PP-HumanSeg: Connectivity-Aware Portrait Segmentation with a Large-Scale Teleconferencing Video Dataset

Lutao Chu, Yi Liu, Zewu Wu, Shiyu Tang, Guowei Chen, Yuying Hao
Juncai Peng, Zhiliang Yu, Zeyu Chen, Baohua Lai, Haoyi Xiong
Baidu, Inc.

Abstract

As the COVID-19 pandemic rampages across the world, the demands of video conferencing surge. To this end, real-time portrait segmentation becomes a popular feature to replace backgrounds of conferencing participants. While feature-rich datasets, models and algorithms have been offered for segmentation that extract body postures from life scenes, portrait segmentation has yet not been well covered in a video conferencing context. To facilitate the progress in this field, we introduce an open-source solution named PP-HumanSeg. This work is the first to construct a large-scale video portrait dataset that contains 291 videos from 23 conference scenes with 14K fine-labeled frames and extensions to multi-camera teleconferencing. Furthermore, we propose a novel Self-supervised Connectivity-aware Learning (SCL) *for semantic segmentation, which introduces a self-supervised connectivity-aware loss to improve the quality of segmentation results from the perspective of connectivity. And we propose an ultra-lightweight model with SCL for practical portrait segmentation, which achieves the best trade-off between IoU and the speed of inference. Extensive evaluations on our dataset demonstrate the superiority of SCL and our model. The source code is available at https://github.com/PaddlePaddle/PaddleSeg.*

1. Introduction

Portrait segmentation [23] has brought great success in various entertainment applications, such as virtual background, beautifying filters, character special effects. Among these applications, video conferencing has become a major scenario for portrait segmentation, where participants could automatically replace their private backgrounds (e.g., ones from private rooms) with virtual scenes.

The outbreak of coronavirus has further accelerated the prevalence of video conferencing to dramatically replace the traditional face-to-face meetings, as working-from-home has been desired [22]. Moreover, compared to the traditional video conferencing that links participants from different offices/conference rooms, the current meeting scenes become much more diverse in surroundings and lighting conditions, as live videos are recorded from each participant's home. Participants may show various postures and actions, and even wear face masks. In addition, participants sometimes access to the teleconferencing using a thin client, such as a webpage for chat based on JavaScript running on a browser, or a chat App running on mobile devices. Thus, there frequently needs to serve portrait segmentation tasks in resource-limited computing platforms (e.g., webpages and smartphones without powerful GPUs) while ensuring real-time performance for teleconferencing on-the-air. All these practical issues in post-COVID-19 video conferencing have brought great challenges and opportunities to the portrait segmentation field.

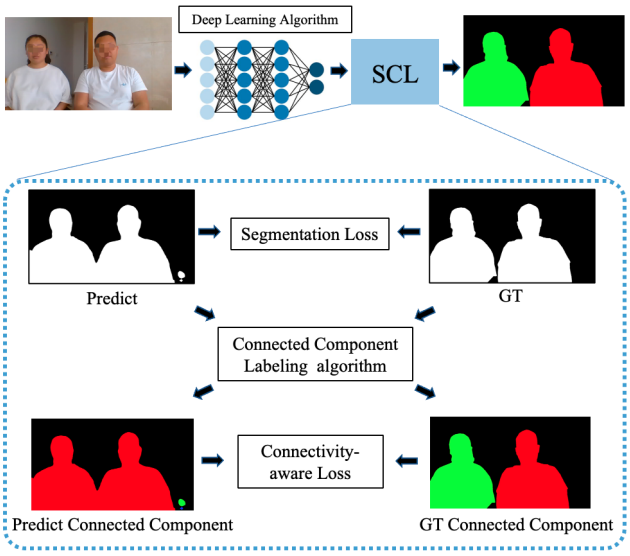

Figure 1. An illustration of the proposed Self-supervised Connectivity-aware Learning (SCL) approach for semantic segmentation, which improves segmentation performance from the perspective of connectivity.

