantic in-

*Yumeng Wang and Bo Xu contribute equally.
†The corresponding author.

formation propagation failures in the matting process, as shown in Figure 1. Second, manually checking the propagated trimaps frame by frame or densely interpolating user-annotated trimaps for reference can be quite burdensome for users.

To address the above issues, we introduce hierarchical semantic guidance in spatial-to-temporal space to guide the alpha prediction instead of the propagated trimaps. We employ the hierarchical memory matching mechanism on top of the Space-Time Correspondence Network (STCN) [3] baseline. Also, we build a novel hierarchical space-time semantic guidance video object matting (HSTSG) framework to achieve effective semantic guidance propagation and temporally coherent alpha prediction. Compared to the previous two-stage VOM methods [28, 34], our algorithm needs the annotated semantic mask of only the first frame as the target foreground reference. Besides, it combines semantic guidance propagation and alpha prediction into one unified task. To enhance the temporal smoothness of the predicted results, we propose a cross-frame attention refinement (CFAR) module that refines the feature representations of multiple adjacent frames (both historical and current frames) based on the spatio-temporal correlation among the cross-frame pixels. The CFAR can also improve the model's robustness when dealing with unfavorable scenarios such as occlusions, new targets, and so on.

To justify our solutions, we conduct extensive experiments on multiple public datasets. The experimental results show that our proposed method surpasses all the state-of-the-art VOM approaches. Overall, the contributions of this paper are as follows:

- We introduce the hierarchical semantic guidance into the spatial-to-temporal space to guide alpha prediction without per-frame trimap generation and achieve better semantic propagation.
- We propose an STCN-based spatio-temporal network with a hierarchical memory matching mechanism to establish stronger temporal coherence for alpha prediction. We also merge the semantic guidance propagation and alpha prediction into one task without relying on redundant trimap generation.
- We propose a cross-frame attention refinement (CFAR) module to improve the temporal smoothness across multiple adjacent frames.
- Extensive experiments demonstrate the effectiveness of our methods, outperforming the state-of-the-art (SOTA) approaches on multiple VOM benchmarks.

2. Related works

2.1. Image matting.

Trimap-based methods. Traditionally, most matting methods require a trimap as auxiliary information to compensate for the ill-posed nature of the matting equation. The trimap is annotated by humans and contains foreground, background, and unknown regions. Traditionally, [4, 7, 8, 25] utilize the sampled pixels color from foreground and background to estimate the alpha value in unknown region. Similarly, [27, 13, 1] determine the alpha matte by propagating it from foreground and background pixels. More recently, deep learning methods have been proposed to solve the matting problem end-to-end. [30, 10, 18, 6] concatenate input image together with its trimap and apply encoder-decoder networks to obtain matting result. [30] also introduce a large matting dataset called Adobe Image Matting (AIM). Despite their great success, these methods are still hard to be deployed in practice because of the high cost to get accurate trimaps for every image.

Background-based methods. Some other approaches try to replace trimap with a relatively inexpensive alternative. [22, 16] propose to determine the foreground through the combination of an image and its background. The background image is fed as a green screen to the network, so it can easily distinguish the foreground. This approach achieves good results but suffers low computational efficiency, and thus can work only on low resolution. [16] reduce its complexity to perform high-resolution and real-time matting. However, both methods have limits when the camera is shaking or the background changes.

Trimap-free methods. Since the trimap is hard to produce, efforts have been made to get rid of it. [21, 33] directly output the matting result from arbitrary images. But due to the lack of prior information, sometimes they would perform below expectation. [35, 26] is designed specifically for human portrait matting, so they can use the semantic information of human portrait. But they may fail in detail regions such as hair.

2.2. Video matting.

Although image-based matting has shown significant success and can be naively applied to videos frame by frame, there are several attempts to make advantage of the temporal correlation of video to improve the matting quality.

