10.2352/EI.2022.34.8.IMAGE-263

Efficient Real-time Portrait Video Segmentation With Temporal Guidance

Weichen Xu[1]***, Yezhi Shen***[1]***, Qian Lin***[2]***, Jan P Allebach***[1]***, and Fengqing Zhu***[1]

[1] ***Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN***
[2] ***HP Labs, Palo Alto, CA***

Abstract

Virtual background has become an increasingly important feature of online video conferencing due to the popularity of remote work in recent years. To enable virtual background, a segmentation mask of the participant needs to be extracted from the real-time video input. Most previous works have focused on image based methods for portrait segmentation. However, portrait video segmentation poses additional challenges due to complicated background, body motion, and inter-frame consistency. In this paper, we utilize temporal guidance to improve video segmentation, and propose several methods to address these challenges including prior mask, optical flow, and visual memory. We leverage an existing portrait segmentation model PortraitNet to incorporate our temporal guided methods. Experimental results show that our methods can achieve improved segmentation performance on portrait videos with minimum latency.

Introduction

Video conferencing has been widely used for remote work and entertainment in recent years. However, the background of the scene may reveal information regarding the participant privacy that could be of concern or nuisance. If the background is cluttered or has a strong light source, the conferencing experience is also less pleasant. Virtual background such as blurring the scene background or replacing it with the participant selected background image or video has been commonly used to address these issues. One challenge with using virtual background is that an accurate segmentation of the participant in each video frame is required. Fig. 1 illustrates the process of enabling the virtual background feature in a video conference.

Current segmentation methods mostly rely on using Convolutional Neural Networks (CNN) and have achieved good accuracy on portrait image segmentation [1]–[3]. However, the performance of image based methods tends to degrade for portrait videos. Artifacts such as flickering edge and false positive prediction in the background can be observed largely due to the lack of consideration for temporal information between frames. Video segmentation methods have been investigated for general object segmentation [4]–[7], but are not designed for portrait videos where noticeable portrait edge detail and inter-frame consistency are crucial. It is worth noting that real-time portrait video segmentation may also have a computation constraint where only CPUs can be used for CNN inference, therefore, efficient methods are much to be desired.

In this paper, we propose several temporal-guided methods for video object segmentation, including prior mask, optical flow, and visual memory. Our methods are implemented based on the portrait image segmentation method PortraitNet [3]. Experimental results show that our methods with temporal guidance outperform those without any temporal guidance for public image and video portrait datasets. In addition, the accuracy and runtime comparison also suggest that our prior mask method is both efficient and accurate among the proposed temporal-guided methods.

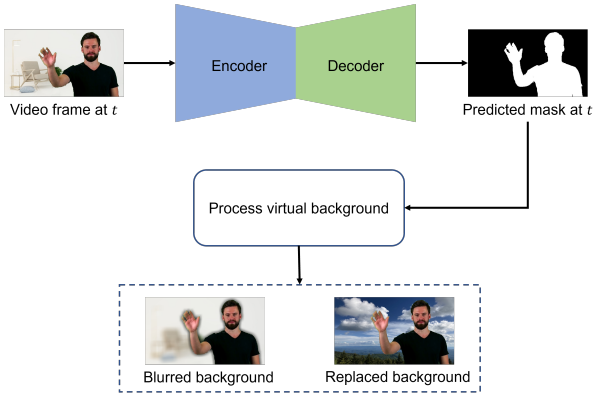


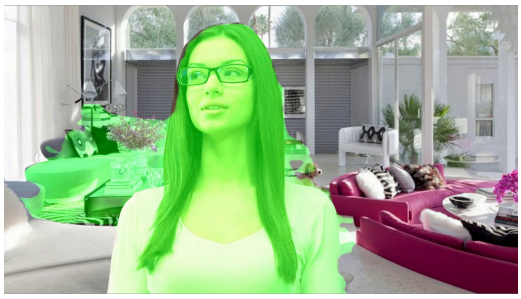
Figure 1. *Illustration of portrait video segmentation to enable virtual background features for real-time video conferencing. The segmentation is trained using an encoder-decoder CNN.*

