PORTRAIT INSTANCE SEGMENTATION FOR MOBILE DEVICES

*Lingyu Zhu**†[1] *Tinghuai Wang**[2] *Emre Aksu*[2] *Joni-Kristian Kämäräinen*[1]

[1] Tampere University, Finland
[2] Nokia Technologies, Finland

ABSTRACT

Accurate and efficient portrait instance segmentation has become a crucial enabler for many multimedia applications on mobile devices. We present a novel convolutional neural network (CNN) architecture to explicitly address the long standing problems in portrait segmentation, *i.e.,* semantic coherence and boundary localization. Specifically, we propose a cross-granularity categorical attention mechanism leveraging the deep supervisions to close the semantic gap of CNN feature hierarchy by imposing consistent category-oriented information across layers. Furthermore, a cross-granularity boundary enhancement module is proposed to boost the boundary awareness of deep layers by integrating the shape context cues from shallow layers of the network. We further propose a novel and efficient non-parametric affinity model to achieve efficient instance segmentation on mobile devices. We present a portrait image dataset with instance level annotations dedicated to evaluating portrait instance segmentation algorithms. We evaluate our approach on challenging datasets which obtains state-of-the-art results.

Index Terms— Convolutional Neural Networks, Semantic Segmentation, Instance Segmentation, Portrait Segmentation

1. INTRODUCTION

The proliferation of digital cameras of mobile devices and the social trend for casually capturing and sharing media have led to an explosive need of portrait photo editing and processing. Portrait segmentation is becoming a crucial enabler to facilitate person-centered photo enhancement, rendering, editing and manipulation. Such applications typically demand very accurate segmentation since the human visual system has a strong sensitivity to artefacts caused by mis-segmentations of portrait images [1]. Yet such mis-segmentations frequently occur when applying general purpose semantic segmentation algorithms on portrait images.

Recently, fully convolutional networks (FCNs) [2, 3, 4, 5, 6] based methods have been proposed to address the semantic segmentation problem. For example, DeepLabv3+ [3] adopts an encoder-decoder architecture with Atrous Spatial Pyramid Pooling (ASPP) [7] for semantic segmentation. Despite of the tremendous progress of the above methods toward semantic segmentation, much remains to be addressed in order to warrant satisfying region coherence and boundary accuracy for artefacts-free portrait photo processing purpose. Moreover, the above networks are trained on general natural image datasets rather than specific portrait images. PortraitFCN+ [8] is proposed as a dedicated portrait segmentation model which is a fine-tuned FCN using portrait images. However, it requires additional prior knowledge of portrait position and shape information which is subject to the availability and effectiveness of face detection results. Furthermore, PortraitFCN+ does not explicitly address the fundamental challenges of portrait segmentation which are deeply rooted in the inherent design limitation of the network architecture.

Portrait segmentation networks assign per-pixel category level labels, *i.e.,* person or background. However, for the scenarios where there are more than one person in the photo, state-of-the-art approaches would fail to distinguish individual instances, due to the inherent limitation of FCNs - translation invariance. Nonetheless, person-specific editing and manipulation in multi-person photos remains useful in various social events [1, 9, 10, 11]. Parametric models for instance-level segmentation [12, 13, 14, 15, 16] have been proposed which, however, require large amount of instance-level ground-truth data for training. Furthermore, embedding an object proposal network in those architectures typically leads to poor boundary delineation due to the cascaded processings upon downsampled feature maps.

Motivated by the above, we propose a novel approach addressing the portrait instance segmentation problem with four main contributions. Firstly, we propose a cross-granularity categorical attention mechanism to close the semantic gap between CNN feature hierarchies in order to achieve semantic region coherence. This novel attention mechanism alleviates the semantic information loss during the feature transformation process which is common in existing FCN based approaches. Secondly, we explicitly incorporate a contour detection functionality in our architecture to extract boundary information and design a boundary enhancement module to propagate the rich boundary information from early layers to later layers. Thirdly, we propose a 'plug-n-play' non-parametric affinity model for portrait instance segmentation which can be easily integrated with any semantic segmentation framework to achieve instance-level segmentation without any instance level training data. Lastly, we present a portrait image dataset with accurate instance level annotations for evaluating portrait instance segmentation.