Video portrait/human matting. Some video-based methods are specifically designed for human matting. [12] first performs single frame matting, then reinforces the consistency of contiguous results by post-processing. However, it cannot handle cases when the human is moving too fast. [17] uses a recurrent neural network to acquire temporal information. With the help of temporal information, it

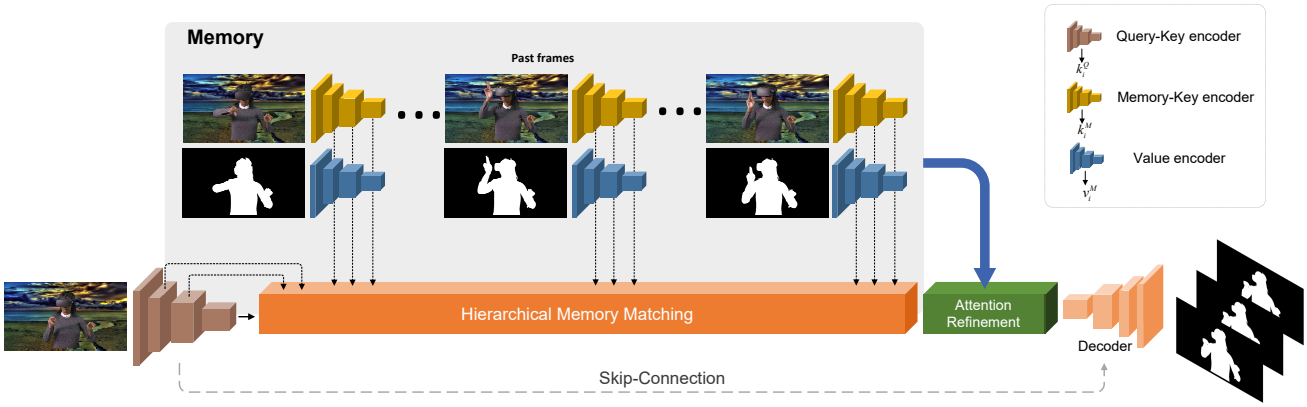


Figure 2. The architecture of our hierarchical space-time semantic guidance video object matting (HSTSG) model. The HSTSG first extracts the hierarchical key features before comparing them with the historical keys. Then a hierarchical memory matching is proposed to retrieve the value features from the memory bank. The cross-video-frame attention refinement network is followed to integrate the feature representations of the spatio-temporal neighborhoods from multiple adjacent frames.

achieves state-of-the-art results in human video matting.

Video object matting. Since the large and robust dataset is hard to obtain, deep learning methods for video object matting haven't been proposed until recent years. [34] release a large benchmark dataset together with its two-stage video matting algorithm. It first generates the trimap for each frame, then aggregates the temporal features using the attention mechanism to predict alpha matte frame by frame. [28] also propagates trimaps across different frames. For matting modules, [28] extracts different spatial and temporal features from multiple frames, which produces spatially and temporally coherent results. Additionally, it also proposed a video object matting dataset. Our work is also evaluated on these two datasets.

3. Architecture

The network of our hierarchical space-time semantic guidance video object matting (HSTSG) is designed to automatically predict the accurate alpha mattes and corresponding semantic masks given only the annotated semantic mask of the first frame as a reference. It can perform both semantic propagation and alpha prediction in the same task without redundant trimap generation. The architecture of the HSTSG network is shown in Figure 2, we first extract the hierarchical key features before comparing them with the historical keys. Then a hierarchical memory matching module is followed to query the value features from the memory bank. We also propose a cross-video-frame attention refinement network to integrate the feature representations of the spatio-temporal neighborhoods across multiple adjacent frames.

3.1. Hierarchical Key and Value Encoders

We design the hierarchical key encoder and value decoder based on STCN [3]. The hierarchical key encoder takes each query frame as input and extracts the hierarchical key features to generate hierarchical spatio-temporal correspondences between the query and memory frames. While trimap propagation struggles to learn the true distribution under certain scenarios ($e.g.$ self-occlusion, perspective change), the semantic mask however can provide robust binary estimation ($i.e.$ foreground or background) during the propagation process [3]. Inspired by this, we utilize the binary semantic mask as the semantic guidance and perform the mask prediction along with the alpha prediction. The hierarchical value encoder is designed to embed the previously predicted alpha mattes and semantic masks into the value features, with hierarchical semantic guidance.

Without loss of generality, We employ ResNet50[9] and ResNet18[9] as the backbones of the hierarchical key encoder and value encoder separately. The hierarchical key features (query key K_i^Q and memory key K_i^M) and value feature V_i^M are extracted from the $i-th$ Res block with the output scale of $\{1/4, 1/8, 1/16\}$ with respect to the query frame, where $i \in \{4, 3, 2\}$. Compared to previous memory-based approaches, we maintain a memory bank of hierarchical features to retrieve corresponding multi-scale value features that can enforce stronger spatio-temporal coherence at both global and local levels.

3.2. Hierarchical Memory Matching Module

Considering the robust temporal coherence both in the global and details, we design a hierarchical top$-k$ filtering memory matching module to exploit complementary semantic information at multiple feature levels.