Actually, many works have been done on both datasets

Figure 2. Examples of our dataset and existing datasets. (a) FVS contains only 4 green-screen videos. Due to the composition effect, the labels are not smooth enough. (b) Maadaa contains a lot of similar images and irrelevant information of the interface of the software, e.g. virtual buttons, small windows. (c) The proposed dataset contains various teleconferencing scenes, various actions of the participants, interference of passers-by and illumination change. Note that all videos with human subjects in the proposed datasets have granted the rights to use and disseminate for scientific research purposes.

and methodologies for portrait segmentation. For datasets, there are EG1800 [23], AISeg [2], FVS [13], Maadaa [1], as shown in figure 2. However, they are rarely applied for video conferencing tasks. The datasets for video conferencing either are with low picture quality and high redundancy or even contain synthesized images. *Thus, a new dataset with real-world teleconferencing videos of high picture quality and fine-grained labels is required.*

In terms of segmentation methods, a great number of works have been proposed to address context information [30, 33, 5], multi-scale adaptation [4, 24], fine edge processing [12, 31, 7, 6, 11], category imbalance loss [3, 19] issues. However, these approaches are designed for generic semantic segmentation yet not optimized for portrait segmentation. Although portrait segmentation is a sub-type of semantic segmentation, it has distinct characteristics comparing with other object segmentation. The person can be regarded as a non-rigid object, so that the postures and appearances are varying, which is challenging in the semantic segmentation task. In addition, generic semantic segmentation is pixel-level classification which ignores the completeness of person instances. *Thus, a new learning approach that takes care of completeness of person instances, subject to varying human actions/postures, is required for portrait segmentation in teleconferencing.*

In addition, to achieve portrait segmentation on mobile devices, several lightweight models for semantic segmentation have been proposed [20, 25]. However, the results of these models [20, 25] evaluated on portrait dataset are unsatisfactory. *Thus, a lightweight model that could deliver real-time portrait segmentation on resource-limited platforms (e.g., mobile devices and browsers) is required.*

Therefore, we introduce an open-source solution for practical portrait segmentation named PP-HumanSeg. In this work, we construct a large-scale video portrait dataset including 291 meeting videos in 23 different scenes. To facilitate researchers in the field, we provide 14,117 fine-annotated images. To improve the completeness of person instances, we propose a new Self-supervised Connectivity-aware Learning (SCL) approach, where the connected component concept is used to represent the completeness of the person. The proposed approach improves the consistent connectivity between the segmentation results and the ground truth. Finally, we propose an ultra lightweight segmentation network using SCL, which achieves the best trade-off among mIoU and the inference speed. The contributions of this paper are as follows:

- We release a large-scale video portrait dataset that contains 291 videos from 23 conference scenes with 14K fine-labeled frames provided, to facilitate the progress

in portrait segmentation in video conferencing. Please refer to figure 2 for comparisons with existing datasets, such as FVS [13] and Maadaa [1].

- We propose a novel Self-supervised Connectivity-aware Learning (SCL) framework for portrait segmentation, which improves segmentation performance from the perspective of connectivity.
- We propose an ultra-lightweight model with SCL for practical portrait segmentation, which achieves the best trade-off between performance and the inference speed. Extensive evaluations on our dataset demonstrate the superiority of SCL and our model.

To the best of our knowledge, it is the first video portrait dataset with various scenes, character appearances and actions for video conferencing, with non-trivial baseline models/algorithms offered.

2. Related Works

While the main contributions of this paper include a new dataset, a new learning framework, and a new lightweight model all for portrait segmentation in teleconferencing setting, we thus introduce and discuss the related works from these three perspectives.

Datasets. There are several popular portrait datasets, such as EG1800 [23], FVS [13], Maadaa [1] and AISeg [2]. Compared to EG1800 [23], AISeg [2], and FVS [13] that provided (self-)portrait images and segmentation labels of persons under various indoor/outdoor or even virtual backgrounds, our work offers massive fine-labeled frames of real-world videos for teleconferencing. Maadaa [1] also provided images collected from video conferencing scenarios, but they were all screenshots from the video conferencing applications that incorporate irrelevant and noisy pixels, such as software interfaces. In addition, all existing datasets do not include persons wearing face masks, which is unavoidable for post-COVID-19 teleconferencing.

