GRIB: Combining Global Reception and Inductive Bias For Human Segmentation and Matting

Yezhi Shen*, Weichen Xu*, Qian Lin†, Jan P. Allebach*, and Fengqing Zhu*
* Purdue University, West Lafayette, IN 47906, USA
† HP Inc., Palo Alto, CA 94306, USA
{shen397, xu1363, allebach, zhu0}@purdue.edu
qian.lin@hp.com

Abstract

Human video segmentation and matting are challenging computer vision tasks, with many applications such as background replacement or background editing. Numerous methods have been proposed for human segmentation and matting in either portrait or first-person view videos. In this paper, we propose a real-time network that performs first-person view hand and manipulated object segmentation as well as second-person view human video matting. We introduce a global reception inductive bias block in the network's encoder that aggregates the pixel features at short, medium, and long ranges. Furthermore, we propose a multi-target optimization method that fully leverages segmentation and matting labels to accelerate training. Our model outperforms existing real-time methods by achieving 93.9% mIoU on HP-Portrait, 95.1% mIoU on VideoMatte as well as 72.7% mIoU on EgoHOS datasets and achieves faster runtime.

1. Introduction

Human segmentation and matting techniques are crucial in applications involving video background blurring and replacement, facilitating the separation of individuals and their surroundings. Portrait image segmentation approaches such as SINet [14] and PortraitNet [31] focus on images with human heads and shoulders exclusively by designing efficient encoders and decoders. While excelling in accuracy for simple portrait images, these methods experience a notable decline in accuracy when confronted with complex images featuring humans with extended limbs and intricate backgrounds. PHOS [20] maximizes the global reception field offered by the vision transformer architecture and introduces a specialized encoder designed for the segmentation of humans and interacting objects, yet only runs real-time on low-resolution images.

Real-time human matting models such as MODNet [8] and RobustVideoMatting (RVM) [11] facilitate general image encoders as their backbones with self-designed decoders. MODNet ensures the alignment between the matting mask and the segmentation mask with the imposition of a consistency loss between its low-resolution segmentation mask and the downsampled matte, but faces limitations in handling videos with pronounced motion blurs due to the absence of temporal information. RVM enhances temporal coherence by integrating gated recurrent unit (GRU) into its decoder modules and introduces an innovative training method that involves utilizing foreground human images combined with indoor background videos.

Existing segmentation and matting methods have predominantly shown promising results on second-person view videos and images, where the entire human contour is captured [26]. This aligns with conventional video composition, where subjects are usually framed within the entirety of the scene, which led to a significant oversight that the first-person perspective has not been adequately addressed. First-person view content, often characterized by partial views of the body, such as hands or objects being manipulated, traditional methods struggle due to the lack of full-contour visibility and the complex interaction of foreground elements with the camera. This has resulted in subpar accuracy levels in human segmentation and matting tasks from a first-person viewpoint, posing a challenge for applications requiring a high degree of immersion and interaction, such as augmented reality (AR) and virtual reality (VR).

In this paper, we aim to bridge the gap by enhancing the segmentation and matting capabilities from the first-person perspective without compromising the quality in second-person view videos. Specifically, we propose an end-to-end model that is capable of performing first and second-person view human and interacting object mask prediction in real-time. To achieve complete segmentation with fine details, we design an efficient encoder block with a wide receptive