Memory read operation with Top-k filtering.

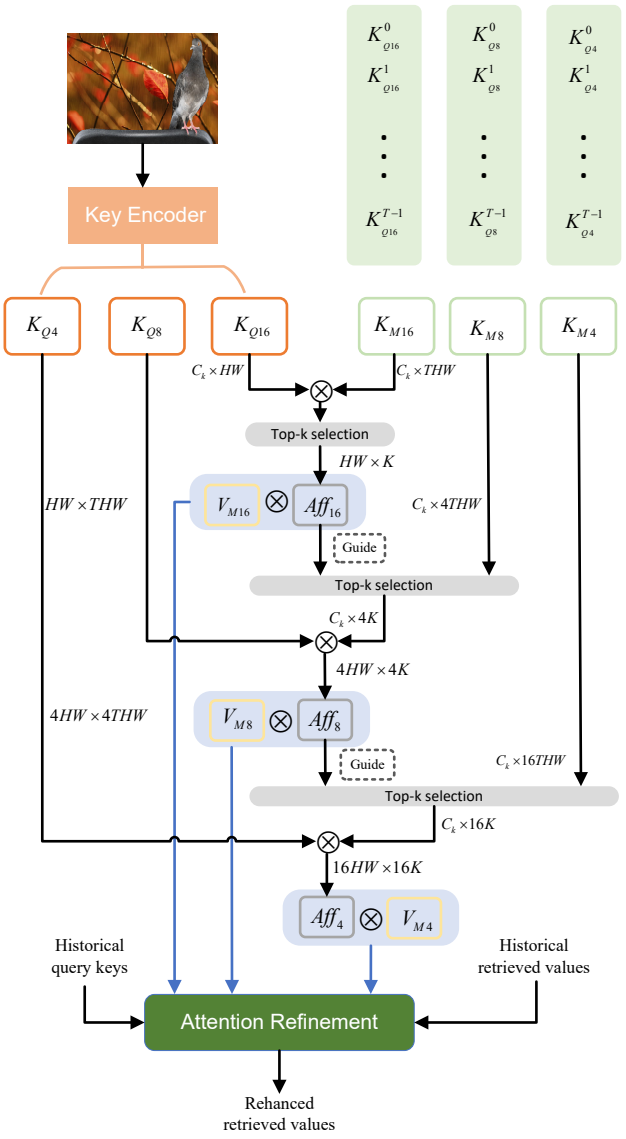


Figure 3. Implenmention of our our hierarchical memory matching module as described in Section 3.2.

As in recent memory read methods [3, 19, 2, 23, 15, 11], the affinity matching between each query and memory pixel is first computed by the memory matching module. However, the hierarchical pixel-to-pixel dense attention map generation comes with highly expensive computational cost.

To address this issue, we introduce top-k filtering in our hierarchical memory matching module and Figure 3 illustrates the detailed implementation of our memory read operation. Given a query frame and T memory frames, we first compute the key and value features of each memory frame. And then the key and value features of different memory frames are concatenated separately along the temporal dimension at each feature level to generate hierarchical memory key and value maps.

At the beginning of our memory read operation, we compute the affinity between each query pixel and memory pixel in s-th features, the pairwise affinity matrix Aff_s at the s level ($s \in \{1/4, 1/8, 1/16\}$) is computed by the dot product as follows:

$$Aff_{ij}^{s} = k_i^{Ms} \cdot k_j^{Qs} \tag{1}$$

where Aff_{ij}^{s} denotes the affinity score between the feature vectors - k_i and k_j at the i, j-th position. Then we define the top-k filtering guided softmax-normalized affinity matrix W^s at the s-th feature level as follows:

$$\mathrm{W}_{ij}^{s} = \frac{\exp(Aff_{ij}^{s})}{\sum_j \exp(Aff_{ij}^{s})}$$
$$\Downarrow$$
$$\mathrm{W}_{ij}^{s} = \begin{cases} \frac{\exp(Aff_{ij}^{s})}{\sum_{i \in \mathrm{Top}_j^k(Aff^s)} \exp(Aff_{ij}^{s})}, if\ i \in \mathrm{Top}_j^k(Aff^s) \\ 0,\ else \end{cases} \tag{2}$$

$$V^{Qs} = V^{Ms}\mathrm{W}^s \tag{3}$$

where $\mathrm{Top}_j^k(Aff^s)$ denotes the set of indices that are top-k in the j-th column of Aff^s at s-th level. The aggregated readout hierarchical feature V^{Qs} for the query frame can be computed as a weighted sum of the memory features with W^s.

