BiEPNet: Bilateral Edge-perceiving Network for High-Resolution Human Parsing

Authors: Qiqi Gong, Yao Zhao, Yunchao Wei

Date: 2023-12-05T00:00:00+00:00

Abstract

Human parsing is a fundamental task aimed at segmenting human images into distinct body parts and holds vast potential applications. Nowadays, the advancement of image-capturing devices has led to a growing number of high-resolution human images. Receptive field, details loss and memory usage are a triplet of contradictions in high-resolution scenarios. Existing human parsing methods designed for low-resolution inputs struggle to process high-resolution images efficiently due to their massive demands for computation and memory. Some methods save resources by overwhelmingly downsampling or encoding high-resolution inputs at the cost of poor performance on details. To resolve the issues above, we propose the Bilateral Edge-Perceiving Network (BiEPNet), consisting of a resources-friendly semantic-perceiving branch to acquire sufficient global information and a simple yet effective edge-perceiving branch used to refine details. The attention mechanism is utilized to simultaneously enhance the perception of context and details, leading to better performance on the boundary regions. To verify the effectiveness of BiEPNet, we contribute a high-resolution human parsing dataset, Human4K, containing 4,000 images with more than five million pixels. Extensive experiments on Human4K demonstrate that our method outperforms state-of-the-art methods while maintaining memory efficiency.

Full Text

Preamble

BiEPNet: Bilateral Edge-perceiving Network for High-Resolution Human Parsing

Qiqi Gong*, Yao Zhao†, Yunchao Wei‡
Institute of Information Science, Beijing Jiaotong University
Beijing Key Laboratory of Advanced Information Science and Network

ABSTRACT

Human parsing is a fundamental task that aims to segment human images into distinct body parts and holds vast potential for applications. With the advancement of image-capturing devices, high-resolution human images have become increasingly common. However, receptive field, detail loss, and memory usage form a triad of competing constraints in high-resolution scenarios. Existing human parsing methods designed for low-resolution inputs struggle to efficiently process high-resolution images due to their massive computational and memory demands. While some methods conserve resources by aggressively downsampling or encoding high-resolution inputs, they sacrifice detail preservation and achieve poor performance on fine-grained structures.

To address these issues, we propose the Bilateral Edge-Perceiving Network (BiEPNet), which consists of a resource-efficient semantic-perceiving branch to capture sufficient global information and a simple yet effective edge-perceiving branch to refine details. An attention mechanism is utilized to simultaneously enhance the perception of context and details, leading to improved performance in boundary regions. To verify the effectiveness of BiEPNet, we introduce Human4K, a high-resolution human parsing dataset containing 4,000 images with over five million pixels. Extensive experiments on Human4K demonstrate that our method outperforms state-of-the-art methods while maintaining memory efficiency.

Index Terms: Human-centered computing—Human Parsing—High-Resolution

1 INTRODUCTION

Human parsing is a fine-grained semantic segmentation task that aims to assign pixel-level labels to input human images [**?**]. As a core component of human understanding tasks, human parsing has drawn significant interest from researchers in the deep learning era and has found widespread applications in fashion, e-commerce, safety, image editing, and beyond.

Despite achieving great success, existing human parsing methods are universally trained and validated on images with approximately 100,000 pixels. However, with the advancement of image-capturing devices, high-resolution human images have proliferated on social networks. Most modern cameras and displays surpass 1080P resolution (1920 × 1080, 2 million pixels), while commercial devices have been upgraded to 4K UHD (3840 × 2160, 8 million pixels) [**?**]. Under these conditions, existing human parsing methods can hardly generalize to high-resolution scenarios due to memory and computational limitations.

DeepLabv3+ [?] is one of the most impactful semantic segmentation models, widely adopted in various tasks [?, ?, ?]. Building upon the atrous spatial pyramid pooling (ASPP) module proposed in [?], DeepLabv3+ introduced an encoder-decoder structure that incorporates low-level features extracted by the backbone into the decoding process. Based on empirical studies, we find that while this skip-layer design helps the model discriminate foreground from background effectively, details within the subject are often ignored. Additionally, the sharp increase in GPU memory usage caused by shortcuts [?] is intolerable in high-resolution scenarios.