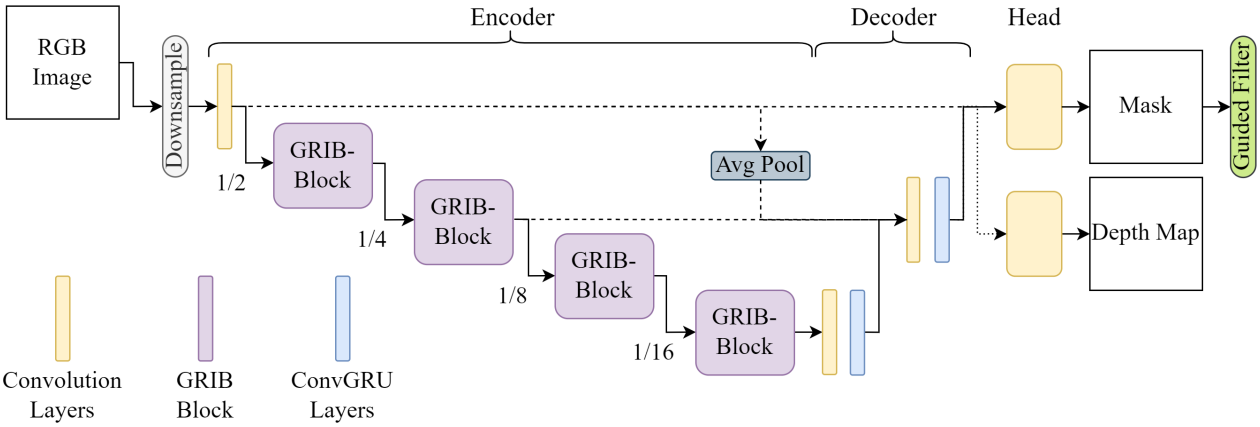


Figure 1. Overview of proposed network architecture, which consists of a hybrid feature extraction encoder, a recurrent decoder, and two output layers for segmentation and depth estimation. Normalization and activation layers are omitted.

field and strong inductive bias using CNN-ViT hybrid architecture. Our model combines the segmentation and matting branches to better leverage the knowledge from all datasets, and distill background knowledge from a pretrained depth estimation model to accelerate the convergence during pretraining. By jointly training our model end-to-end using segmentation, matting, and depth prediction datasets, the model converges faster and achieves better accuracy. Our method can be applied to mixed reality applications to perform passthrough function for humans in the surroundings. To summarize, our contributions are as follows:

- We design a lightweight global reception inductive bias block for the encoder, which leverages features from short, medium, and long ranges.
- We propose to combine the segmentation and matting branches for more efficient training.
- We develop a novel network architecture that efficiently decodes extracted features for foreground mask prediction on both first-person and second-person perspective videos.

2. Related Works

Portrait Segmentation Portrait segmentation methods [14, 29, 31] favors simple input images where human occupies over 50% of the entire image. These methods feature meticulously crafted lightweight encoders aimed at drastically reducing parameters while maintaining the desired accuracy in image portrait segmentation. However, they tend to yield subpar results when applied to images with intricate backgrounds or when dealing with elongated human limbs. PHOS [20], a recent segmentation approach designed for video segmentation of human upper body, adopts a larger encoder based on vision transformers along with a recurrent decoder.

Human Matting Methods for human matting [8, 10, 11, 13, 17] produce nuanced labels for human foregrounds to enhance video background replacement and editing. These techniques typically employ standard image encoders as their backbones, which are comparatively sluggish and less effective than purpose-built encoders. Notably, recent advancements such as MODNet [8] and RVM [11] offer end-to-end training for human matting tailored to video applications, eliminating the need for supplementary inputs. However, it is important to note that MODNet and RVM are not capable of processing first-person-view videos.

Egocentric Hand Segmentation Hand segmentation methods from egocentric perspectives [5, 9, 30] extract hand regions from first-person view videos, serving multiple purposes such as human activity categorization and hand gesture recognition. Among them, fine-grained EgoHOS [30] stands out as the sole method capable of accurately segmenting both hands and manipulated objects. However, its reliance on a multi-stage segmentation process renders it unsuitable for real-time applications.

3. Method

We propose a novel network that is capable of performing human segmentation and matting mask prediction for first-person and second-person view videos. GRIB, a novel feature extractor block is designed to efficiently explore spatial information at all distances to achieve complete and accurate mask prediction. As shown in Figure 1, our model has a shared encoder-decoder with two output heads to map the output features to a foreground mask and a depth map. We pair our model with a guided filter [22] to perform inference on images with resolutions higher than 256×256.