We follow [24] to utilize the selected k best matching memory pixels in the coarser level (low-resolution, $e.g.$ $1/16$) attention maps to guide $4k$ pixels in the finer level (higher-resolution, $e.g.$ $1/8$)) attention maps for significantly lower computational costs. The retrieved value V_t^{Qs} for the current frame are concatenated with the historical adjacent query values to produce the cross-frame retrieved value $\bar{V}_t^{Qs}$ and then fed into the cross-frame attention refinement module for temporal smoothness improvement.

3.3. Cross-frame attention refinement.

To enhance the temporal smoothness, we propose a transformer-based [29] cross-frame attention refinement (CFAR) module that leverages the spatio-temporal neighborhoods from the multiple adjacent frames (both historical and current frames) to refine $\bar{V}_t^{Qs}$ based on the spatio-temporal correlation. We concatenate the current query key K_t^{Qs} with K_{t-1}^{Qs} and K_{t-2}^{Qs} along the temporal dimension to produce cross-frame query key $\bar{K}_t^{Qs}$. We first compute the spatio-temporal affinity among the pixels of adjacent query key maps at the s-th feature level:

$$Aff_{cross}^{s} = \mathrm{Attn}\left(\bar{K}^{Qs}, \bar{K}^{Qs}\right) \tag{4}$$

Then the cross-frame retrieved value $\bar{V}_t^{Qs}$ can be enhanced as follows:

$$V_{attn}^{s} = \bar{V}_t^{Qs} + softmax\left(Aff_{cross}^{s}\right) \odot L_1\left(\bar{V}_t^{Qs}\right) \tag{5}$$

Method	Trimap Setting	VMD				
		MSE	MAD	SSDA	dtSSD	MESSDdt
DIM [30]	full-trimap	9.99	44.38	61.85	34.55	2.82
IndexNet [18]	full-trimap	9.37	43.53	58.83	33.03	2.33
GCA [14]	full-trimap	8.20	40.85	55.82	31.64	2.15
TCVOM(GCA) [34]	full-trimap	**7.07**	**37.65**	**50.41**	**27.28**	**1.48**
TCVOM [34]	1-trimap	22.15	57.40	77.23	32.18	2.97
HSTSG(Ours)	1-mask	**12.48**	**37.97**	**56.09**	**28.03**	**1.86**

Table 1. Results of our HSTSG versus state-of-the-art methods on VideoMatting108 test set with the medium trimap setting. "full-trimap" means frame-by-frame user-annotated trimaps. TCVOM(GCA) means TCVOM[34] utilizes GCA [14] as backbone.

Method	Trimap/Mask Setting	DVM					
		MSE(10^{-3})	MAD	Grad	Conn	dtSSD	MESSDdt
DIM[30]	full-trimap	30	54.55	35.38	55.16	23.48	0.53
IndexNet[18]	full-trimap	28	53.68	27.52	54. 44	19.5	0.49
Context-Aware[10]	full-trimap	27	51.78	28.57	49.46	19.37	0.5
GCA[14]	full-trimap	22	47.49	26.37	45.23	18.36	0.33
DVM[28]	full-trimap	14	40.91	19.02	40.58	15.11	**0.25**
MG Mating[32]	full-mask	19	43.28	25.14	43.96	19.41	0.42
DVM[28]	20-trimap	16	43.66	26.39	42.23	16.34	0.28
HSTSG(Ours)	1-mask	**5**	**26.74**	**11.11**	**22.68**	**13.92**	0.26

Table 2. Results of our HSTSG versus state-of-the-art methods on DVM test set. "full-trimap" and "20-trimap" means user-annotated trimaps are provided frame-by-frame and every 20 frames respectively.

where $\odot$ indicates an element-wise multiplication, $L_1(\cdot)$ is the L1 normalization which normalizes along the memory dimension. Finally, we adopt the standard FFN to strengthen the feature representation ability of the attention value:

$$V_{cross}^{Qs} = FFN\left(V_{attn}^{s}\right) + V_{attn}^{s} \tag{6}$$

These processes impose the locality constraint on global frames by applying an attention mechanism. The enhanced retrieved value V_{cross}^{Qs} are then fed into the decoder through shortcut connections at the corresponding scale to predict the refined semantic masks and alpha mattes of current and adjacent query frames. The corresponding results are also updated in the memory bank.