Learning Methods and Lightweight Models. The existing learning algorithms for semantic segmentation mainly incorporate cross entropy loss, lovasz loss [3], dice loss [19], and RMI loss [35] for training. In addition, upon these training methods, the multi-branch networks have been proposed to improve the lightweight models [29, 20, 21, 18] for generic segmentation problem. Compared to these works, we propose a new SCL framework that incorporates a new loss, namely *self-supervised connectivity-aware loss*, to improve the completeness of segmentation results for person instances and introduce a new model design, namely *ConnectNet* to facilitate ultra-lightweight connectivity-aware portrait segmentation.

Note that some face-related libraries, such as [26, 34], also include face detection modules that can improve the performance of portrait segmentation. Due to the page limits, we do not include the discussion on them here.

3. The Proposed Dataset

In this section, we introduce the ways we collect and label images and videos for portrait segmentation in real-world teleconferencing settings.

3.1. Data Collection

In order to get closer to the real video conference data distribution, we collect the videos in 23 common conference scenes including meeting rooms, offices, public office areas, living room, classrooms, etc. In addition, the participants perform various actions, e.g. waving hands, getting up and sitting down, drinking water, using mobile phones, shaking, etc. We also collected a large number of pictures of people wearing masks. Finally, we get a large-scale dataset of 291 videos with 1280x720 resolution. In order to reduce redundancy, we extract frames from the videos at a low frame rate of 2.5 FPS to get 14117 HD images. The diversity of collected images is shown in figure 2(c).

3.2. Data Labeling

We recruited several professional annotators to label the collected data. They provide high-quality labels of our dataset in both pixel level and video level.

3.2.1 Pixel-Level Labeling

In fact, the annotation of portrait segmentation usually has two ambiguous instances, 1) hand-held items, 2) distant passerby or people with backs. The annotation of them depends on the practical applications, as well as the definition of foreground and background. In video conferencing, the purpose of portrait segmentation is highlighting participant-related parts rather than the surroundings. The hand-held items highly related to the activities of participants, such as mobile phone, glasses and cup. However, distant passerby or people with backs are not participants of the video conference, which should be ignored. Therefore, all hand-held items are labelled together with human body. Distant passers-by or people with backs are not labelled, even though they are usually labelled in other applications of portrait segmentation.

3.2.2 Video-Level Labeling

Following the practice of VOC [9] and PSS [32], we annotate our videos based on the objects appeared in the video. Each video clip has multi-class attributes, e.g. the scene id, the number of participants, the activity of participants,

Figure 3. Examples of composited videos for teleconferencing.

wearing face mask, passers-by. The video-level annotation can be used to video description and multi-task learning, which also provides a good starting point to human activity analysis study in video conferencing.

3.3. Video Synthesis for Teleconferencing

Besides the 14K fine-labeled images, we also collected pure-background images in 90 different video conferencing scenes. Then we use a simple video composition strategy to augment the dataset further. The high-quality portrait masks allow us to extract the portrait parts precisely, and much more labeled images is composed of the extracted portrait parts and pure-background images. Through data composition, we generate around one million images eventually. Due to high-quality annotation, the edges of the composition data are smooth and look natural, as shown in figure 3.

4. SCL: Self-supervised Connectivity-aware Learning for Portrait Segmentation

In this section, we present the design of Self-supervised Connectivity-aware Learning (SCL) framework (shown in figure 1) for semantic segmentation. To improve the completeness of segmentation results, we define a new concept namely *semantic connectivity* to represent the portrait segmentation results and ground truth. Specifically, in addition to using traditional semantic labels as supervision, SCL extracts the connected components from semantic labels and uses them as the supervision signal via a Semantic Connectivity (SC) loss. Note that SCL framework complement with other deep neural architectures (e.g. CNNs, Transformer-based Networks [36, 15, 27, 8]) to boost the performance of portrait segmentation.

4.1. Semantic Connectivity between Components in Segmentation

In this work, we use the connected components to represent the completeness of the portrait segmentation. In topology, connected component is a maximal subset of a topological space that cannot be covered by the union of disjoint subsets. In portrait segmentation, we take the region of a person instance as a connected component. Figure 4 shows an example for connected components calculation and matching.