3.1. Encoder Architecture

Global Reception Inductive Bias (GRIB) Block: The relationship between the person's belongings and the hu-

man is important information that needs to be extracted to ensure correct segmentation. Our feature extraction encoder is designed to achieve a global receptive field, while still maintaining profound capability of local feature extraction.

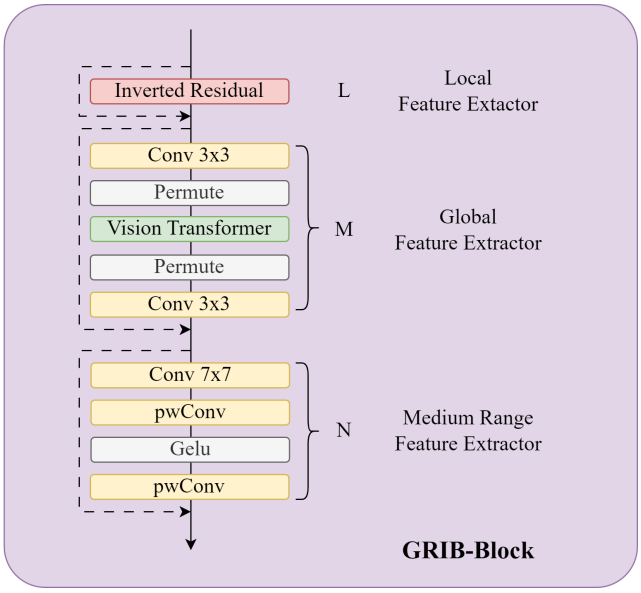


Figure 2. Architecture of our Global Reception Inductive Bias (GRIB) block, consisting of local feature extractors, global feature extractors, and midium range feature extractors.

We design the Global Reception Inductive Bias (GRIB) block, shown in Figure 2, which aims to provide our model with balanced global reception and inductive bias [1, 27] while maintaining reasonable scalability. GRIB enjoys a CNN-ViT hybrid design with sequentially connected Transformer layers to obtain a global receptive field, large kernel convolution layers to exploit mid-range inductive bias, and Inverted Residual blocks[7] to efficiently extract details from local features.

The GRIB block leverages local features to predict the mask for the finest details including hair, fingers, and small objects. A 3×3 convolution layer in the Inverted Residual block performs feature extraction in the local regions. Inverted Residual block features the squeeze and excite (SE) design, which captures channel-wise dependencies in feature maps. By recalibrating the contribution of feature map channels through the SE layer, Inverted Residual blocks become more effective at learning discriminative features within the data. As a result, the manipulated objects are better separated from similar objects in the background.

The global feature extractor starts with a 3×3 convolution layer to distribute the extracted local information. The following M Mobile Vision Transformer (MViT) [12] layers are responsible for extracting global features by performing patch-wise attention. The global representation transformer module is capable of encoding long-range information across the entire input, which helps to associate the manipulated objects with the human. The vision transformer with global reception provides us the privilege of only using 4-layer encoders without the need to apply additional bottleneck blocks after the last layer of the feature encoder. Even though MViT is a lightweight neural network designed for mobile applications, as a global attention Transformer, the MViT layer still suffers from limited scalability of depth M due to the time complexity of H^2W^2, where H and W denotes the height and width of inputs respectively.

The GRIB block uses the design of N convolution groups consisting of a large kernel convolution, two pointwise convolutions (pwConv), and an activation function to extract medium-range features. The 7×7 convolution provides the model with a reception field of 10% on feature maps of resolution 64×64 and 20% on resolution 32×32. In human segmentation and matting tasks, medium-range reception corresponds to an understanding of individual body parts including head, shoulders, arms, and legs. Having medium-range reception provides stable segmentation results when limbs not connected to the torso appear in video frames.