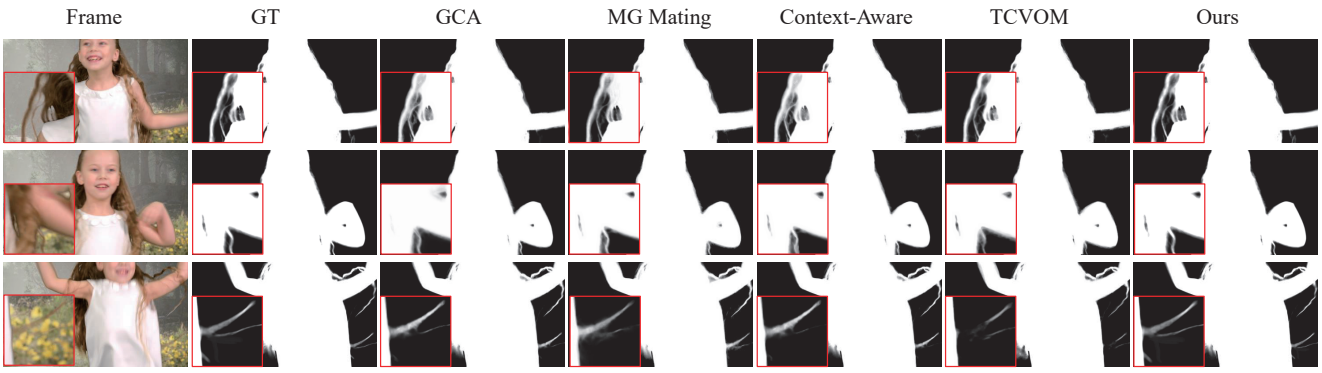
4. Experiments

In this section, we first describe the datasets used for training and testing. Subsequently, we compare our results with existing state-of-the-art (SOTA) foreground matting algorithms. Finally, we analyze the effectiveness of each component in our hierarchical space-time semantic guidance video object matting (HSTSG).

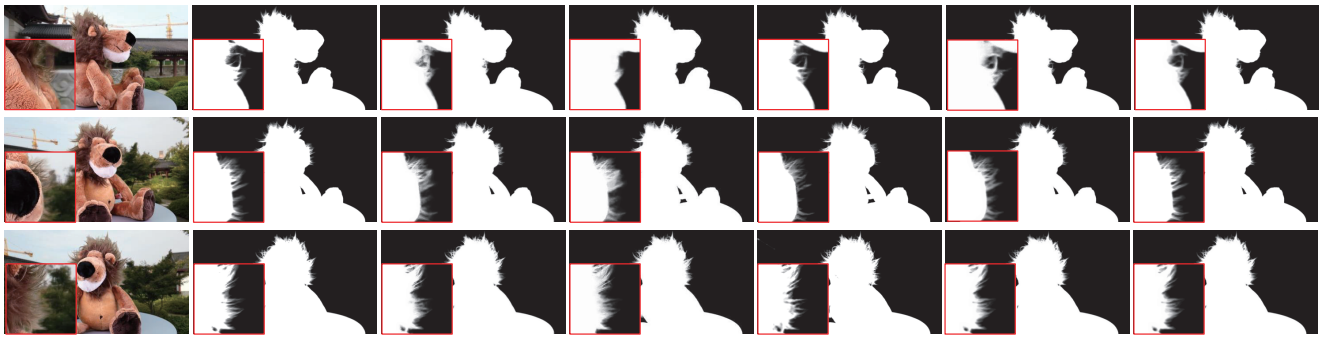
4.1. Datasets and experimental settings.

VideoMatting108. The VideoMatting108 dataset [34] consists of 108 video clips with 1080p resolution. The dataset relies on green-screen video footage to extract ground-truth alpha mattes, where 68 high-quality (1080p and 4K) ones from the Internet and 40 green-screen ones by self-collection to supplement the object types such as fur, hair, and semi-transparent. The dataset is split into 80 clips in the training set and 28 clips in the validation set. The ground truth trimaps of VideoMatting108 are generated and dilated on the fly with the random-sized kernel (from 1×1 to 51×51) during training. **DVM.** DVM [28] collects over 6500 various real-life videos of natural scenarios as background. Also, it includes green screen video clips from the Internet to serve as targets for foreground color and alpha matte generation. The training set contains 6400 videos by compositing foregrounds from 325 images plus 75 green screen videos, with 16 natural background videos. The test set contains 248 videos by compositing foreground from 50 images plus 12 green screen videos, with 4 natural background videos.