Downsampling and sliding windows represent two intuitive solutions to this dilemma. As shown in Fig. 1, downsampling-based methods first reduce high-resolution images to low resolution, then directly upsample the low-resolution predictions to obtain high-resolution segmentation maps. Sliding-window-based methods crop the entire image into several patches and make predictions for each independently. However, both approaches suffer from critical drawbacks. Simple downsampling of high-resolution images leads to detail loss, particularly for small-scale objects, resulting in performance degradation. Sliding-window methods feed the network with only a patch of the whole image, saving memory at the cost of contextual consistency.

Recently, several resource-efficient segmentation models have been proposed, inspiring research on high-resolution image processing. Among these, dual sub-network architectures are particularly compelling. BiSeNet [?] was initially designed for real-time semantic segmentation, but its excellent trade-off between memory usage and performance makes it an ideal baseline for high-resolution semantic segmentation. Building upon bilateral path design, ISDNet [?] combines a deep segmentation network (where high-resolution images are downsampled) with a shallow network, gradually fusing features to generate final high-resolution outputs. Similarly, PGNet [?] introduces a pyramid grafting network to fuse features from deep convolutional networks and transformer architectures for high-resolution salient object detection.

We argue that these methods are sub-optimal. The context path in BiSeNet overwhelmingly encodes the input to 1/32 of its initial size, failing to maintain details. Meanwhile, the "DeepNet+ShallowNet" structures in ISDNet and PGNet can confuse the network during feature fusion due to the independent semantics of the two branches. Additionally, simple downsampling for computational reduction in the deep network reintroduces detail loss.

To tackle these problems, we propose a novel Bilateral Edge-Perceiving Network (BiEPNet, illustrated in Fig. 1(c)) for high-resolution human parsing that can be trained and validated end-to-end. Our BiEPNet comprises a memory-efficient semantic-perceiving branch to provide abundant global context information and an efficient edge-perceiving branch to ensure focus on details. Unlike typical bilateral structures that heavily downsample or encode high-resolution input images, we preserve resolutions for both input images and feature maps, striking a balance among receptive field, details, and memory usage. Moreover,

our edge-perceiving branch effectively learns representations of high-frequency regions, thereby augmenting attention to details and yielding significant improvements in overall high-resolution human parsing. Compared to generic semantic segmentation methods, we consume less GPU memory while achieving better performance. To verify the effectiveness of BiEPNet, we contribute a high-resolution, well-annotated human parsing dataset called Human4K and conduct comprehensive experiments on it.

The contributions of this paper can be summarized as follows:

- We propose a resource-efficient yet effective framework called Bilateral Edge-Perceiving Network (BiEPNet) that directly processes high-resolution human images while maintaining both sufficient receptive field and attention to details.
- We contribute Human4K, a high-resolution human parsing dataset including 4,000 human images with over 5 million pixels and accurate annotations.
- Extensive experiments demonstrate that our method achieves superior performance compared to existing methods while consuming fewer memory resources during inference.

2.1 Semantic Segmentation

Semantic segmentation is a significant task in computer vision that has achieved remarkable success. Deep convolutional neural networks have enabled the development of FCN-based [**?**] semantic segmentation models, which have been tested on various benchmarks and yielded impressive results [**?**, **?**, **?**]. DeepLabv3+ [**?**] utilizes an atrous spatial pyramid pooling module with an encoder-decoder structure to acquire sufficient information from multiple levels. PSPNet [**?**] introduces a pyramid pooling module to capture multi-scale context. Despite achieving extraordinary performance, these methods require massive spatial and temporal computational costs due to their architectures. The linear increase in computational cost with input size causes these architectures to struggle with effectively processing high-resolution images.

To address the problems raised by heavyweight models, several lightweight yet competitive segmentation models have been proposed. BiSeNetV1 [**?**] introduces a bilateral structure composed of a context path and a spatial path for high-level information and low-level details, respectively. Building upon BiSeNet, STDC [**?**] presents a lightweight backbone to achieve faster speed and higher accuracy. However, these methods do not robustly handle high-resolution segmentation. In contrast, we propose a novel high-resolution human parsing framework that incorporates a semantic-perceiving branch and an edge-perceiving branch, yielding competitive or even superior results compared to both lightweight and heavyweight models.