Encoder: Each layer of the encoder is constructed using a GRIB block after the initial 3×3 convolution layer used for down-samplings. Inspired by lightweight mobile encoders, we create two variants of our network: Ours-Small (Ours-s), and Ours-Extra Small (Ours-xs) for different resource use cases with differently configured encoding channel sizes.

3.2. Decoder and Output Heads

Decoder: Our decoder features a two-layer design with skip connections from the encoder and input, which reduces the number of parameters yet is still able to maintain precision during decoding. Our decoder block design is inspired by RVM [11] and Xu et al. [25], where the authors propose using ConvGRU, which is a gated recurrent unit (GRU) paired with 2D convolution layers. Each decoder stage is constructed using a combination of two layers of convolutional layers with ReLu activation function, a ConvGRU layer, and a bilinear upsampling layer. Our decoder retains the advantage of pure CNN decoders including the ability of spatial information decoding and efficient feature fusing, yet captures the temporal information from video sequences. To compensate for the loss of details due to skipping the $\frac{1}{2}, \frac{1}{8}$ layer, our $\frac{1}{4}$ layer fuses feature map from the first layer downsampled using average pooling in addition to the skip connection from the encoder side.

Output Head: Our network architecture comprises two distinct output heads: one is dedicated to generating mask predictions and the other focuses on producing depth predictions. Both output heads utilize the fused feature map

from the original image and the decoder output as their input. Sequential combinations of convolution, normalization, and activation layers are employed within the output heads to project the decoded features into the final output predictions.

4. Experiment

4.1. Datasets

Six image portrait segmentation datasets, one video human foreground dataset, one video segmentation dataset, and one hand-object video segmentation dataset are used in our experiments.

Image datasets: The image datasets BaiduV1 [23], BaiduV2 [23], EG1800 [18], Supervisely [3], and HP-Multi-Person [24] are cropped to contain upper body only. Depth maps of the above portrait image segmentation datasets are generated using MiDaS V3.0 [15] DPT-L [16] model. These datasets are used during the pertaining stage to accelerate the convergence of our network and avoid overfitting. We use the HP-Portrait[19] image dataset, which contains images of the upper body of humans with manipulated objects, for both training and testing. We divide the HP-Portrait image segmentation datasets in an 8:1:1 ratio for training, validation, and testing as described in PHOS[20].

Video dataset: Since there are no publicly available portrait video segmentation datasets, we follow PHOS [20] to use VideoMatte240K (VideoMatte) [11] as our foreground video dataset. The HP-Portrait is also used as an image segmentation foreground dataset. We prepare a background dataset consisting of 17,000 self-collected background images and 3,000 background videos. The frames used for training, validation, and testing are generated at runtime by compositing the selected foreground image or video sequences with the selected background image or video sequences. The 484 clips in VideoMatte are divided following the original paper [11] into 474:4:5 for training, validation, and testing, and all frames are used. When VideoMatte labels are used for segmentation evaluation, a sigmoid function is applied to all masks.

The combination of all clips in YoutubeVis [28] containing humans is used as a human video segmentation dataset during the pre-training stage. In addition, we filter the EgoHOS [30] hand and object video segmentation dataset so that all arms and interacting objects share the same foreground label. The training, validation, and testing subsets are kept as provided with an approximate 8:1:1 ratio.