2. METHOD

We firstly introduce our novel architecture specialized to achieve accurate portrait segmentation, followed by a non-parametric affinity model which facilitates our ultimate goal of portrait instance segmentation.

* Equal contribution
† This work is done in Nokia Technologies

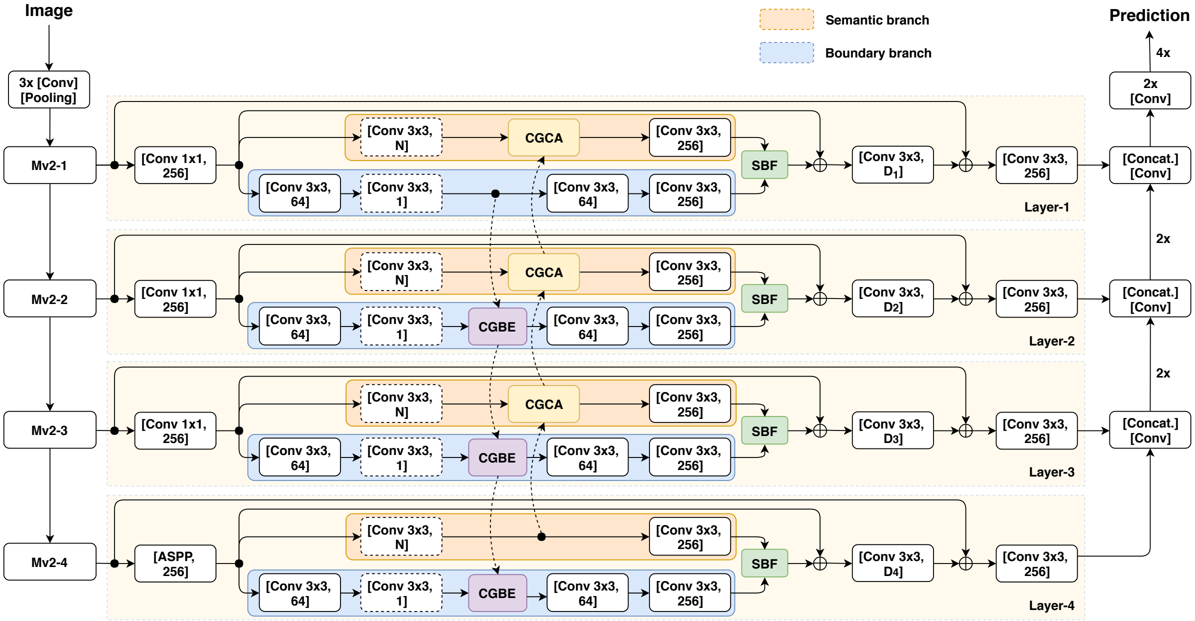


Fig. 1: An overview of our proposed architecture which mainly comprises two parts: an encoder and a decoder. Four layers take as input of the feature maps from encoder: Mv2, which propagate consistent categorical information and boundary enhancement across feature hierarchy in opposite directions. The dashed blocks in each layer are associated with the auxiliary loss functions during training.

2.1. Deep Portrait Segmentation

Our proposed architecture for portrait segmentation mainly comprises two parts: an encoder and a decoder, as illustrated in Figure 1. We adopt MobileNetv2 (Mv2) [17] as the encoder due to its efficiency and light weight. Mv2 consists of four layers according to the size of feature maps, namely Mv2-i ($i = 1, 2, 3, 4$). At each layer, Mv2-i interfaces with the proposed decoder by providing the feature map as input to its corresponding four layers which are denoted as Layer-i in this work. Layer-i firstly projects the feature to 256 dimensions using either an 1×1 convolutional layer (for Layer-1, -2 and -3) or an ASPP [7] (for Layer-4). Thereafter, the feature maps are fed into two branches inside each layer, *i.e.,* semantic branch and boundary branch.

2.1.1. Semantic branch

Semantic branch aims to project feature into categorical space and enhance its level of semantic information by imposing a consistent semantic flow from deeper layers. On the semantic branch, the feature map passes though a 3×3 convolution operator with N neurons, where N is the categories number of training dataset, whereby the *categorical feature* is guided with category-wise deep supervisions. The supervised *categorical feature* is then forwarded to the proposed *cross-granularity categorical attention* module (Section 2.1.4) to be semantically enhanced by fusing with the categorical information of deeper layers, which is consecutively projected back to the same dimension space with filters 256.