2.2 Human Parsing

Human parsing has experienced significant development in the deep learning era. Context learning has emerged as a prominent paradigm for human parsing. Chen et al. [**?**] exploited the attention mechanism for modeling human structures, and subsequently Chen et al. [**?**] recalibrated regional feature weights for segmentation results. Wang et al. [**?**] designed a parallel-connected network to aggregate features at different resolutions for stronger context information. Other studies have attempted to use auxiliary information as additional supervision to refine human parsing results. Ruan et al. [**?**] proposed CE2P, leveraging global context information and edge details to benefit human parsing. Due to its superb performance, CE2P has been widely used as a baseline for subsequent works. Zhang et al. [**?**, **?**] introduced pose information as auxiliary supervision to further help the model understand relationships among human parts. SCHP [**?**] utilized self-correction methods to reduce annotation noise in LIP [**?**], achieving 1st place in the CVPR2019 LIP Challenge. CDGNet [**?**] suggested that human parts are highly related to their positions, generating horizontal and vertical class distribution labels as supervision signals.

2.3 High-Resolution Image Processing

While downstream tasks in computer vision have been extensively researched, most focus has been on low-resolution input and output, leaving high-resolution scenarios insufficiently explored. CascadePSP [**?**] is one of the earliest works to study high-resolution image processing. It is a model-agnostic structure that continuously refines output from the previous stage. Additionally, some works [**?**, **?**, **?**] have attempted to refine boundary performance to promote overall accuracy. Recently, PGNet [**?**] introduced a framework fusing features from both CNN and Transformer to leverage their advantages for high-resolution salient object detection. Although these works have shown promising results in binary classification tasks, the dynamics change when dealing with multi-class classification, primarily due to inter-class conflicts.

The global-local framework is a popular paradigm for solving high-resolution segmentation. GLNet [**?**] combines context information from the global branch with details from the local branch to improve segmentation results. Similarly, PPN [**?**] proposed a classification branch to select important patches and fuse them with global images. MagNet [**?**] progressively refines rough results using multi-scale processing. ISDNet [**?**] proposed a "DeepNet+ShallowNet" structure to learn information from different scales. However, these solutions involve redundant computation and slow prediction speed, and the information interaction between branches can confuse the network. In contrast, we endow our

network with independent abilities that contribute to segmentation results in a resource-efficient manner.

3 HIGH-RESOLUTION HUMAN PARSING DATASET

Existing Human Parsing Datasets. Current publicly available human parsing datasets are relatively scarce compared to generic semantic segmentation datasets. ATR [**?**] and LIP [**?**] are two standard datasets commonly used in human parsing research. The LIP dataset comprises 50,462 images divided into 30,462/10,000/10,000 for training/validation/testing and labeled with 20 semantic classes. In contrast, the ATR dataset consists of 17,700 images distributed across 18 categories. While these datasets are large in scale, they typically exhibit low resolution, with most images containing at most hundreds of thousands of pixels. Moreover, they suffer from several drawbacks that impede representation learning (see Fig. 2). First, labels are often noisy. Second, humans displayed in some images are incomplete or fragmented. Third, there exists a mismatch between images and annotations where the number of objects differs from the number of annotations.

Human4K: A High-Resolution Human Parsing Dataset. To bridge the gap between high-resolution scenarios and human parsing, we contribute Human4K, a high-resolution human parsing dataset. This dataset was initially collected by Maadaa Data and includes over 10,000 images. We selected 4,000 images to form the Human4K dataset. Each image contains at least 5 million pixels, approaching 4K resolution, and some images exceed 20 million pixels, meeting the standards of contemporary commercial cameras. We follow the category definitions from LIP for the first 20 categories and add two additional categories: Torso-skin and Else (initially labeled as necklace, bracelet, etc.), resulting in 22 categories total. Images are randomly split into 2,500/500/1,000 for training/validation/testing.

Unlike LIP, where person instances are cropped from Microsoft COCO [**?**], all images in Human4K are captured in real-world scenes with varying complexity. Each sample in Human4K strictly displays only one person, avoiding mismatches between human instances and annotations. To make our network more human-centric, nearly all images include the Face category, while exhibiting diverse appearances, postures, and viewpoints. Our Human4K addresses the aforementioned drawbacks of existing datasets.

4.1 Architecture Overview

Based on Human4K, we propose Bilateral Edge-Perceiving Network (BiEPNet). As shown in Fig. 3 [Figure 3: see original paper], BiEPNet comprises a semantic-perceiving branch and an edge-perceiving branch. The two branches are de-

signed with distinct objectives but collaborate closely to contribute to the parsing result. The semantic-perceiving branch aims to provide sufficient semantic and context information using a segmentation network (Section 4.2), while the edge-perceiving branch enhances attention to boundary regions (Section 4.3). Information from the two branches is fused using an attention-based method to produce the final result. Unlike ISDNet [**?**], we directly feed both branches with high-resolution input images without downsampling.