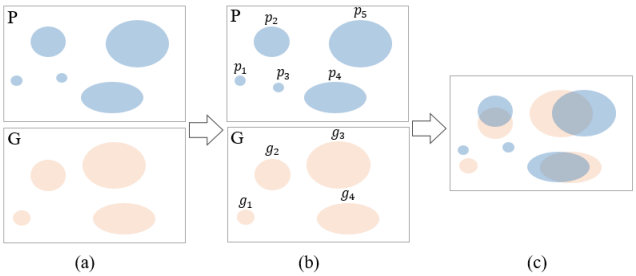

Figure 4. Connected components calculation and matching. (a) It indicates prediction and ground truth, i.e. P and G. (b) Connected components are generated through the CCL algorithm [10], respectively. (c) Connected components are matched using the IoU value.

We find the connected components of predictions (P) and ground truth (G), respectively. Connected components calculation is a fundamental principle in image processing, where there are many methods, e.g. connected component labelling (CCL) and edge thinning. In our approach, we use a CCL algorithm to calculate the connected components, because of its robustness [10]. We then traverse all connected components of G and P to find all pairs that intersect with each other. In figure 4, there are three pairs, i.e. $[g_2, p_2]$, $[g_3, p_5]$, $[g_4, p_4]$, and three isolated components, i.e. p_1, p_3, g_1. Note that a connected component in G could have intersections with multiple connected components in P, which is not be indicated in the figure.

Assuming g_i is paired with $\{p_1, p_2, ..., p_k\}$, the connectivity of g_i is denoted as C_i, which is calculated with the equation as follows.

$$C_i(P) = \frac{1}{k}\Sigma_{k=1}^{k}\text{IoU}(g_i, p_k) \in (0, 1] \tag{1}$$

$$\text{IoU}(g_i, p_k) = \frac{|g_i \cap p_k|}{|g_i \cup p_k|} \tag{2}$$

In particular, when g_i is only paired with one connected component in P, e.g. p_j, C_i equals to IoU between g_i and p_j. If g_i is an isolated component, C_i equals to 0.

Finally, we define the semantic connectivity (SC) of the entire image given the graph of components in the ground truth G and the graph in the prediction P as the follow.

$$\text{SC}(P, G) = \frac{1}{N}\Sigma_{i=1}^{N}C_i(P) \tag{3}$$

where N is the total number of both pairs and isolated components. Note that for $\forall P, G$ we have $\text{SC}(P, G) \in [0, 1]$.

4.2. Learning with SC Loss

To enable the self-supervised connectivity-aware learning, the SCL frameworks uses a novel loss function based on the proposed semantic connectivity, which minimize the

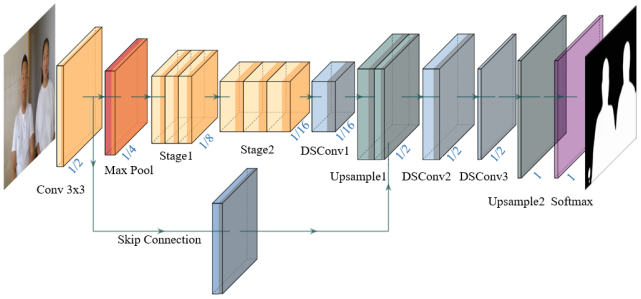

Figure 5. ConnectNet: an Ultra-lightweight Model for Portrait Segmentation.

inconsistency of connectivity between the prediction and the ground truth. In addition, when no intersection between the prediction and the ground truth, we use an area-based loss function to better optimize the model.

The mathematical notation is the same as in the previous section, we denote Semantic Connectivity-aware (SC) Loss as L_{SC}. If there is at least a pair between P and G, L_{SC} is defined as follow.

$$L_{\mathrm{SC}}(P, G) = 1 - \mathrm{SC}(P, G)\ , \tag{4}$$

where for $\forall P, G$ we have $L_{\mathrm{SC}}(P, G) \in [0, 1]$.