4.2. Experiment Setup

Training loss: We impose different loss functions during the model training with different selections of datasets. We train our mask prediction branch with binary cross-entropy (BCE) loss defined in Equation 1 on the human segmentation datasets, where y is the prediction and y' is the label.

$$y_s = sigmoid(y) \tag{1}$$

$$\mathcal{L}_{bce} = y'(-log(y_s)) + (1 - y')(-log(1 - y_s)) \tag{2}$$

During training with the VideoMatte dataset, we generate our pseudo-segmentation label by applying the sigmoid function to the matting label and apply the BCE loss. To fully leverage the temporal information and matting label introduced by the VideoMatte datasets, we apply the additional L1 loss and the laplacian pyramid loss reported by [4, 6] and a coherent loss from [21], where y is the prediction and y' is the label.

$$\mathcal{L}_{l1} = |y - y'| \tag{3}$$

$$\mathcal{L}_{lap} = \sum_{s=1}^{5} \frac{2^{s-1}}{5} |L^s_{pyr} y - L^s_{pyr} y'| \tag{4}$$

$$\mathcal{L}_{coh} = |\frac{dy}{dt} - \frac{dy'}{dt}|^2 \tag{5}$$

We apply the L1 loss to the depth estimation branch during pre-training.

Table 1. Model configuration

Model	Encoder-Ch	Decoder-Ch
Ours-xs	[16, 24, 48, 64]	[64, 32]
Ours-s	[32, 48, 64, 80]	[80, 32]

Training strategy:

Our model training is pipelined into three stages. They are designed so that our network progressively learns from image segmentation tasks jointly with video matting tasks to accelerate training. Our models are trained on two NVIDIA RTX 3090 GPUs parallel using Adam optimizer with the encoder and decoder channels listed in Table 1. We configure our models with L = [2, 1, 1, 1], M = [0, 1, 2, 0], and N = [0, 1, 2, 2] as in Figure 2 for each encoder layer.

Stage 0: Our model uses this pre-training stage since it leverages a self-designed feature extraction encoder, which lacks pre-trained weights. The model is trained for 10 epochs on the resolution of 256×256. All the image datasets are used for training the mask prediction branch and depth estimation branch with a learning rate of $1e^{-4}$. The video human segmentation dataset YoutubeVis is only used for training the mask prediction branch.

Stage 1: The model is trained for 15 epochs on the resolution of 256×256. The HP-portrait image segmentation dataset, EgoHOS hand object segmentation video dataset, and VideoMatte video matting dataset are used for training

Table 2. Segmentation and Matting Results evaluated on HP-Portrait, VideoMatte and EgoHOS datasets

Method/Datasets	HP-Portrait (mIoU) ↑	VideoMatte-Seg (mIoU) ↑	VideoMatte (MAD) ↓	EgoHOS (mIoU) ↑
MODNet [8]	88.16%	92.20%	9.12	48.34%
RVM [11]	89.36%	94.56%	6.04	53.40%
PHOS [20]	91.85%	93.71%	/	58.26%
Ours-xs	92.37%	94.08%	7.78	66.83%
Ours-s	**93.89%**	**95.14%**	**5.51**	**72.65%**

the mask prediction branch, and videos with a length of 15 are sampled. Training stage 1 uses a fixed learning rate of $5e^{-5}$.

Stage 2: The model is trained for another 10 epochs on the resolution of 1920×1080. The datasets used are consistent with the previous training stage, but videos with a length of 35 are sampled. Training stage 2 uses a fixed learning rate of $2e^{-5}$.

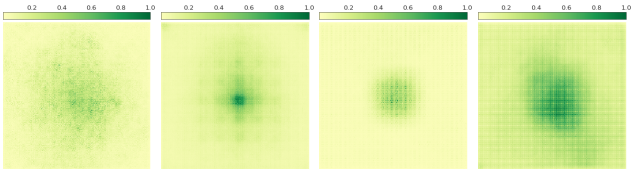
4.3. Experimental Results and Discussion

In our evaluation, we measure the performance of our approach against contemporary real-time portrait video segmentation and matting techniques such as MODNet [8], RVM [11], and PHOS[20]. We assess the networks' segmentation and matting accuracy, temporal consistency, and visual quality. To ensure a fair comparison, we retrain all the compared methods using the datasets outlined in section 4.1 with their provided training methods. Notably, for MODNet and RVM, which leverage pre-trained backbones on ImageNet-1K, pre-training stage is excluded during their training process.