4.2 Semantic-perceiving Branch

Our semantic-perceiving branch serves as a strong global context information perceptor, following the design of DeepLabv3 [**?**]. Given an image $I \in \mathbb{R}^{3\times H\times W}$, where H and W are the height and width respectively, the backbone outputs a feature map from the last block denoted as $f \in \mathbb{R}^{C\times h\times w}$ where $(h, w) = (\frac{H}{8}, \frac{W}{8})$. The feature f is then decoded with an ASPP head [**?**] to obtain strong semantic information $f_{sp} \in \mathbb{R}^{C_{sp}\times h\times w}$. An FCN [**?**] head can be added on top of the semantic-perceiving branch to acquire the coarse segmentation map $S_{deep} \in \mathbb{R}^{cls\times h\times w}$, where cls is the number of categories.

Compared with the context branch in BiSeNet [**?**] and the deep net in ISDNet [**?**], our semantic-perceiving branch captures abundant semantic information without compromising dense prediction requirements. BiSeNet overwhelmingly encodes the input image with an output stride of 32 (ours is 8), while ISDNet directly downsamples the input image into $I \in \mathbb{R}^{3\times \frac{H}{4}\times \frac{W}{4}}$ to preserve memory, degrading details in high-resolution input images.

4.3 Edge-perceiving Branch

Prediction quality on boundary regions is one of the most distinctive challenges in human parsing [**?**, **?**, **?**, **?**]. Compared to complex semantic information, edge information is relatively low-level. Traditional operators [**?**, **?**] can obtain coarse edge maps from RGB images. In the deep learning era, these operators are simulated using non-learnable convolutional operators [**?**, **?**]. We argue that edge information can be effectively perceived with just a few convolutional layers, and leveraging this information can significantly enhance overall parsing results.

Based on empirical studies, we observe that low-level features in DeepLabv3+ [**?**] struggle to recognize details within effectively discriminated foreground and background regions (see Fig. 4 Figure 4: see original paper). In BiSeNet's spatial path [**?**], which consists of several convolutional layers, edge parts are perceived but make weak contributions to subsequent processing (see Fig. 4(b)). To enhance the utilization of boundary information for refining parsing results, we employ several convolution layers as a boundary extractor and reinforce

the features with a self-attention mechanism, thereby constructing our edge-perceiving branch. Let features extracted by consecutive convolution layers be denoted as $f_s \in \mathbb{R}^{C_s \times h \times w}$.

To relate detailed edge information with other spatial positions both near and far, we utilize the interlaced sparse self-attention mechanism [**?**] on f_s to finally acquire features with strong edge information $f_{ep} \in \mathbb{R}^{C_{ep} \times h \times w}$ (see Fig. 4(c)). Unlike [**?**, **?**] where edge detector parameters are fixed, all parameters in our edge-perceiving branch are learnable.

Considering that features from the semantic-perceiving branch f_{sp} and edge-perceiving branch f_{ep} are responsible for perceiving different information and contribute unequally to parsing results, we follow [**?**] to fuse them using a channel-wise attention method that exploits the relationship between semantic information and edge details. Finally, the predicted human parsing result is produced by a standard segmentation head.

4.4 Loss Functions

Human Parsing Loss. Standard cross-entropy loss is used for the final parsing results (L_{hp}) as well as auxiliary supervision on the semantic-perceiving branch S_{deep} (L_{aux}).

Edge Perceiving Loss. Following [**?**], we adopt a weighted binary cross-entropy loss (L_{ep}) to supervise training of the edge-perceiving branch. L_{ep} is defined as follows:

$$L_{ep} = - \sum_{\{1:b_i>t\}} \left(s_{i,c} \log \hat{s}_{i,c} \right)$$

where t is a predefined threshold, b_i is the output of the edge prediction head, and $s_{i,c}$ and $\hat{s}_{i,c}$ are the ground-truth and prediction result of the i-th pixel for class c, where $c \in (0, 1)$. Ground-truth edge maps are generated following [**?**].