Related Work

Portrait Image Editing. Portrait image editing has attracted great research interests in recent years. Portrait typically refers to the upper-half region of a person, which is commonly featured in photography. To enable editing for portrait images, an accurate pixel-wise segmentation mask of the person in the image is required. With the advance of convolutional neural network architectures, improved performance has been achieved for semantic segmentation. PortraitNet [3] proposed a lightweight U-shape encoder-decoder model architecture. It uses the MobileNetV2 architecture [8] as the lightweight encoder to extract image features and replaces the standard convolution in the decoder with depth-wise separable convolution, which had fewer parameters and lower inference latency. SINet [1] designed an extremely lightweight portrait segmentation network with only 86.9K parameters and achieved comparable results with less than 1% mIoU degradation than the state-of-the-art PortraitNet on a portrait image dataset EG1800 [2]. EG1800 is the first public portrait image dataset, which provides 1,800 portrait images and manually an-

notated pixel-wise masks. Other public portrait image datasets including Supervise-Portrait [9] and BaiduV1&V2 [10], which were generated by cropping the upper half region from the original annotated full-body person images.

(a) Motion blur



(b) Complicated background

Figure 2. *Example artifacts for portrait video segmentation using image based methods. The overlay of the predicted portrait mask and the original video frame are shown. (a) Motion blur causes segmentation error on the arm. (b) False positive prediction in a complicated background.*

Portrait Video Segmentation. Portrait video segmentation is closely related to portrait image segmentation, since a portrait video can be processed as a series of portrait images. Even though the current best portrait image segmentation networks can achieve more than 95% mIoU accuracy on test images, their performance tends to degrade on real-world videos due to the lack of considering temporal information. Typical artifacts, including mask inconsistency between adjacent frames and segmentation error in the background, greatly reduce the quality of the user experience. Examples of such artifacts are shown in Fig. 2. A straightforward and effective method of incorporating temporal information is optical flow. Jain et al. proposed a fusion model, which had two CNN branches to process the RGB image and the optical flow, respectively, and fused feature maps from both branches [4]. In a recent work [11], Kuang et al. proposed using a processed optical flow result as an auxiliary input to the segmentation network, which emphasized the region showing large motion. However, an accurate optical flow requires additional computation, which introduces non-negligible latency to the real-time applications. A segmentation mask from the previous frame concatenated with the RGB frame as the network input has significantly improved the robustness of video segmentation [5], [12]. Recurrent modules such as LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) have also been proved effective in several video tasks [6], [13], [14].

Portrait Video Datasets. Although there are several public video object segmentation datasets that include labelled human videos such as YouTube-VOS [7], DAVIS [15]. Videos in these datasets are for general object segmentation, thus a dedicated portrait video dataset is preferred in this task. However, such a portrait video segmentation dataset is difficult to obtain due to the high cost of manually annotating each video frame. Kuang, et al. introduced the ConferenceVideoSegmentation-Dataset [11], which contains 6 synthesized videos of online conferencing scenes. Lin et al. proposed the VideoMatte240K matting dataset [16], which contains 484 high-resolution videos of the human foreground and the alpha matte generated from green-screen videos. Thus, a diverse set of video datasets can be generated by composing a person foreground with different background images.

Methods

In this section, we first introduce PortraitNet [3], which we use as the baseline model. We modify its network input and model architecture to enable efficient learning from video temporal information.

PortraitNet Architecture

PortraitNet [3] achieves the state-of-the-art accuracy on real-time portrait image segmentation. It uses a U-shape encoder-decoder structure, which is similar to U-Net [17]. To further reduce inference latency without degrading the segmentation quality, several efficient lightweight model designs were adopted. First, it uses MobileNetV2 [8] as the encoder to extract dense image features from a relatively-small input image size 224×224. Compared with other powerful feature extractors like ResNet [18], MobileNetV2 originally designed for mobile devices can achieve comparable performance with far fewer model parameters and less inference latency. Furthermore, it introduces a lightweight residual block for the decoder transition module, which is used to reconstruct image features. The major improvement of the proposed residual block is replacing the normal convolution with the depth-wise separable convolution, which decomposes a normal convolution into a depth-wise convolution and a point-wise convolution. Additionally, it introduces two auxiliary training losses, namely boundary loss and consistency constraint loss to improve performance.