Note that there is a special case that no pair exists between P and G, and connectivity is becoming to be 0. It could happen in the beginning of training, due to random initialization of parameters. However, 0-connectivity in SCL would lead to zeros gradients, and the weights cannot not be updated accordingly the connectivity. For such special case, we design a non-trivial loss function to cold start the process. Specifically, to ensure the continuity and differentiability of the loss function in the cold-start phase, we write the SC loss L_{SC} as follow.

$$L_{SC}(P, G) = \frac{|P \cup G|}{|I|}, \tag{5}$$

where I represents the image and $|\cdot|$ represents the area of the region (total number of pixels in the region), and for $\forall P, Q$ we have $L_{SC}(P, G) \in (0, 1]$.

Finally, SCL incorporates the SC Loss as a regularizer to complement with the segmentation losses (denoted as L_S. e.g. cross entropy loss) in the form of $L = L_S + \lambda * L_{\mathrm{SC}}$ to optimize the model. The hyper-parameter λ denotes a weight to make trade-off between the SC loss and the segmentation loss.

5. ConnectNet: an Ultra-lightweight Neural Network for Portrait Segmentation

We propose an ultra-lightweight segmentation network to work with SCL, namely *ConnectNet*, as shown in figure 5. ConnectNet adopts an encoder-decoder structure. The encoder follows an inverted bottleneck block [16] design with channel-shuffle operation to extract features efficiently. To reduce the computation loads while maintaining high resolutions, ConnectNet compresses the number of stages and channels, where every stage is stacked by multiple inverted bottleneck blocks. Moreover, ConnectNet incorporates depth-wise separable convolution to improve the decoding efficiency in the decoder, where depth-wise separable convolution decomposes the ordinary convolution into depth-wise convolution and point-wise convolution so as to further reduce computation loads.

With features extracted in an encoder-decoder network with bottleneck layers, the encoder would lower the resolution of the feature map and lose the spatial details. Spatial information is critical in segmentation tasks. Therefore, the proposed network connects the encoder and decoder across layers through a skip connection to integrate the underlying texture features, which is more conducive to generating fine masks. At the same time, the skip connection directly reuses the features extracted by the encoder without additional computation costs.

6. Experiments

6.1. Experiment settings

All of our experiments are conducted on two Tesla V100 GPUs of 32GB using PaddlePaddle[1] [17]. Code and pretrained models are available at PaddleSeg[2] [14]. During training, we use polynomial decay with power equal to 0.9, and the learning rate equals to 0.05 and 0.025 for HRNet-W18-small and other networks respectively. We use SGD as our optimizer with weight decay parameter being 0.0005. We apply data augmentation methods including scale, crop, flip, and color distortion for training. We use BBDT algorithm [10] for connected component labeling.

In order to avoid similar images in the validation set and test set, we divide the dataset by scene level. The proposed dataset is randomly divided into a training set with 11 scenes and 9006 images, a validation set with 6 scenes and 2549 images, and a test set with 6 scenes and 2562 images. We train our model with the batch size of 128. For all experiments, we take mIoU and pixel accuracy as evaluation metrics.

6.2. Experiment Results

6.2.1 Hyper-parameter

SCL optimizes the network using a weighted combination of cross entropy loss and SC loss. Different combination coefficient may bring different effects. In order to show that the SC loss is parameter in-sensitive and robust. We conduct

[1]https://github.com/PaddlePaddle/Paddle
[2]https://github.com/PaddlePaddle/PaddleSeg

5 experiments with different weight coefficients, i.e. 0.01, 0.05, 0.1, 0.5, and 1.0.

As shown in Table 1, the connectivity of model's prediction is improved on different combination of SC loss and segmentation loss. We set λ as 1.0 in the following experiment settings.

loss ratio	baseline	0.01	0.05	0.1	0.5	1.0
mIoU	93.0	94.2	93.9	94.5	92.6	**94.6**

Table 1. Robustness of SC loss under different ratios

6.2.2 Ablation study on various models

We evaluate the effectiveness of our SC loss on light-weight networks including HRNet-W18-small [25], BiseNetV2 [28] and ConnectNet. As shown in table 2, SC Loss is effective across these networks, where the mIoU metric improves in HRNet-W18-small, BiseNetV2, and ConnectNet respectively. Through enhancing the connectivity of the connected components, the models obtain the better segmentation performance.