Fusion Loss. Features from the semantic-perceiving branch and edge-perceiving branch are fused and enhanced by attention. Auxiliary supervision is imposed on these enhanced features, deriving the fusion loss L_{fuse}:

$$L_{fuse} = L_{hp}^{fuse} + L_{ep}^{fuse}$$

Total Loss. The total loss L is a weighted combination of the losses mentioned above:

$$L = \lambda_1 L_{hp} + \lambda_2 L_{ep} + \lambda_3 L_{fuse}$$

Note that all prediction heads except the final segmentation head only operate during the training phase.

5.1 Implementation Details

Our method comprises a semantic-perceiving branch and an edge-perceiving branch. The semantic-perceiving branch employs DeepLabv3 [**?**] with ResNet-18 [**?**], while the edge-perceiving branch consists of four convolution blocks by default. ResNet-18 parameters are initialized using a pretrained ImageNet model [**?**]. All loss weights are set to 1.0 in Equation 3.

We use the mmsegmentation framework [**?**] as our codebase. Data augmentations for human parsing differ slightly from generic semantic segmentation. Considering memory usage during training and the continuity of human body structures, we resize the longest edge of an input human image to 1024 while preserving its aspect ratio. Next, we randomly scale the image with a scaling factor varied in (0.5, 2). Furthermore, we crop a 1024×1024 patch from the scaled image, balancing human body continuity and training data abundance. Notably, distinguishing left/right pairs in human images is a special challenge; label pairs should be swapped if horizontal flipping occurs. Images are padded to 1024×1024 during inference.

BiEPNet parameters are optimized using SGD with momentum 0.9 and weight decay 5e-4. The initial learning rate is set to 0.01, and we apply a warmup strategy for the first 1K iterations with a warmup factor of 0.1. The learning rate is then adjusted using a polynomial decay with parameter 0.9, with maximum iterations set to 80K. Experiments are conducted with a batch size of 8 for training on 4 NVIDIA GeForce RTX 3080Ti GPUs. Memory usage is measured with batch size 1 on an RTX 3080Ti GPU.

5.2 Experiments on Human4K

We evaluate our framework on Human4K. First, we compare BiEP with several reproduced generic semantic segmentation methods on Human4K. In addition to the standard evaluation metric mean Intersection over Union (mIoU), we focus on boundary performance—a key challenge in human parsing—by averaging boundary IoU [**?**] for each class to obtain mean boundary IoU (mBIoU). For fair comparison, all methods in Table 1 (except STDC [**?**]) use ResNet-18 [**?**] as the backbone with input size 1024×1024.

Compared to the approaches listed in Table 1, our method not only achieves the highest mIoU but also performs best on boundary regions. Based on quantitative results, we draw two conclusions: (1) Benefiting from the edge-perceiving branch design, BiEPNet outperforms others on challenging categories such as

'Socks', 'Left/Right pairs', and especially 'Scarf', where the improvement is overwhelming; (2) The semantic-perceiving branch yields competitive understanding for common categories like 'Hair' and 'Pants'. We also report average memory consumption during inference for each method to demonstrate our efficiency. CE2P is our main competitor in terms of performance, but it consumes 41.17% more memory than our method due to redundant usage of low-level features from the backbone.

Fig. 5 [Figure 5: see original paper] presents qualitative results, primarily comparing our method with the baseline DeepLabv3+ and state-of-the-art ISDNet. DeepLabv3+ performs well at locating the subject in the image, but pixels within semantic regions are often misclassified (manifested as spots within regions). Suffering from extreme downsampling in the deep branch, ISDNet exhibits poor performance on details, leading to blurred boundary recognition. Both quantitative and qualitative experiments demonstrate that BiEP successfully balances memory consumption and performance.

5.3 Ablation Studies

To illustrate the effectiveness of each module in BiEPNet, we conduct comprehensive ablation studies focusing primarily on our edge-perceiving branch.

Ablations of Compositions. Our baseline method uses the naive semantic-perceiving branch with DeepLabv3 and ResNet-18, with channel-wise attention as the default fusion method. Table 2 shows that edge details provided by the edge-perceiving branch boost overall performance by a large margin, and the self-attention mechanism adopted on top of the edge-perceiving branch further improves performance.

Ablations of Convolution Layer Count. Table 3 presents results with different numbers of convolution layers in the edge-perceiving branch. When the number is fewer than three, large feature maps cause out-of-memory errors. The third line demonstrates that more convolution layers can degrade overall performance, so we use four layers by default.

Ablations of Fusion Methods. Similar to [**?**], we compare different fusion methods in Table 4 . ADD directly adds same-sized features from the semantic-perceiving and edge-perceiving branches and passes them through a convolution block, while CAT concatenates the two features and passes them through a convolution block (requiring twice the input channels of ADD). Comparison results demonstrate the effectiveness of the channel-wise attention fusion method.