Evaluation metrics. To numerically evaluate the algorithm, we employ SSDA(average sum of squared difference), MESSDdt (mean squared difference between the warped temporal gradient), and dtSSD(mean squared difference of direct temporal gradients) as the temporal coherency metrics [5]. In addition, we adopt MAD (mean absolute difference), MSE (mean squared error), Grad (gradient error), and Conn (connectivity error) to evaluate the



(a) Dancing girl



(b) Lion

Figure 4. Comparison of alpha predictions with state-of-the-art methods on VideoMatting108 dataset. GCA and CAM take frame-by-frame trimaps as inputs, and MG takes frame-by-frame masks as inputs.

Pretrained	Module		VMD				
	HMMM	CFAR	MSE	MAD	SSDA	dtSSD	MESSDdt
			32.14	72.48	72.36	43.96	3.34
✓			28.43	64.88	70.15	36.49	2.91
✓	✓		18.20	53.15	68.11	35.46	2.82
✓		✓	19.41	52.65	64.28	33.50	2.43
	✓	✓	15.96	30.17	64.67	33.31	2.46
✓	✓	✓	12.48	37.97	56.09	28.03	1.86

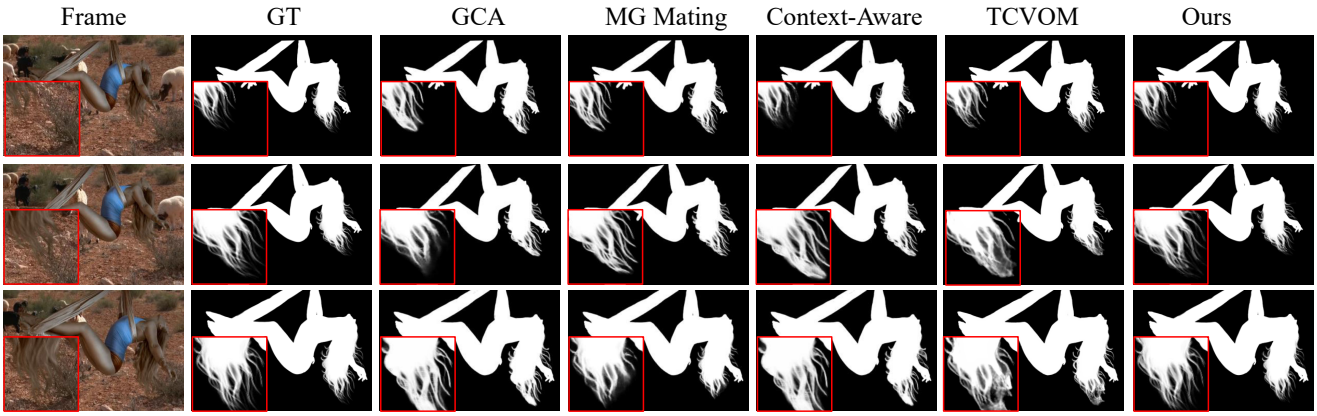
Table 3. Ablation on modules and training stages. "Pretrained" means model is initilized with the pretrained weights on the DIM dataset. HMMM and CFAR represents our hierarchical memory matching module and cross-frame attention refinement.

per-pixel accuracy [30].