Segmentation and Matting Accuracy: We conduct a thorough assessment of segmentation accuracy by comparing it to existing methods at a resolution of 256×256 on the HP-Portrait and VideoMatte datasets, as detailed in Section 4.1. The evaluation is based on mean intersection over union (mIoU), defined in Equation 6, where y represents the predicted mask and y' is the label. Additionally, we scrutinize the accuracy of first-person view hand object segmentation, comparing it to all methods at a resolution of 256×256 on the EgoHOS dataset, measured in terms of mIoU.

$$IoU = \frac{y \cap y'}{y \cup y'} \tag{6}$$

To gauge the segmentation accuracy of matting methods on the VideoMatte-Seg dataset, we apply a sigmoid function to all matting predictions and labels. Furthermore, we assess the matting accuracy at full HD resolution on the VideoMatte dataset, comparing it to the matting output of MODNet and RVM, using Mean Absolute Difference (MAD) as the metric.



(a) MODNet (b) RVM (c) PHOS (d) Ours

Figure 3. The Effective Receptive Field (ERF) of MODNet, RVM, PHOS and Ours respectively. A more widely distributed area indicates a larger ERF, while a darker area indicates more attention. The ERF of our method shows a good combination of global receptive field, uniform mid-range reception and strong inductive bias.

Results are reported in Table 2, as well as in Figure 5 and 6 , demonstrating that our method excels in both segmentation and matting tasks. Our model outperforms others quantitatively across all comparisons, showcasing commendable scalability. Figure 3 plots the effective reception field (ERF) using the method in [2], showing that our method achieves a good combination of the global receptive field, uniform medium-range reception, and strong inductive bias. The experiment results and plotted ERF further illustrate the direct correlation between limb segmentation accuracy (EgoHOS dataset) and medium-range feature extraction capability, while matting accuracy is directly linked to the model's proficiency in local feature extraction.

Temporal consistency: The temporal consistency of all the compared methods is assessed using the proposed technique [20] by PHOS on ten videos, each consisting of 900 frames. The evaluation focused on interframe (IF) mean Intersection over Union (mIoU) and interframe pixel accuracy. As outlined in Table 4, our method, labeled as "Ours-s," attains the highest IF mIoU and the second-highest IF pixel accuracy. The performance differentials between RVM, PHOS, and our approach are minimal, primarily attributed to the presence of limited motion in our daily scenario test videos and the advantageous integration of the Gated Recurrent Unit (GRU) in the decoder for all methods. Visual examples in Figure 7 show that our method produces consistent and accurate mask predictions for human videos

Model Size and Runtime: Table 5 presents a compre-

Table 3. Segmentation Results evaluated in mIoU on different datasets

Method/Datasets	HP-Portrait (mIoU) ↑	VideoMatte-Seg (mIoU) ↑	EgoHOS (mIoU) ↑
(a) **Ours-s**	**93.89**%	**95.14**%	**72.65**%
(b) w/o global feature extractor	86.96%	92.08%	54.23%
(c) w/o medium range feature extractor	88.27%	89.75%	61.36%

Table 4. Temporal Consistency Evaluation

Method	IF mIoU ↑	IF Pixel Accuracy ↑
MODNet [8]	93.79%	97.21%
RVM [11]	96.37%	98.55%
PHOS [20]	96.55%	98.48%
Ours-xs	96.40%	98.34%
Ours-s	**96.69**%	**98.61**%

hensive evaluation of our proposed approach alongside existing methods, focusing on parameters, floating-point operations (FLOPs), and frames per second (fps) at FHD resolution with a down-sample ratio of 0.25. The speed assessment is conducted using an NVIDIA T1000 GPU with 2.5 TFLOPs computation power, known for its performance comparable to the Qualcomm Snapdragon 8 Gen1 mobile SoC. The results reveal that our method boasts the fewest parameters and demonstrates the capability to achieve real-time performance, even on mobile devices.