---

6 CONCLUSION AND LIMITATIONS

This paper presents a high-resolution human parsing framework consisting of a semantic-perceiving branch and an edge-perceiving branch, endowing the network with superb abilities to process both global and local information simultaneously. The two branches cooperate to produce final parsing results through a channel-wise attention mechanism. Our memory-friendly method surpasses several state-of-the-art methods on our contributed Human4K dataset.

However, due to category imbalance, we notice that training results can vary with random seeds, and this problem is not consistently resolved in this paper. Systematic exploration of more robust architectures and methods for preprocessing human images warrants further research.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (No. 2021ZD0112100), National NSF of China (No. U1936212, No. 62120106009).

REFERENCES

[1] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986.
[2] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.
[3] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2016.
[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pp. 801–818, 2018.
[5] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8924–8933, 2019.
[6] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi. Spgnet: Semantic prediction guidance for scene parsing. In *IEEE International Conference on Computer Vision*, pp. 5218–5228, 2019.
[7] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.

[8] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8890–8899, 2020.
[9] M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.
[11] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. Rethinking bisenet for real-time semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725, 2021.
[12] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 932–940, 2017.
[13] S. Guo, L. Liu, Z. Gan, Y. Wang, W. Zhang, C. Wang, G. Jiang, W. Zhang, R. Yi, and L. Ma. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4361–4370, 2022.
[14] H. Harkat, J. Nascimento, and A. Bernardino. Fire segmentation using a deeplabv3+ architecture. In *Image and Signal Processing for Remote Sensing XXVI*, vol. 11533, pp. 134–145. SPIE, 2020.
[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision*, pp. 770–778, 2016.
[16] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang. Interlaced sparse self-attention for semantic segmentation. arXiv preprint arXiv:1907.12273, 2019.
[17] C. Huynh, A. T. Tran, K. Luu, and M. Hoai. Progressive semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16755–16764, 2021.
[18] N. Kanopoulos, N. Vasanthavada, and R. L. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988.
[19] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.
[20] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.
[21] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2402–2414, 2015.
[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P.

Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.
[23] K. Liu, O. Choi, J. Wang, and W. Hwang. Cdgnet: Class distribution guided network for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4473–4482, 2022.
[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.
[25] H. Polat. Multi-task semantic segmentation of ct images for covid-19 infections using deeplabv3+ based on dilated residual network. *Physical and Engineering Sciences in Medicine*, 45(2):443–455, 2022.
[26] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.
[27] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4814–4821, 2019.
[28] S. A. Taghanaki, A. Bentaieb, A. Sharma, S. K. Zhou, Y. Zheng, B. Georgescu, P. Sharma, Z. Xu, D. Comaniciu, and G. Hamarneh. Select, attend, and transfer: Light, learnable skip connections. In *Machine Learning in Medical Imaging*, pp. 417–425. Springer, 2019.
[29] C. Wang, P. Du, H. Wu, J. Li, C. Zhao, and H. Zhu. A cucumber leaf disease severity classification method based on the fusion of deeplabv3+ and u-net. *Computers and Electronics in Agriculture*, 189:106373, 2021.
[30] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, and X. Wang. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.
[31] T. Wu, Z. Lei, B. Lin, C. Li, Y. Qu, and Y. Xie. Patch proposal network for fast semantic segmentation of high-resolution images. In *AAAI Conference on Artificial Intelligence*, pp. 12402–12409, 2020.
[32] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li. Pyramid grafting network for one-stage high resolution saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11717–11726, 2022.
[33] J. Xu, Z. Xiong, and S. P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired from pid controller, 2022.
[34] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision*, pp. 325–341, 2018.
[35] Y. Yuan, J. Xie, X. Chen, and J. Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pp. 489–506, 2020.
[36] Z. Zhang, C. Su, L. Zheng, and X. Xie. Correlating edge, pose with parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8900–8909, 2020.

[37] Z. Zhang, C. Su, L. Zheng, X. Xie, and Y. Li. On the correlation among edge, pose and parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8492–8507, 2021.
[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.
[39] P. Zhou, B. Price, S. Cohen, G. Wilensky, and L. S. Davis. Deep-strip: High-resolution boundary refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Note: Figure translations are in progress. See original paper for figures.