Prior Mask Guided PortraitNet

In order to efficiently obtain temporal information from the prior segmentation mask, we modify the first convolutional layer of PortraitNet to allow 4-channel input, which are the RGB frame and prior mask based on [5], [12]. We call the modified model *MaskPortraitNet*. Due to the absence of a valid prior mask for the initial frame during video inference, the initial prior mask needs to be set as an empty mask, as is shown in Fig. 3.

In our experiment, we propose two strategies for training MaskPortraitNet with static image and video datasets, respectively. As described in [5], motion information can be learned even from static images, where the prior mask of an image can be simulated by applying affine transformation and thin plate spline smoothing to the ground truth mask to simulate object motion and camera position changes. This simulation technique offers more flexibility for model training since we can simulate the temporal information from images instead of restricting training data to

videos only. Compared to learning the temporal information from static images with simulated motion, training with videos benefits from learning more natural motion, which could make the trained model better adapt to real-world video scenarios.

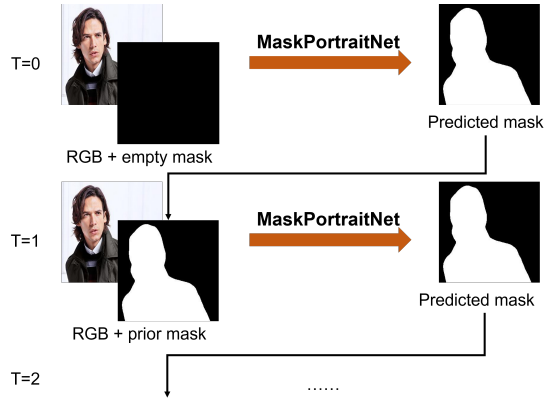


Figure 3. *Illustration of the inference process for a portrait video with MaskPortraitNet, which adds the prior mask as the auxiliary input to incorporate temporal information.*

PortraitNet with Gated Recurrent Unit (GRU)

We propose an RNN-CNN model, *GRUPortraitNet*, which is mostly inherited from PortraitNet except the original decoder transition modules are replaced with efficient ConvGRU modules introduced in [13]. The ConvGRU module can effectively incorporate temporal features represented by the hidden state in the GRU module while processing spatial features in the convolutional layers. However, due to the heavy computation load of the ConvGRU module, only half of the feature maps are sent into the ConvGRU module and the rest are later concatenated with the output feature maps of the ConvGRU module, as is shown in Fig 4.

Compared to other methods which utilize processed temporal guidance as the auxiliary model input to improve the video segmentation performance, RNN-CNN models can effectively incorporate temporal information with only the RGB frame as the model input.

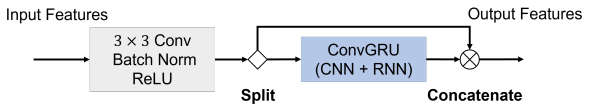


Figure 4. *Decoder transition module with ConvGRU used to reconstruct image features from both spatial and temporal features in GRUPortraitNet. For efficient usage, only half of feature maps are processed in ConvGRU.*

Flow Guided PortraitNet

To enable learning from optical flow features, We utilize the architecture proposed by [11], which introduces a feature mask of large motion as an auxiliary input in addition to the RGB frame to improve the segmentation performance. We call this optical flow based method *FlowPortraitNet*. The pipeline includes two major parts, motion feature extraction and segmentation. First, an accurate optical flow result between two adjacent frames is obtained by running a pre-trained CNN-based optical flow estimation model PWC-Net [19]. Then, a binary feature map is generated by the motion extraction function using the estimated optical flow. For fair comparison, we adopt PortraitNet as the segmentation network with the RGB frame and motion feature map as the network input.

