GRPSNET: MULTI-CLASS PART PARSING BASED ON GRAPH REASONING

Njuod Alsudays *Jing Wu* *Yu-Kun Lai* *Ze Ji*

Cardiff University, UK
{alsudaysn, wuj11, laiy4, jiz1}@cardiff.ac.uk

ABSTRACT

Multi-class part parsing is a dense prediction task that decomposes objects into semantic components with multi-level abstractions. Despite the importance of this problem, it remains challenging due to the presence of both part-level and class-level ambiguities. In this paper, we propose GRPSNet network which integrates graph reasoning to capture relationships between parts for part segmentation. These captured relationships help to enhance the recognition and localization of parts. We also propose to exploit the relationships of part boundaries to further enhance the accuracy of part segmentation. The experimental results demonstrate the effectiveness of the proposed method and show that it achieves state-of-the-art performance on the benchmark datasets.

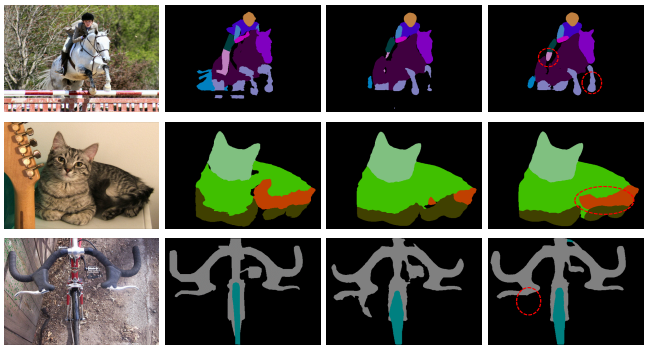
Index Terms— Part parsing, Semantic segmentation, Graph Reasoning, Deep Learning.

1. INTRODUCTION

Semantic part parsing aims to simultaneously detect multiple object classes in the scene and accurately segment the parts within each object class. Understanding parts within each object class is important for many fine-grained tasks, such as object detection [1], fine-grained action detection [2], pose estimation [3], and categorization [4, 5]. However, multi-class part parsing has been considered only recently [6, 7, 8, 9, 10]. It is a challenging task that needs to tackle both part-level ambiguity and object-level ambiguity.

In particular, one challenge is that some parts occupy a small area in the scene, making them difficult to detect. For example, as shown in Fig. 1 (c), even the state-of-the-art method [10] struggles to accurately detect the human and horse legs and the cat tail. Another challenge is inaccurate boundary detection due to the clutter of several classes in the scene, which often leads to occlusions. An example is the segmentation of the bike body, as shown in Fig. 1.

Graph reasoning to infer relational information has proven to be useful for many computer vision tasks. Recently, graph-based methods have achieved considerable improvements by modeling the relationships between distant regions and reasoning using global information [11, 12, 13]. This global perspective can improve the ability of a model to handle complex scenes, where local information alone might not be sufficient. Inspired by the achievements, in this paper, we propose to integrate graph reasoning into part parsing to tackle the above challenges. We construct two graphs, namely, the part graph and the boundary graph. The part graph aims to extract a broad range of relationships such as spatial and appearance between entire parts to enhance the recognition and localization of parts in the scene, while the boundary graph focuses on extracting adjacency information, such as edges and transitions between parts, to enhance the detection and understanding of part boundaries.



(a) Input (b) GT (c) AFPSNet (d) Ours

Fig. 1. Challenges of multi-class part parsing. (a) Input images. (b) Ground truth masks. (c) Results from the current state-of-the-art method [10]. (d) Results from our method.

In addition to graph reasoning, we adopt a multi-task framework, which includes part awareness, boundary awareness, and object awareness branches. Part awareness integrates the part graph to capture relationships between parts. Boundary awareness enforces boundary constraints to refine the boundaries of detected parts. We also integrate the boundary graph to help the model better detect the adjacency of parts in the image. Moreover, as object-level ambiguity is also a challenge, we propose the object awareness branch to improve the localization of parts within each object class.

The contributions of this work are as follows:

- We propose to integrate part and boundary graphs to extract part relationships with the goal of reducing ambiguities in part parsing.

- We propose a multi-task framework to perform boundary and object awareness as auxiliary tasks to improve the part localization within each object class.
- The proposed method achieves state-of-the-art performance on the benchmark part parsing datasets.