2.1.2. Boundary branch

Boundary branch embeds a boundary detection functionality for extracting boundary features which are enhanced by the rich boundary information propagated from shallow layers. In the boundary branch, the feature map is projected to 64 dimensions and then 1 dimension gradually with 3×3 convolution layers. Similar to semantic branch, deep supervision is adopted in the boundary branch to generate the 1-channel feature map, supervised by the boundary information generated by running a sober edge operator on segmentation groundtruth. The learned *boundary feature* is then passing to the proposed *cross-granularity boundary enhancement* module (Section 2.1.5) though the network from shallow layer to deeper layer to enhance the boundary awareness, which is consecutively projected back to the same dimension space with filters 64 and 256.

2.1.3. Semantic and boundary fusion unit

The semantic branch output from each Layer-i is fused with the feature map from corresponding boundary branch in the proposed *semantic and boundary fusion* (SBF) module. The SBF integrates the learned boundary information with the semantic features by passing the boundary feature through a sigmoid operator, an element-wise multiplication and a summation with the semantic feature. The fused feature is thereafter projected to D_i which denotes the feature dimension from Mv2-i. Instead of learning a direct mapping, two residual connections [18] are incorporated to each Layer-i whereby 256 and D_i feature maps are summed respectively to avoid feature degradation and promote network convergence. The feature map is then projected to 256 channel for all layers with 3×3 convolution operation as output.

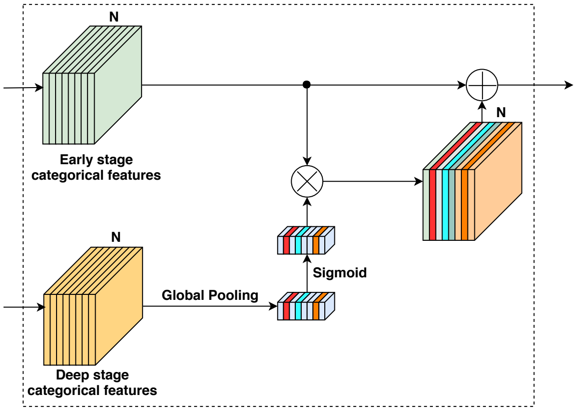


Fig. 2: The architecture of proposed cross-granularity categorical attention module.

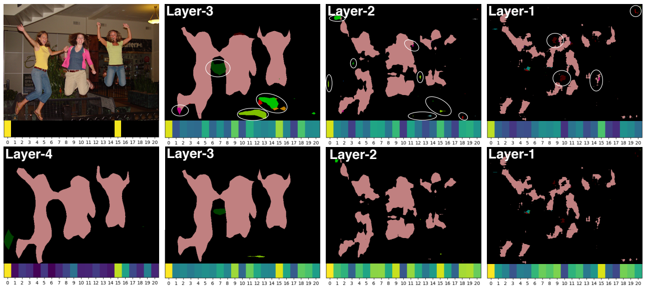


Fig. 3: Intermediate semantic labellings and categorical attentions before (first row) and after (second row) cross-granularity categorical attention (CGCA). There are considerable amount of mis-labelings before CGCA caused by the semantic ambiguities. These ambiguities can be observed from the color bar indicating the corresponding categorical attention weight vector below the label map. The color bar shows weak categorical attentions on categories 0 and 15 (*i.e.,* background and person respectively in Pascal VOC 2012 dataset) before CGCA. After CGCA, the semantic ambiguity issue residing in the lower-layer features is largely resolved indicated by both the corrected mis-labelings and the increased attentions on consistent categories, *i.e* background and person, illustrated in the color bar.

2.1.4. *Cross-granularity categorical attention*

In CNNs, deeper layers learn rich semantic information while shallow layers lack of sufficient semantic knowledge to make accurate semantic prediction despite of their high spatial resolution. Most recent semantic segmentation networks suffer the challenges of semantic incoherence issue caused by this semantic inconsistency among feature hierarchies. We propose a cross-granularity categorical attention (CGCA) module to adopt the category-oriented information encoded from deeper layers to guide the early layers.