Experiments

In this section, we first describe datasets used for our model training and evaluation. Then, we discuss the experiment's setup and training strategies for each implemented method.

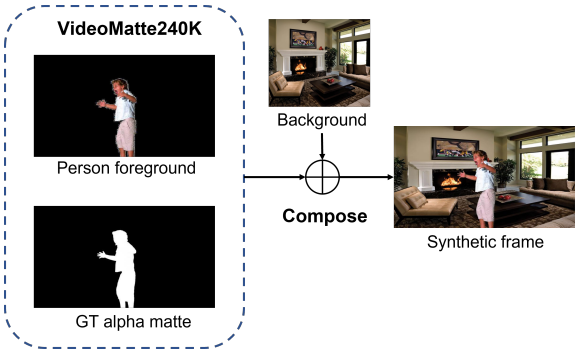


Figure 5. *Example of a synthetic video frame generated by composing a person foreground from VideoMatte240K dataset [16] with a random background image.*

Table 1: Image and video datasets used in our experiments.

Dataset	Type	Train / Test
EG1800 [2]	image	1500 / 300
Supervise-Portrait [9]	image	1858 / 400
BaiduV1&V2 [10]	image	8451 / 2112
ConferenceVidSeg [11]	video	4 / 2 (clips)
VideoMatte240K [16]	video	479 / 5 (clips)

Datasets

The original PortraitNet uses the EG1800 [2] and Supervise-Portrait [9] image datasets. In our implementation, two additional portrait image datasets, BaiduV1&V2 [10] are also used. The total number of training images exceeds 10K, which is desired for video segmentation to cover more diverse scenes.

For training with videos, the VideoMatte240K [16] is used since it provides a large number of videos with detailed person foreground and associated alpha matte. These details also allows use to generate synthetic videos without apparent edge artifacts. Due to licensing restrictions, only 193 background images are released compared to more than 8K background images used in model training reported in the paper [16]. Additionally, we collect 4,018 background images sourced from the Internet, representing different levels of scene complexity. The process of generating synthesized video frames is illustrated in Fig 5.

To evaluate our implementation, we combine the test videos from ConferenceVideoSegmentation [11] and VideoMatte240K. We remove two videos from the test set of VideoMatte240K since

they feature the full-body region and the presence of multiple persons, which is not the target of portrait segmentation.

Training

For model training, we use the binary cross entropy (BCE) loss:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p_i) + (1-y_i) \cdot log(1-p_i) \quad (1)$$

where p_i and y_i denote the predicted probability and the ground truth label for the ith pixel, respectively.

Mean intersection over union (mIoU) is used as the quantitative metric for evaluating segmentation quality:

$$mIoU = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{Pred_i \cap GT_i}{Pred_i \cup GT_i} \quad (2)$$

where $Pred_i$ and GT_i are the predicted and the ground truth masks for the ith images.

All of our experiments are conducted on a single NVIDIA GTX TITAN V graphics card with 12GB memory. For efficiency, all training and testing inputs are resized to 224×224. To make trained models generalizable for different scenes not in our original training data, we apply several augmentation methods including affine transformation, cropping, scale change, horizontal flip, image noise, color shift, brightness change and contrast change. Compared to the offline augmentation method, which takes too much disk space, we adopt the more flexible online augmentation method in our implementation.

PortraitNet is our baseline model, which does not incorporate any temporal information in processing portrait videos. We train the model on all the image datasets listed in Table 1 for 200 epochs with batch size 64. The initial learning rate is set to 0.0001 and is reduced to 0.00001 after 100 epochs. The Adam algorithm is applied as our training optimizer.

MaskPortraitNet takes the segmentation mask from the previous frame as the auxiliary input. We train this model on image datasets and video datasets respectively. For training this model on image datasets, training details are set to be the same as that in the baseline PortraitNet except that the training prior mask is randomly generated from the ground truth mask. We also train this model on the VideoMatte240K video dataset, where the training prior mask comes from the ground truth mask of the previous frame. We only train the model for 50 epochs with a fixed learning rate of 0.00001, which shows good convergence. Since a prior mask is not always available in real-world application, the trained model should work without a valid prior mask. Therefore, we apply an empty mask as the training prior mask for 30% of the training data to improve the robustness of the model.