Model	mIoU	Pixel Acc
HRNet-W18-small	93.0	97.2
HRNet-W18-small + SCL	**94.5**	97.8
BiseNetV2	85.8	94.2
BiseNetV2 + SCL	**87.5**	94.8
ConnectNet	94.1	97.6
ConnectNet + SCL	**94.6**	97.6

Table 2. Ablation study on light-weight networks

6.2.3 Comparision with other SOTA losses

In this section, we prove the superiority of SC loss over other state-of-the-art losses including lovasz loss [3], and RMI loss [35]. We conduct these experiments on HRNet-W18-small with learning rate being 0.5. For fair comparison, we set the coeffecient of the compound losses as 0.01 for all of the experiments.

As shown in Table 3, the SC loss we propose outperforms other loss methods. The experiment with SC loss has the best score in mIoU and pixel accuracy. These loss focus on different aspects of semantic segmentation like class imbalance and structural information. The evaluation result shows the SC loss has SOTA performance for portrait segmentation.

6.3. Effectiveness of ConnectNet

In order to validate the performace of our proposed model, we compare its performance compared with other light-weight state-of-the-art models, including BiseNetV2 [28], Fast SCNN [20], HRNet [25]. As shown in Table 4, our model is faster and more effective than other SOTA light-weight models. Compared with HRNet-W18-small, our model has greater performance and 41% faster. Compared with Fast SCNN and BiseNetV2, our model is 1.5-3ms slower, but 1.2% and 8.5% higher in mIoU than BiseNetV2 and Fast SCNN, respectively.

Loss	mIoU	Pixel Acc
CE Loss (baseline)	93.0	97.2
CE Loss + Lovasz Loss	93.0	97.2
CE Loss + RMI Loss	94.3	97.7
CE Loss + SC Loss	**94.5**	**97.8**

Table 3. Comparision with SOTA losses

The experimental results show that our model outperforms BiseNetV2 and Fast SCNN to a great extend, but only have less than 10% of their parameters. This is crucial in mobile and web applications considering that the storage requirement is rather strict.

Model	mIoU	Pixel Acc	Infer Time	Params
BiseNetV2	85.8	94.2	10.0	2.32
Fast SCNN	85.7	93.9	**8.6**	1.44
HRNet-W18-small	93.0	97.2	19.76	3.95
ConnectNet	**94.2**	**97.6**	11.5	**0.13**

Table 4. Benchmark on the state-of-the-art lightweight models. The unit of inference time is ms and the unit of Params is M.

6.3.1 Qualitative Comparison

In order to qualitatively show the performance of our network, We visualize the predictions of different networks on test images. As shown in figure 6, our model has better completeness than other models, and it is less prone to make disperse predictions.

7. Conclusion

To facilitate the progress in portrait segmentation in a video conferencing context, we introduce an open-source solution named PP-HumanSeg. In this work, we first construct a large-scale video portrait dataset that contains 291 videos from 23 conference scenes with 14K fine-labeled frames provided. To improve the completeness of segmentation results, we propose a Self-supervised Connectivity-aware Learning (SCL) framework incorporating a novel Semantic Connectivity (SC) loss. Such SC loss models the topology of portrait segmentation as a graph of connected components and measures the inconsistency between the graphs (i.e., connectivities) extracted from the ground truth labels and the prediction results as the loss. Furthermore,