The CGCA, as illustrated in Figure 2, encodes the global categorical attention from the N-channel categorical feature map of deeper layer by a global average pooling followed by a sigmoid operator. The CGCA is achieved by multiplying the global categorical attention informed by later layer, *i.e* the weight vector in Figure 3, with the feature map from current layer to adjust channel-wise interpretations. Finally, the category-enhanced feature is summed with the current feature map. This attention plays a crucial role in maintaining consistent categorical information across feature hierarchy and bridging the semantic gaps. Figure 3 shows the intermediate semantic labellings as well as the categorical attentions before and after the CGCA module, where we can see that CGCA significantly resolves the semantic ambiguity issue residing in the lower-level features and makes the features more attentive on consistent categories 0 and 15 (*i.e.,* background and person respectively in Pascal VOC 2012 dataset) in this example.

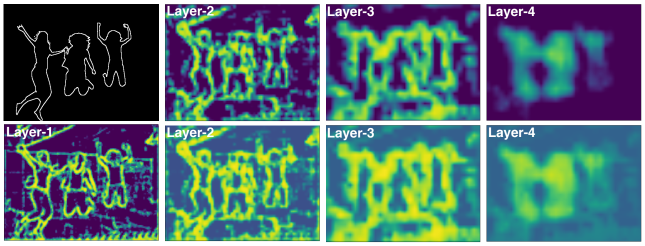


Fig. 4: Boundary prediction before (first row) and after (second row) cross-granularity boundary enhancement (CGBE). CGBE enhances the boundary awareness of the deep-layer features, especially the boundary attention to Layer-4.

2.1.5. *Cross-granularity boundary enhancement*

Boundary information is proven effective for localizing objects in space and scale which in turn provide better boundary delineation and shape context cues for semantic segmentation task against within-class variations. Encoder extracts appearance and contextual information at various hierarchies, with decreasing spatial details and increasing semantic information from Mv2-1, -2, -3 to Mv2-4. Figure 4 shows the detected boundaries from Layer-1, -2, -3 to Layer-4 which shows that the boundary information is mainly preserved in earlier layers whilst the boundary information detected in later layers is very weak. Therefore, we propose a simple yet effective cross-granularity boundary enhancement (CGBE) module to enhance the boundary awareness of deeper layers features. Specifically, CGBE achieves boundary enhancement by element-wise summation between the learned boundary information from early layer and the boundary detection of current layer. Figure 4 shows the element-wise boundary enhancement of four layers respectively, which demonstrates that the boundary branch from each layer is able to capture object boundaries despite of its simplicity and the proposed CGBE module significantly enhances the boundary awareness especially the deeper layers, *e.g.,* Layer-4, despite of its simplicity.

2.2. Portrait Instance Segmentation

The proposed network architecture assigns per-pixel binary category labels, *i.e.,* person or background, rather than individual person instances. Yet, portrait instance segmentation remains useful in personalized editing or manipulation of multi-person photos taken in various social events. To this end, we introduce an instance segmentation method

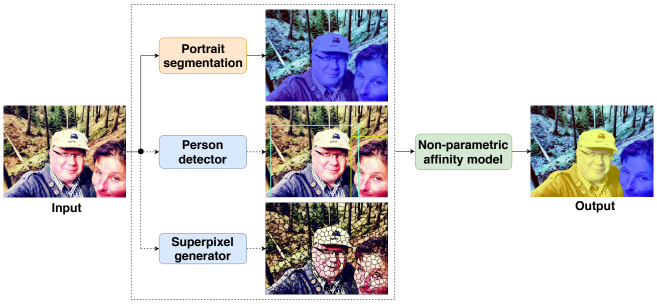


Fig. 5: The proposed method for semantic instance segmentation.

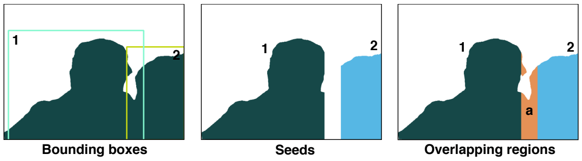


Fig. 6: The left figure illustrates the portrait segmentation map and detected person instances associated to bounding boxes. The middle figure shows the generated seeds for each person instance (*i.e.,* known identity) and the right figure shows the overlapping regions which comprise pixels with unknown identities.

to combine both the instance- and category-level features, which turns out to be effective and efficient in resolving inter-instance ambiguities. The proposed method is schematically illustrated in Figure 5, where the multi-person photo is firstly processed by a portrait segmentation network, a person detector and a superpixel generator, producing person/non-person labels, person bounding boxes and superpixels respectively. The instance segmentation is sequentially produced by the proposed non-parametric affinity model.