GRUPortraitNet learns from video sequences. For every epoch, we randomly pick 20 different subsets of video frames from each video in the VideoMatte240K dataset. Each subset includes continuous frames of a certain number. We apply a temporal augmentation method on video sequences including randomly adjusting the video frame rate and reversing the video sequence, in addition to the contextual augmentation for individual frames such as rotation, resizing, and etc. We set the initial learning rate to 0.0001 for the first 50 epochs where the video subset length is set to 15. Then the learning rate is reduced to 0.00001 for another 50 epochs where the video subset length is increased to 25.

FlowPortraitNet also takes an additional fourth channel input, which is the processed motion feature map. Unlike other methods we introduced in this paper that do not require precomputed results, training for FlowPortraitNet relies on the time-consuming process of optical flow estimation and motion feature extraction. If trained on the VideoMatte240K dataset, which has around 240K video frames in total, the computation is too large. Instead, we follow the original training setup in [11], which trains the model on the relatively small ConferenceVideoSegmentation-Dataset. Due to the small dataset size, the model converges in 10 epochs.

Table 2: mIoU results on the evaluation videos using different methods. PortraitNet is the baseline method without any temporal guidance. The other methods are based on PortraitNet and are modified to utilize different temporal guidance.

Method	Dataset	Guidance	mIoU
PortraitNet	image	None	90.5%
MaskPortraitNet	image	Prior mask	92.6%
MaskPortraitNet	video	Prior mask	**96.7%**
GRUPortraitNet	video	Memory	95.4%
FlowPortraitNet	video	Optical flow	94.3%

Results and Discussion

We report the mIoU results on the evaluation videos using different methods in Table 2. Since the baseline PortraitNet does not incorporate any temporal information to improve segmentation robustness for video frames, it has the lowest test accuracy among all methods we implemented. Using the same training image dataset and strategy, MaskPortraitNet which utilizes the piror mask as the auxiliary input channel achieves a 2.1% increase on the test accuracy compared to the baseline PortraitNet trained with the same image dataset. In addition, MaskPortraitNet trained on the video dataset achieves the best result, which shows that training on video datasets helps further improve the performance compared to training on image datasets. This is likely due to the benefit of learning from real prior masks in the video dataset compared to the simulated prior masks from images. GRUPortraitNet and FlowPortraitNet both show improved results compared to the baseline method. A visual comparison of the segmentation quality among these methods is shown in Fig 5.

To compare the efficiency of different methods in our experiment, we also conduct the runtime analysis as is shown in Table 3. For the runtime measurement, an Intel i7-1185G7@3GHz CPU is used where the latency is obtained by averaging over 100 iterations after the initial 10 warmup iterations. Since these methods are all based on PortraitNet, their model sizes are expected to be very close. FlowPortraitNet requires an additional CNN model PWC-Net [19] to estimate optical flow, thus we report their combined model size. It is also worth noting that PWC-Net is not originally designed for mobile devices and can only achieve real-time inference on high-end GPUs. Future work can focus on implementing other optical flow estimation methods, which achieve a balance between the computation and quality.