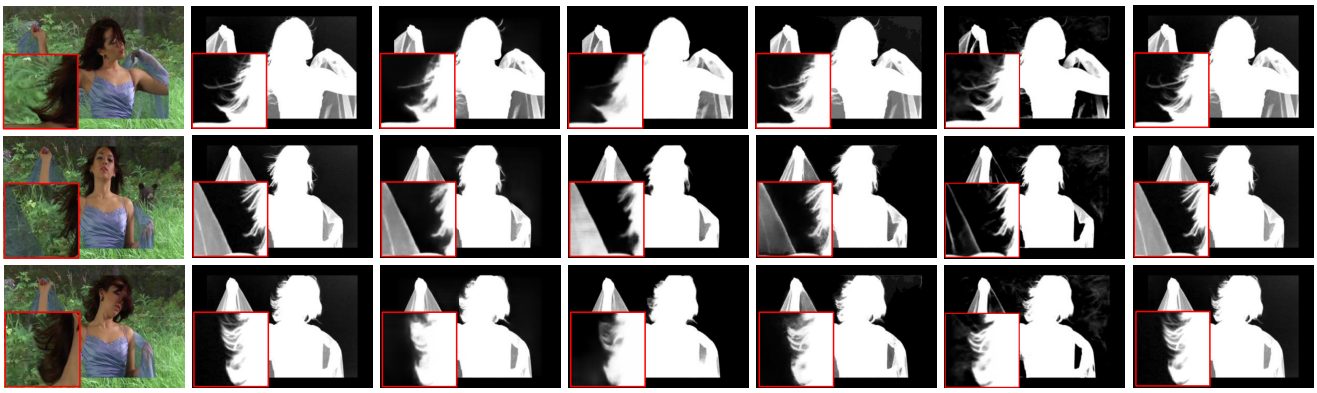
Implementation details. The training of our video matting network consists of two stages. Before Stage 1, we initialize our baseline model with the weights of STCN [3] trained on the segmentation datasets [20, 31]. In the first stage, we pretrain our model on the DIM dataset [30]. Then we proceed to train our model on the video matting datasets [34, 28] in Stage 2. We re-scale all video frames to 512×512 pixel patches in training. For inference, we utilize full-resolution as inputs. To properly manage the duration of training samples and memory bank, we mainly follow the implementation details of STCN [3]. More implementation details are provided in the supplementary material.

4.2. Comparison with State-of-the-Art Methods.

To evaluate the performance of our video matting method, we compare our HSTSG with the state-of-the-art trimap-based or mask-guided image matting methods: DIM[30], IndexNet[18], Context-Aware[10], GCA[14], and MG Mating[32], all of which need user-annotated trimaps or masks frame-by-frame; trimap-based video object matting: DVM[28] and TCVOM [34] that need trimap propagation after given one or several user-annotated trimaps. We follow the trimap propagation strategy in TCVOM [34] to generate trimaps when such annotation is provided. We report MSE, MAD, SSDA, dtSSD, and MESSDdt, and Grad and Conn between predicted and



(a) Yoga



(b) A woman with a scarf

Figure 5. Visual comparisons with the state-of-the-art methods on the DVM [28] dataset.

Feature Stride(m)			VMD				
16	8	4	MSE	MAD	SSDA	dtSSD	MESSDdt
✓			78.23	114.33	110.68	46.15	3.60
	✓		52.14	89.37	86.54	37.17	2.88
		✓	24.82	67.41	79.78	32.00	2.58
✓	✓		39.43	74.03	79.83	30.59	2.46
✓		✓	19.69	58.56	70.20	29.12	2.18
	✓	✓	17.35	41.10	64.17	28.96	1.91
✓	✓	✓	12.48	37.97	56.09	28.03	1.86

Table 4. Performance comparison of Hierarchical Memory with different scales.

ground truth alpha mattes. To fairly compare, we fine-tune these image-based methods on the VideoMatting108 [34] and DVM [28] benchmarks. For trimap-based methods, we measure errors only on the unknown regions, while we measure the global errors for our trimap-free HSTSG.

VideoMatting108. Table 1 shows the quantitative results of our HSTSG and other SOTA models on the VideoMatting108 dataset with medium trimap setting for trimap-based methods. We observe that our HSTSG shows superiority over DIM [30] and IndexNet [18] on $4/5$ metrics, and over GCA on $3/5$ metrics. Our HSTSG only under-performs by a small margin compared to TCVOM(GCA) [34] with a full-trimap setting that requires expensive manual annotation. While given only one user-annotated trimap or mask, our HSTSG outperforms the state-of-the-art VOM method (TCVOM(GCA) [34]) by a large margin, which demonstrates that our hierarchical space-time semantic guidance mechanism can effectively maintain the completeness and temporal coherence of the semantic propagation. Some visualizations on VideoMatting108 [34] are provided in Figure 4.

DVM. Table 2 shows the quantitative results of our

Memory management	VMD				
	MSE	MAD	SSDA	dtSSD	MESSDdt
Every 5 frames	57.19	77.31	84.35	34.31	2.97
Every 3 frames	56.54	78.16	82.63	33.2	2.62
First + Every 3 frames	38.62	52.46	68.17	28.64	2.18
prev 2 frames	25.38	42.13	65.64	28.72	1.95
First+ prev 2 frames	12.48	37.97	56.09	28.03	1.86

Table 5. Comparison of different memory management strategies.