2. RELATED WORK

2.1. Part Parsing

Existing research in the recent literature has shown effective performance in accurate segmentation of parts of one specific category, such as human bodies [14, 15], vehicles [16, 17] and animals [18, 19]. However, existing models mainly assume a single object of interest per image, where the object is well-localized beforehand and with no occlusions. In contrast, detecting multiple object classes in the scene and simultaneously parsing the parts within each class is a relatively new problem that has been explored only recently [6, 7, 8, 9, 10]. Zhao *et al.* [6] proposed a joint boundary-semantic awareness framework to enhance the part localization and the expression of class-relevant feature channels. Michieli *et al.* [7] proposed a framework consisting of three subnetworks with a graph-matching module. Object-level segmentation maps are used as guidance for part-parsing within the object, while the graph matching technique is used to preserve the relative spatial relationships between the predicted parts and ground truth. Michieli *et al.* [9] improved their framework by integrating edge information to enhance edge localization. Tan *et al.* [8] introduced a framework with a confident semantic ranking loss function to model the pixel relationships among intra and inter parts. Alsudays *et al.* [10] proposed a framework based on scaled attention and feature fusion to capture more part details from finer scales and to effectively fuse different scales of features. Despite the remarkable performance of these methods, relation reasoning has not been explored in multi-class part parsing. Therefore, we propose to integrate graph reasoning into the part segmentation task to enhance the recognition and localization of parts in the scene.

2.2. Graph Reasoning

In recent years, graph-based methods [20, 11, 12, 13] have been widely used for relational reasoning. Kipf *et al.* [20] proposed a graph convolution network that introduces the basis of feature embedding on graph-structured data for semi-supervised classification. Later studies [11, 12, 13] introduced graph reasoning for visual recognition tasks due to its ability to capture global information in graph propagation. However, the approaches of multi-class part parsing lack the investigation of the relationships among part regions. Therefore, we propose to integrate graph reasoning to capture and model these relationships for improved part segmentation.

3. METHOD

3.1. Overview

GRPSNet uses DeepLab v3+ [21] as a backbone. The same encoder as in [21] is used as a shared layer for the subsequent three branches. The object and part awareness branches then follow the typical design of AFPSNet [10], which consists of Atrous Spatial Pyramid Pooling (ASPP), Attention Refinement Module (ARM) and Feature Fusion Module (FFM) to help the model selectively emphasize important contextual information at different scales. Finally, graph reasoning units (GRU) are built in the part and boundary branches to enable the reasoning of part relationships at part-level and boundary-level, as shown in Fig. 2.

3.2. Graph Reasoning Unit (GRU)

In multi-class part parsing methods, extracted regions are analyzed separately without considering the dependencies between them, leading to limited performance when handling part-level ambiguities, occlusions, and tiny parts. Therefore, we propose to integrate graph reasoning to extract relationships between regions and to enhance the features of these regions with the extracted relationships. To achieve that, we construct two graphs, i.e., part graph and boundary graph. Part graph performs graph reasoning at part level, which aims to extract broad relational information between parts, such as their spatial layouts and appearance similarity. Boundary graph performs graph reasoning at the boundary level. It has a more specific focus on the adjacency between parts.

In these two graphs, a Graph Convolution Network (GCN) [20, 22] is employed as in [12] to refine the model understanding of relationships within a graph by adjusting edge weights in the adjacency matrix during training. GCN is applied via two $1D$ convolution layers along different directions, i.e., node-wise and channel-wise, as shown in Fig. 3. In the following, we describe the construction of the two graphs.

Construction of the part graph. We build the part graph $G_p = (V_p, E_p, A_p)$ as adapted from [12] to extract the part relationship information from the feature map $F \in \mathbb{R}^{H \times W \times C}$, which is the output of the shared encoder. The part graph G_p consists of a set of nodes V_p and edges E_p. A_p denotes the adjacency matrix. In the part graph, the input feature map F is divided into meaningful and distinct regions. Each node in V_p represents one of these regions, and the edges between nodes denote relationships between the corresponding regions.