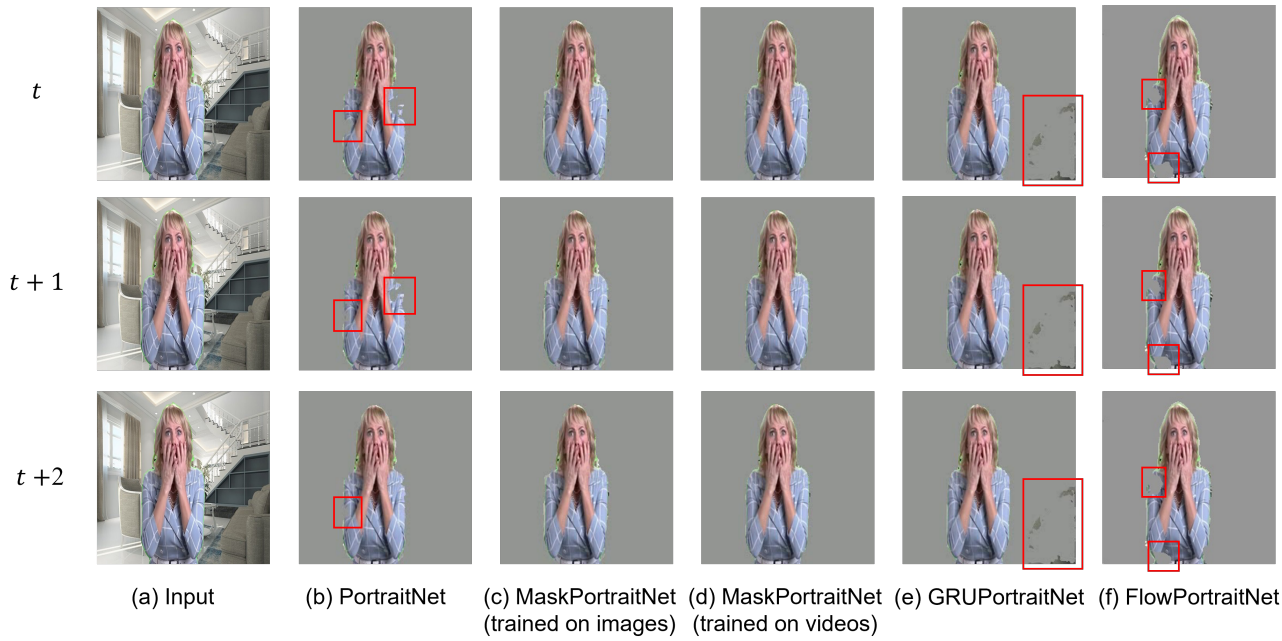


Figure 6. *Visual comparison of segmentation results of a sample video using the different methods proposed. Apparent segmentation regions of error are indicated with red blocks.*

Table 3: Comparison of model size and inference latency of the proposed methods on an Intel i7-1185G7@3GHz CPU. Optical flow estimation in FlowPortraitNet requires an additional computation-intensive PWC-Net, thus the latency is reported on an NVIDIA GTX TITAN V GPU.

Method	Size	Latency
PortraitNet	1.9M	57ms
MaskPortraitNet	1.9M	58ms
GRUPortraitNet	2.3M	66ms
FlowPortraitNet	10.6M	45ms (GPU)

Conclusion

In this paper, we proposed different strategies to incorporate temporal information in portrait video segmentation such as optical flow, visual memory, and prior mask guidance. Our methods are implemented on the image based PortraitNet. Experimental results show that the use of a prior mask will not introduce significant latency and shows the best improvement compared to an image based method without temporal guidance. In addition, training on video datasets shows a future advantage compared to using image datasets when incorporating prior mask for video segmentation.

Our experiments are conducted on public video datasets that are generated from green-screen videos and composited on background images. They may not represent the more dynamic and complex background scenes in real-world portrait videos. Our future work will include a dedicated real-world portrait video dataset to evaluate the performance of proposed portrait video segmentation methods.

References

[1] H. Park, L. Sjosund, Y. Yoo, N. Monet, J. Bang, and N. 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*, pp. 2066–2074, Mar. 2020, Snowmass Village, CO. [Online]. Available: `https://dx.doi.org/10.48550/arXiv.1911.09099`.

[2] X. Shen, A. Hertzmann, J. Jia, *et al.*, "Automatic portrait segmentation for image stylization," *Computer Graphics Forum*, vol. 35, no. 2, pp. 93–102, May 2016. [Online]. Available: `https://dx.doi.org/10.1111/cgf.12814`.