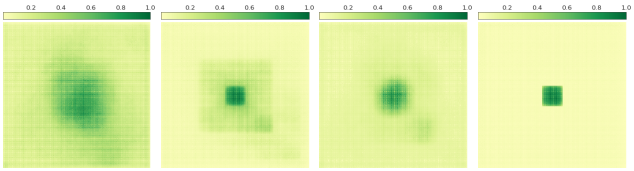
Table 5. Model Size and Runtime Evaluation

Model	Parameters ↓	FLOPs ↓	FPS ↑
MODNet [8]	6.49M	9.69G	23.07
RVM [11]	3.75M	**3.08G**	**35.81**
PHOS [20]	1.23M	6.64G	18.39
Ours-xs	**0.65M**	4.45G	30.52
Ours-s	1.27M	6.57G	25.65

5. Ablation Study

We evaluate the effectiveness of our designed encoder and decoder through ablation studies on HP-Portrait, VideoMatte, and EgoHOS datasets. We create three sub-variants of Ours-s model by (a) removing the global feature extractors in the GRIB blocks, (b) removing the medium range feature extractors in the GRIB blocks, and (c) removing both the global and medium range feature extractors in the GRIB blocks.

Figure 4 shows the effective receptive field of our model and its sub-variants respectively. Our local feature extractor within the GRIB block adeptly captures dense local features, while our medium-range feature extractor harnesses



(a) (b) (c) (d)

Figure 4. The Effective Receptive Field of (a) our full model, (b) our model without (w/o) global feature extractor, (c) our model without medium range feature extractor, and (d) our model without both global and medium range extractor respectively.

features spanning the medium range, and our global feature extractor integrates information across the entirety of the image. Table 3 demonstrates that the global feature extractor exhibits significant importance in the HP-portrait and EgoHOS datasets, characterized by expansive human upper body and arm presence spanning large image areas respectively. Conversely, the medium-range feature extractor showcases its efficacy in the VideoMatte dataset, notable for numerous frames featuring complete human figures and extended limb coverage occupying approximately half of the image.

6. Conclusion

This paper introduces a novel approach for predicting human and manipulating object masks in both first and second-person view videos. To capture relationships between pixels at global, medium-range, and short-range scales, we devise a global reception inductive bias block that combines vision transformer and convolution layers. To enhance model convergence and address the absence of pretrained weights, we conduct pretraining on human segmentation videos and distilled knowledge from depth estimation models. Through extensive experiments, we demonstrate the high accuracy and temporal consistency of our method on segmentation and matting datasets for both first-person and second-person view videos. Our approach surpasses existing real-time human segmentation and matting methods in terms of both accuracy and efficiency.

Figure 5. Visual examples of the human foregrounds extracted using predicted masks from MODNet, RVM, PHOS, Ours-s, and ground truth (GT). The frames are sampled from the EgoHOS video segmentation dataset as a representation of first-person view videos. Ground truth

Figure 6. Visual examples of the human foregrounds extracted using predicted masks from MODNet, RVM, PHOS, and Ours-s. Ground truth are not presented as the YoutubeVis label does not contain manipulated objects. The frames are sampled from the YoutubeVis video segmentation dataset as a representation of second-person view videos.

(a) A portrait video example from the VideoMatte test set without object manipulation.

(b) A full body video example from the VideoMatte test set with object manipulation.

Figure 7. Visual examples of consistent and accurate mask results of sample videos from VideoMatte test set. The predicted mask is overlayed in green on the video frames extracted.