There are mainly two scenarios in portrait instance segmentation given the spatial layout of people appearing in the image. We utilize person detector to roughly determine the spatial location and range of person instances, based on which two approaches are proposed to tackle these two scenarios respectively.

In the first scenario, the person instances do not overlap or occlude each other. The occlusion relationships can easily be determined by examining the detected bounding boxes. If there is no overlap among the bounding boxes from person detector, the isolated regions generated by the portrait segmentation network can be used as the final instance segmentation.

The major challenge in instance segmentation is resolving the occlusion or overlap between people which is not addressed in portrait segmentation network. It can be generally associated to the scenario where the detected person bounding boxes are overlapping. Figure 6 illustrates an example of overlapping bounding boxes generated from person detector. The bounding boxes generated from object detector identify each object instance with its location and spatial range, which naturally assigns non-ambiguous instance labels to non-overlapping regions. The instance labels inside the areas where multiple bounding boxes overlap remain uncertain. The instance segmentation task can now be formulated as a problem to assign the generated pixels inside the overlapping regions (Figure 6 right) with labels with respect to the known identities (*i.e., Seeds* in Figure 6 middle).

In order to categorize the uncertain pixels into the existing corresponding identities, we propose a novel method to compute affinities of uncertain pixels with respect to all identities. We correlate the affinity field with the distance in both of appearance and spatial location of each pixel w.r.t. each known identity, which can be characterized as the geodesic distance from the known identities to classify the pixels in the overlapping regions. The geodesic distance is the smallest integral of a weight function over all possible paths from the seeds area to each uncertain pixel. To enable efficient computation and improve the robustness against local noise, we utilize superpixels rather than pixels. The geodesic distance between two superpixels on spatial-aware feature maps is formulated in Equation 1, where $s_a = I(x_a, y_a)$ and $s_b = I(x_b, y_b)$ are connective superpixels along the path $S_{s_1 s_2}$ on feature map I which is the normalized histogram of each superpixel, and the weights $W_{s_a s_b} = \frac{1}{2}\sum\frac{(s_a-s_b)^2}{s_a+s_b}$ are defined as χ^2 distance between the histogram of neighbor superpixels. Afterwards, we obtain the smallest summation of a weight function over all possible paths between superpixel s_1 and s_2. The affinity values can be defined disproportionally to the computed weighted distance, *i.e.,* smaller distance indicates higher affinity between two superpixels.

$$d(s_1, s_2) = \min_{S_{s_1 s_2}} \sum_{s_a s_b} W_{s_a s_b} \quad s_a, s_b \in S_{s_1 s_2} \tag{1}$$

Therefore, the affinities between pixels in overlapping region and all the seeds are calculated and each pixel with unknown identity is consequently assigned with the instance label with which it has the highest affinity.

3. EXPERIMENTS

We evaluate the performance of our proposed portrait segmentation architecture and non-parametric affinity instance segmentation algorithm on three datasets. The performance of portrait segmentation is measured in terms of pixel mean Intersection-over-Union (mIoU) while portrait instance segmentation is measured with Average Precision (AP).

3.1. Datasets

PASCAL VOC 2012: The PASCAL VOC 2012 [19] dataset consists of 20 foreground object classes and one background class. The original dataset includes 1,464 (*train*), 1,449 (*val*) and 1,456 (*test*) images with pixel-level annotations. The dataset is augmented by [20], contributing 10,582 (*trainaug*) training images.
Portrait: The portrait dataset [8] consists of 1800 (1500 for training and 300 for validation) images with resolution 800 × 600.
Our dataset: In order to evaluate portrait instance segmentation, we present a new dataset containing 50 multi-person portrait style images with instance level annotations. Those images are mostly captured with mobile front cameras with large variations in poses, scenes, scales, people count and

Method	mIoU (%)
DeepLabv3-Mv2 [7]	75.32
Our method	**76.80**

Table 1: Performance on PASCAL VOC 2012 *val* set.