[3] S.-H. Zhang, X. Dong, H. Li, R. Li, and Y.-L. Yang, "PortraitNet: Real-time portrait segmentation network for mobile device," *Computers & Graphics*, vol. 80, no. 1, pp. 104–113, Mar. 2019. [Online]. Available: `https://dx.doi.org/10.1016/j.cag.2019.03.007`.

[4] S. D. Jain, B. Xiong, and K. Grauman, "FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2126, Jun. 2017, Honolulu, HI. [Online]. Available: `https://dx.doi.org/10.48550/arXiv.1701.05384`.

[5] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2663–2672, Jun. 2017, Honolulu, HI. [Online]. Avail-

able: https://dx.doi.org/10.48550/arXiv.1612.02646.

[6] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i-Nieto, "RVOS: End-to-end recurrent network for video object segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5277–5286, Jun. 2019, Long Beach, CA. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1903.05612.

[7] N. Xu, L. Yang, Y. Fan, *et al.*, "Youtube-VOS: Sequence-to-sequence video object segmentation," *Proceedings of the European Conference on Computer Vision*, pp. 585–601, Sep. 2018, Munich, Germany. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1809.00461.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Jun. 2018, Salt Lake City, UT. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1801.04381.

[9] Supervise.ly, https://supervise.ly/.

[10] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," *2014 22nd International Conference on Pattern Recognition*, pp. 1538–1543, Aug. 2014, Stockholm, Sweden. [Online]. Available: https://dx.doi.org/10.1109/ICPR.2014.273.

[11] Z. Kuang and X. Tie, "Flow-based video segmentation for human head and shoulders," *arXiv preprint arXiv:2104.09752*, Apr. 2021. [Online]. Available: https://dx.doi.org/10.48550/arXiv.2104.09752.

[12] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7376–7385, Jun. 2018, Salt Lake City, UT. [Online]. Available: https://dx.doi.org/10.1109/CVPR.2018.00770.

[13] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 238–247, Jan. 2022, Waikoloa, HI. [Online]. Available: https://dx.doi.org/10.48550/arXiv.2108.11515.

[14] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4481–4490, Jun. 2017, Honolulu, HI. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1704.05737.

[15] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool, "The 2019 Davis challenge on VOS: Unsupervised multi-object segmentation," *arXiv preprint arXiv:1905.00737*, May 2019. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1905.00737.

[16] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8762–8771, Jun. 2021. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1911.09099.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, Nov. 2015, Munich, Germany. [Online]. Available: https://dx.doi.org/10.1007/978-3-319-24574-4_28.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Jun. 2016, Las Vegas, NV. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1512.03385.

[19] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," pp. 8934–8943, Jun. 2018, Salt Lake City, UT. [Online]. Available: https://dx.doi.org/10.48550/arXiv.1709.02371.

Author Biography

Weichen Xu *is pursuing a Ph.D. degree in Electrical and Computer Engineering at Purdue University and working as the graduate research assistant in the Video and Image Processing Laboratory. He received his B.Eng degree in Automation from Northeastern University, Shenyang, China. His research interests include video processing and deep learning.*

Yezhi Shen *is a Ph.D. student of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. He received the B.S.E.E. degree in Electrical and Computer Engineering from Purdue University in 2021. His major research area is in computer vision.*

Qian Lin *is an HP Fellow working on computer vision and deep learning research. She is also an adjunct professor at Purdue University. She joined Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. She is inventor/co-inventor for 45 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS&T) in 2012, Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013, and the Society of Women Engineers Achievement Award in 2021.*

Jan P. Allebach *is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, the IS&T/OSA Edwin Land Medal, the IS&T Johann Gutenberg Prize, is a Fellow of the National Academy of Inventors, and is a member of the National Academy of Engineering.*

Fengqing Zhu *is an Assistant Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. Dr. Zhu received the B.S.E.E. (with highest distinction), M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University in 2004, 2006 and 2011, respectively. Her research interests include image processing, computer vision, video compression and digital health. Prior to joining Purdue in 2015, she was a Staff Researcher at Futurewei Technologies (USA). She is a senior member of the IEEE.*