Practically, the input feature maps are reshaped by using two convolutional layers to reduce the input dimension and obtain the node features F_{V_p}. That is, as Fig. 3 shows, the feature map F is passed through a 1×1 convolutional layer to generate the node mask $B \in \mathbb{R}^{H \times W \times N}$. The input feature map F is also passed through another 1×1 convolutional layer to reduce the channel number size of F and generate $F\prime \in \mathbb{R}^{H \times W \times C\prime}$. Then, a channel-wise multiplication of $F\prime$

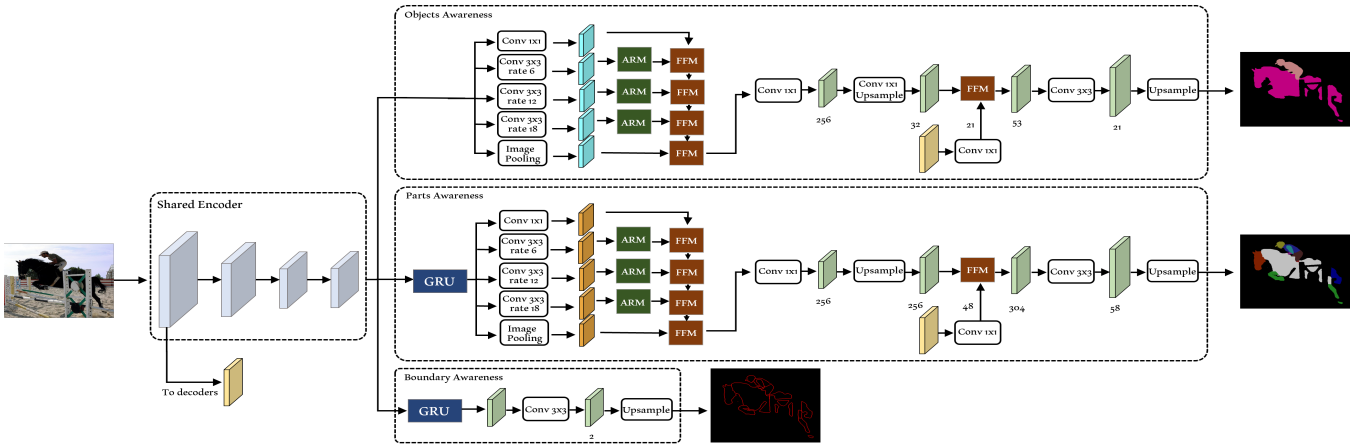


Fig. 2. An overview of the proposed GRPSNet framework, which consists of three branches to perform multiple tasks simultaneously. Two Graph Reasoning Unit (GRU) are integrated into the part and boundary awareness to capture the relationships between part classes. The object awareness aims to capture auxiliary object information and supervise the part parsing.

and B is performed to compute the feature map of nodes F_{V_p}. The node features are represented by $F_{V_p} = f(F) \in \mathbb{R}^{N \times C\prime}$, where $f(.)$ refers to the construction function that is performed to construct the feature of nodes from the feature map, N is the number of nodes extracted from F, and $C\prime$ is the reduced number of channels. A softmax function is performed along the $H \times W$ dimension of B. By applying the softmax function to the node mask B, we introduce a probabilistic element into the assignment of classes to nodes in the part graph. The softmax probabilities can influence the graph structure by guiding the connections and relationships between nodes. Edges between nodes may be influenced by the confidence or probability scores associated with the corresponding classes. This can be particularly useful in scenarios where parts may exhibit uncertainty or ambiguity in their classification, allowing for a more nuanced representation of the parts and their relationships in the constructed graph. After transforming the feature map F into a graph representation F_{V_p}, the relationships between nodesare computed through the Graph Convolution Network (GCN).

Construction of the boundary graph. We build the boundary graph $G_b = (V_b, E_b, A_b)$ from the same feature map $F \in \mathbb{R}^{H \times W \times C}$. The boundary graph consists of a set of nodes V_b and edges E_b. Each node in V_b represents a detected region from the input feature map F. These regions are extracted by performing a 1×1 convolution layer on the feature map F to generate a node mask B as in the part graph. A sigmoid function is applied to the extracted nodes in the node mask B to enhance the representation of boundary strength by providing probabilistic attributes. The sigmoid probabilities provide semantic information about the likelihood of nodes being part of boundaries. This information can be included as node attributes in the boundary graph and can influence the graph structure by guiding the connections and relationships between nodes. Edges between nodes may be influenced by the confidence or probability scores associated with each node being part of a boundary. This can be valuable in scenarios when dealing with ambiguous or uncertain boundaries, allowing for more flexible and nuanced modeling of the relationships between nodes and their association with boundaries. Different from the part graph, the edges in the boundary graph have binary values representing the presence or absence of boundaries between the detected regions in the node features map F_{V_b}. Edges connect nodes and represent relationships between detected adjacent boundaries. The presence of an edge indicates a spatial connection between two boundary points. The visual correlations between boundary nodes are captured by passing the node features F_{V_b} through the graph convolution network (GCN).