References

[1] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition. *arXiv preprint arXiv:2110.13083*, 2021. 3

[2] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. Louisiana, USA. 5

[3] D Drozdov, M Kolomeychenko, and Y Borisov. Supervisely. *supervise. ly, https://supervise. ly*, 2020. 4

[4] Marco Forte and François Pitié. f, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020. 4

[5] Ester Gonzalez-Sosa, Guillermo Robledo, D Gonzalez-Morin, Pablo Perez-Garcia, and A Villegas. Real time egocentric object segmentation for mixed reality: Thu-read labeling and benchmarking results. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 195–202, 2022. 2

[6] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4130–4139, 2019. 4

[7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. Seoul, South Korea. 3

[8] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. *2022 AAAI/AAAI Conference on Artificial Intelligence*, 36(1):1140–1147, 2022. Virginia, USA. 1, 2, 5, 6

[9] Fanqing Lin, Brian Price, and Tony Martinez. Ego2hands: A dataset for egocentric two-hand segmentation and detection. *arXiv preprint arXiv:2011.07252*, 2020. 2

[10] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 2

[11] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3132–3141, 2022. Hawaii, USA. 1, 2, 3, 4, 5, 6

[12] Sachin Mehta and Mohammad Rastegari. Mobilevit: lightweight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 3

[13] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11696–11706, 2022. Louisiana, USA. 2

[14] Hyojin Park, Lars Sjosund, YoungJoon Yoo, Nicolas Monet, Jihwan Bang, and Nojun Kwak. Sinet: Extreme lightweight portrait segmentation networks with spatial squeeze module and information blocking decoder. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2066–2074, 2020. Colorado, USA. 1, 2

[15] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 4

[16] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. Montreal, Canada. 4

[17] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020. 2

[18] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. *Computer Graphics Forum*, 35(2):93–102, 2016. 4

[19] Yezhi Shen, Weichen Xu, Qian Lin, Jan P. Allebach, and Fengqing Zhu. Depth assisted portrait video background blurring. *Electronic Imaging*, 35(7):273–1–273–1, 2023. California, USA. 4

[20] Yezhi Shen, Weichen Xu, Qian Lin, Jan P. Allebach, and Fengqing Zhu. Real-time end-to-end portrait and in-hand object segmentation with background fusion. *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 242–247, 2023. Brisbane, Australia. 1, 2, 4, 5, 6

[21] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6975–6984, 2021. 4

[22] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2018. Munich, Germany. 2

[23] Zifeng Wu, Yongzhen Huang, Yinan Yu, Liang Wang, and Tieniu Tan. Early hierarchical contexts learned by convolutional networks for image segmentation. *2014 22nd International Conference on Pattern Recognition*, pages 1538–1543, 2014. Stockholm, Sweden. 4

[24] Weijuan Xi, Jianhang Chen, Qian Lin, and Jan P Allebach. High-accuracy automatic person segmentation with novel spatial saliency map. *Proceedings of the International Conference on Image Processing*, pages 1560–1564, 2019. Taipei, Taiwan. 4

[25] Weichen Xu, Yezhi Shen, Qian Lin, Jan P Allebach, and Fengqing Zhu. Efficient real-time portrait video segmentation with temporal guidance. *Electronic Imaging*, 34:1–7, 2022. California, USA. 3

[26] Weichen Xu, Yezhi Shen, Qian Lin, Jan P. Allebach, and Fengqing Zhu. Exploiting temporal information in real-time

portrait video segmentation. *Proceedings of the 4th International Workshop on Human-Centric Multimedia Analysis*, page 33–39, 2023. Ottawa, Canada. 1

[27] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021. 3

[28] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. Seoul, South Korea. 4

[29] Ruifeng Yuan, Yuhao Cheng, Yiqiang Yan, and Haiyan Liu. Real-time segmenting human portrait at anywhere. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2196–2202, 2023. Vancouver, Canada. 2

[30] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. *European Conference on Computer Vision*, pages 127–145, 2022. Tel Aviv, Israel. 2, 4

[31] Song-Hai Zhang, Xin Dong, Hui Li, Ruilong Li, and Yong-Liang Yang. Portraitnet: Real-time portrait segmentation network for mobile device. *Computers & Graphics*, 80:104–113, 2019. 1, 2