Method	mIoU (%)
FCN (Person Class) [8]	73.09
PortraitFCN [8]	94.20
PortraitFCN+ [8]	**95.91**
Our method (Person Class)	92.50
Our method with Portrait	95.37

Table 2: Quantitative Comparison results of different semantic segmentation methods on portrait [8] *val* set.

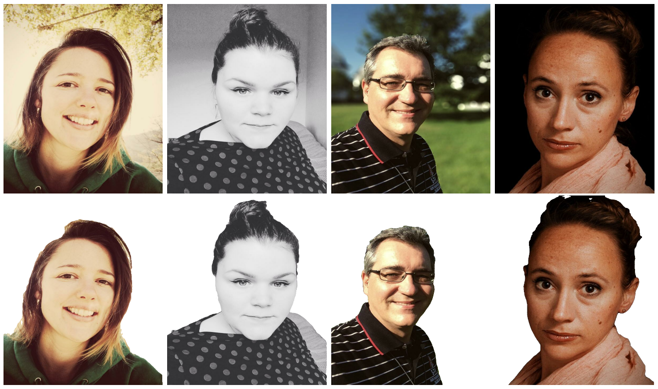
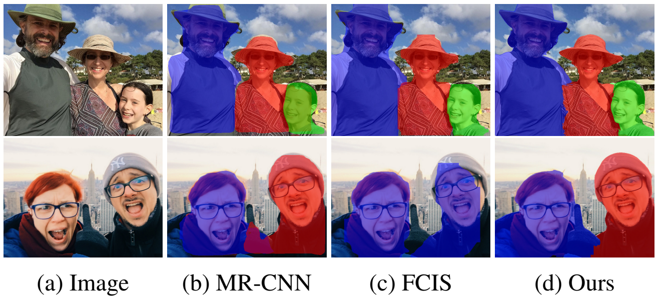


Fig. 7: Visualization examples of portrait segmentation



(a) Image (b) MR-CNN (c) FCIS (d) Ours

Fig. 8: Comparison of portrait instance segmentations.

occlusion relationships which pose challenges for evaluating the generalization and robustness of our proposed portrait instance segmentation approach.

3.2. Results

3.2.1. *Deep Portrait segmentation*

We train the proposed portrait segmentation architecture on PASCAL VOC 2012 dataset and the augmented dataset provided by [20]. Our architecture achieves performance of 76.80% on PASCAL VOC 2012 *val* set which outperforms the state-of-the-art MobileNetv2 based architecture DeepLabV3-Mv2 [21] with a considerable margin of 1.48%, as reported in Table 1.

The portrait dataset [8] consists of 1800 images with 1500 for training and 300 for validation. However, the *train* and *val* lists used by [8] for training and evaluating PortraitFCN+ are not publicly available. In addition, due to online data missing there are only 1695 valid images retrieved from the portrait dataset. Therefore we randomly split them into a 1500 image *train* dataset and 195 image *val* dataset to retain the same volume of training images. We firstly evaluate the person class predictions with our proposed model which is pre-trained on PASCAL VOC 2012 dataset (without fine-tuning on portrait dataset) on portrait *val* set. As reported in Table 2, our method on person class reaches performance of 92.50%. It gains huge improvement of 19.41% compared to FCN (person class) 73.09%. We further finetune our network on the portrait *train* set without any additional data and post processing. The same data augmentations as in [8], *i.e.,* scales with $\{0.6, 0.8, 1.0, 1.2, 1.5\}$, ransom rotations in $\{-45, -22, 0, 22, 45\}$ and 5 gamma values in $\{0.5, 0.8, 1.0, 1.2, 1.5\}$, are applied during training. Our method achieves performance of 95.37% on the *val* set, and attains improvement of 1.17% compared with PortraitFCN. Note, our architecture achieves comparable results with PortraitFCN+ which is not end-to-end trainable and will totally fail when human face is not detected as it relies on prior knowledge of person provided by face detection. These limitations make PortraitFCN+ impractical for mobile applications where the computational resources are constrained and face detection frequently misses out people in natural photos. Some portrait segmentation results are visualized in Figure 7.