Top-k	VMD				
k at $1/16$ level	MSE	MAD	SSDA	dtSSD	MESSDdt
16	24.34	41.78	62.54	29.51	2.11
8	12.48	37.97	56.09	28.03	1.86
4	31.05	45.62	56.76	29.82	2.08

Table 6. Performance comparison of different k in top-k filtering of memory bank at the $1/16$ feature level.

HSTSG and other SOTA models on the DVM dataset with a medium trimap setting for trimap-based methods. Our method shows significant superiority over all competing methods, on both trimap-based or mask-based cases including full-trimap or full-mask settings and VOM ones with 1-trimap setting. The quantitative results also demonstrate that our HSTSG can achieve more accurate alpha prediction and stronger temporal coherence, thanks to the hierarchical space-time semantic guidance and cross-frame attention refinement. We provide some comparisons in Figure 5 to illustrate the smoothness improvement of our HSTSG compared to other VOM methods.

4.3. Ablation study.

Module ablation. We conduct an ablation study on our proposed hierarchical memory matching module (HMMM) and analyze the significance of different components. As shown in Table 3, we observe that pretraining on the image matting dataset can reach faster convergence than the random initialization, because pretraining can accelerate the learning process to quickly find more meaningful semantic features. When applying our designed hierarchical memory matching module on the STCN baseline, the matting performance is significantly improved. Our HMMM+baseline outperforms the baseline by 36% on MSE and 18.1% on MAD, which demonstrates that our proposed hierarchical memory matching mechanism can contribute to establishing stronger temporal coherence. After adding the CFAR module, the performance is improved further and our model surpasses the baseline by 56.1% and 41.5% on MSE and MAD. The CFAR module can benefit the feature representations across multiple adjacent frames, meanwhile, it can also enhance the temporal smoothness of the predicted frames.

Different settings of HMMM. We investigate the effectiveness of the hierarchical memory matching nodule under two factors: 1) the selection of feature scales, 2) memory management, and the top-k guidance setting. We set the memory read at certain scales and select different hierarchical combinations as ablation. The experiments are conducted on the VideoMatting108 validation set. As shown in Table 4, for the single-scale memory read settings, the finer scale achieves better performance, where the best setting is stride 4. Note that the performance is further improved after applying the hierarchical memory reading, which demonstrates that the hierarchical feature representations are beneficial to combine both global semantic information on a coarse scale and object details on a fine scale. Table 5 shows the performance difference under different memory management settings in the reference stage, we observe that the setting of taking the first user-annotated frame and its most recent two previous frames in the memory bank can achieve the best performance. Table 6 shows the ablation results of different top-k settings. We observe that the dense memory reads may not always bring performance gain, likely because denser memory matching may introduce an unnecessary semantic noise. Conclusively, adjusting the appropriate k value ($e.g.$, setting $k = 8$ at the $1/16$ scale) can effectively improve the model performance.

5. Conclusion

In this paper, we adapt the hierarchical memory matching mechanism into the space-time baseline to build an efficient and robust framework for semantic guidance propagation and alpha prediction. To enhance the temporal smoothness, we also propose a cross-frame attention refinement (CFAR) module that can refine the feature representations across multiple adjacent frames (both historical and current frames) based on the spatio-temporal correlation among the cross-frame pixels. Extensive experiments demonstrate the effectiveness of hierarchical spatio-temporal semantic guidance and the cross-video-frame attention refinement module, and our model outperforms the state-of-the-art VOM methods. For future work, it may be possible to extend the video object matting method to video 3D object reconstruction.

References

[1] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[2] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021.

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021.

[4] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[5] Mikhail Erofeev, Yury Gitman, Dmitriy S Vatolin, Alexey Fedorov, and Jue Wang. Perceptually motivated benchmark for video matting. In *BMVC*, pages 99–1, 2015.

[6] Marco Forte and François Pitié. f, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020.

[7] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010.

[8] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. IEEE, 2011.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4130–4139, 2019.

[11] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021.

[12] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1140–1147, 2022.

[13] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007.

[14] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11450–11457, 2020.

[15] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020.

[16] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021.

[17] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022.

[18] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019.

[19] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[20] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[21] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020.

[22] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020.

[23] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *European Conference on Computer Vision*, pages 629–645. Springer, 2020.

[24] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021.

[25] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013.

[26] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European conference on computer vision*, pages 92–107. Springer, 2016.

[27] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM SIGGRAPH 2004 Papers*, pages 315–321. 2004.

[28] 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.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017.

[31] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021.

[33] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019.

[34] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5128–5137, 2021.

[35] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 297–305, 2017.