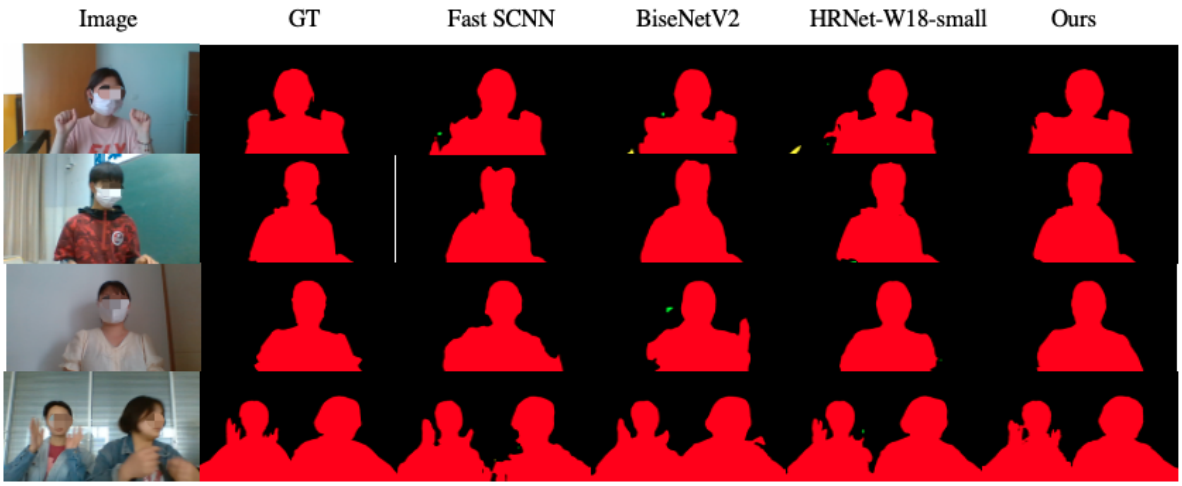

Figure 6. Semantic segmentation results of different light-weight networks

we propose an ultra-lightweight model, namely ConnectNet, with SCL for practical portrait segmentation. The proposed solution achieves the best trade-off between IoU and inference time in the dataset. Extensive evaluations on our dataset demonstrate the superiority of SCL and ConnectNet. The comparisons with other algorithms also show the advantage of proposed datasets from the perspectives of coverage and comprehensions.

Acknowledgement

This work was supported by the National Key Research and Development Project of China (2020AAA0103500).

References

[1] Human-centric video matting challenge, maadaa.ai. https://maadaa.ai/, 2021.

[2] Portrait segmentation dataset, aiseg. https://github.com/aisegmentcn/matting_human_datasets, 2021.

[3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[6] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15334–15342, 2021.

[7] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *European conference on computer vision*, pages 660–676. Springer, 2020.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[10] Costantino Grana, Daniele Borghesani, and Rita Cucchiara. Optimized block-based connected components labeling with decision trees. *IEEE Transactions on Image Processing*, 19(6):1596–1609, 2010.

[11] Yuying Hao, Yi Liu, Zewu Wu, Lin Han, Yizhou Chen, Guowei Chen, Lutao Chu, Shiyu Tang, Zhiliang Yu, Zeyu Chen, et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1551–1560, 2021.

[12] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

[13] Zijian Kuang and Xinran Tie. Flow-based video segmentation for human head and shoulders. *CoRR*, abs/2104.09752, 2021.

[14] Yi Liu, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Baohua Lai, and Yuying Hao. Paddleseg: A high-efficient development toolkit for image segmentation, 2021.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[16] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision*, pages 116–131, 2018.

[17] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. Paddlepaddle: An open-source deep learning platform from industrial practice. 1(1):105–115, 2019.

[18] Davide Mazzini. Guided upsampling network for real-time semantic segmentation. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 117. BMVA Press, 2018.

[19] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[20] Rudra Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. In Kirill Sidorov and Yulia Hicks, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 187.1–187.12. BMVA Press, September 2019.

[21] Rudra P. K. Poudel, Ujwal Bonde, Stephan Liwicki, and Christopher Zach. Contextnet: Exploring context and detail for semantic segmentation in real-time. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 146. BMVA Press, 2018.

[22] Libby Sander. Coronavirus could spark a revolution in working from home. are we ready. *The Conversation*, 11, 2020.

[23] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, pages 93–102. Wiley Online Library, 2016.

[24] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[25] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[26] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face.evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021.

[27] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[28] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, pages 1–18, 2021.

[29] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[30] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[31] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pages 489–506. Springer, 2020.

[32] Yu Zhang, Chang-Bin Zhang, Peng-Tao Jiang, Feng Mao, and Ming-Ming Cheng. Personalized image semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[33] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[34] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

[35] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 11115–11125, 2019.

[36] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.