The inference time of our architecture on a 800×600 portrait image is ~ 42 milliseconds on a commodity PC with one Nvidia 1080Ti GPU. The corresponding portrait segmentation mobile application is developed on Nokia 8 (Snapdragon 835, Adreno 540 GPU). The inference time on a 225×225 image is ~ 1.67 second.

3.2.2. *Portrait Instance segmentation*

We report the AP_r at different IoU thresholds r and the AP (average from AP_{50} to AP_{95}) of our proposed non-parametric affinity modeling for portrait instance segmentation on our new dataset. Our proposed approach outperforms the state-of-the-art methods MR-CNN [22] and FCIS [23] by significant margins, particularly at high IoU thresholds, as listed in Table 3. Some visual comparisons with MR-CNN and FCIS are shown in Figure 8 which confirm the superior boundary snapping and instance-awareness of our results. On the contrary, MR-CNN produces poor boundaries on all evaluated images; FCIS sporadically misses person instances (2nd row) and is prone to producing unnatural straight-line-like boundaries. More comparisons are provided in the supplementary material.

4. CONCLUSION

We proposed a novel end-to-end portrait segmentation architecture with unique cross-granularity categorical attention and boundary enhancement mechanisms in a unified framework. It tackles the problems of semantic incoherence and poor boundary delineation by imposing semantic attention and enhancing boundary awareness across multi-granularity feature

Method\AP$_r$(%)	AP$_{50}$	AP$_{60}$	AP$_{70}$	AP$_{80}$	AP$_{90}$	AP
MR-CNN [22]	**97.8**	**97.8**	**97.8**	90.7	34.2	78.2
FCIS [23]	93.0	90.5	90.5	88.7	62.7	78.6
Our method	97.5	96.5	95.4	**92.0**	**67.8**	**85.1**

Table 3: Comparison of AP at various IoU thresholds for portrait instance segmentation with MR-CNN and FCIS on our new dataset.

hierarchy respectively. Moreover, we presented a simple yet effective non-parametric affinity model for portrait instance segmentation with enhanced instance-awareness. We have performed extensive evaluations of our approach and obtained state-of-the-art performance on challenging datasets, including our new multi-person portrait dataset with instance level annotations.

5. REFERENCES

[1] Tinghuai Wang, John P. Collomosse, Andrew Hunter, and Darryl Greig, "Learnable stroke models for example-based portrait painting," in *BMVC*, 2013.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[4] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Learning a discriminative feature network for semantic segmentation," *arXiv preprint arXiv:1804.09337*, 2018.

[5] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei, "Fully convolutional adaptation networks for semantic segmentation," in *CVPR*, 2018, pp. 6810–6818.

[6] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *arXiv preprint arXiv:1808.00897*, 2018.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[8] 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*. Wiley Online Library, 2016, vol. 35, pp. 93–102.

[9] Tinghuai Wang, Bo Han, and John P. Collomosse, "Touchcut: Fast image and video segmentation using single-touch interaction," *CVIU*, vol. 120, pp. 14–30, 2014.

[10] Tinghuai Wang, Huiling Wang, and Lixin Fan, "A weakly supervised geodesic level set framework for interactive image segmentation," *Neurocomputing*, vol. 168, pp. 55–64, 2015.

[11] Tinghuai Wang, Huiling Wang, and Lixin Fan, "Robust interactive image segmentation with weak supervision for mobile touch screen devices," in *ICME*, 2015, pp. 1–6.

[12] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," *arXiv preprint arXiv:1712.04837*, 2017.

[13] Qizhu Li, Anurag Arnab, and Philip HS Torr, "Holistic, instance-level human parsing," *arXiv preprint arXiv:1709.03612*, 2017.

[14] Huiling Wang, Tinghuai Wang, Ke Chen, and Joni-Kristian Kämäräinen, "Cross-granularity graph inference for semantic video object segmentation," in *IJCAI*, 2017, pp. 4544–4550.

[15] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao, "Weakly supervised instance segmentation using class peak response," *arXiv preprint arXiv:1804.00880*, 2018.

[16] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018, pp. 8759–8768.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[20] Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011, pp. 991–998.

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*. IEEE, 2017, pp. 2980–2988.

[23] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei, "Fully convolutional instance-aware semantic segmentation," in *CVPR*, 2017, pp. 4438–4446.