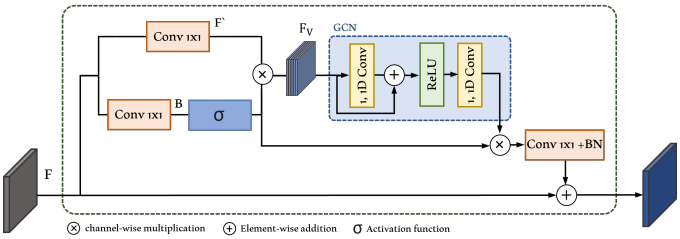


Fig. 3. The architecture of the Graph Reasoning Unit (GRU), which is adapted from [12]. We propose normalizing the constructed nodes by performing an activation function before passing them to the Graph Convolution Network (GCN).

3.3. Joint Parsing Framework

GRPSNet consists of four components: (1) an encoder shared by all three branches, (2) an object awareness branch, (3) a part awareness branch, and (4) a boundary awareness branch.

Shared encoder. We use the same encoder as in DeepLab v3+ [21], which is the ResNet-101 model [23] pre-trained on

ImageNet dataset [24].

Object awareness branch. The object segmentation branch is built based on AFPSNet [10] which is the state-of-the-art part segmentation network. As shown in Fig. 2, The object awareness branch takes the shared representation from the shared encoder and passes it through the Atrous Spatial Pyramid Pooling (ASPP) unit, which includes the attention module (ARM) and feature fusion module (FFM). Thus, object information is captured at different scales. The extracted object information is used to enhance the localization of parts within each object class. The cross-entropy loss is used for training this branch:

$$L_{object}(y,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\cdot\log(p_{i,c}), \qquad (1)$$

where $y_{i,c}$ is a one-hot encoding scheme of ground truth labels and $p_{i,c}$ is a matrix element corresponding to the predicted value for each object. Here, i is the sample index, c is the part index and C is the number of objects.

Part awareness branch. As shown in Fig. 2, this branch also follows the design of AFPSNet [10] with integration of part graph. The ASPP module in the model captures contextual information across various scales. The cross-entropy loss is again used for training part branch.

$$L_{part}(y,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\cdot\log(p_{i,c}), \qquad (2)$$

Boundary awareness branch. The boundary branch consists of boundary graph, followed by two convolution layers to obtain the boundary features, as shown in Fig. 2. A binary cross-entropy loss is used for training this branch.

$$L_{edge}(y,\widehat{y}) = -(y\log(\widehat{y}) + (1-y)\log(1-\widehat{y})), \qquad (3)$$

where $\widehat{y}$ refers to the predicted boundary map and y refers to the ground truth label.

GRPSNet is trained to perform multiple tasks simultaneously. The loss function is adjusted to determine how the model balances learning across these tasks. The final loss function L_{total} of our framework is a weighted sum of individual losses for each task,

$$L_{total} = \lambda_e\cdot L_{edge} + \lambda_o\cdot L_{object} + L_{part}, \qquad (4)$$

where $\lambda_{\{e,o\}}$ are weights to balance the boundary and object losses to control the contribution of each task to the overall training objective.

4. EXPERIMENTS

4.1. Implementation Details

Dataset. The widely used PASCAL-Part [4] and the large-scale ADE20K-Part [25] datasets are used to train and evaluate the proposed method. PASCAL-Part includes PASCAL-Part-58, PASCAL-Part-108, and PASCAL-Person-Part. Both PASCAL-Part-58 and PASCAL-Part-108 contain 10103 images of varying sizes, along with 58 (PASCAL-Part-58) or 108 (PASCAL-Part-108) part-level annotations of 21 semantic object classes, including the background class. We use 4998 images for training and 5105 images for testing as the original split in [4]. PASCAL-Person-Part contains 3533 images of multi-person on various scales and with 7 part-level annotations, including the background class. We use 1716 images for training and 1817 images for testing as the original split in [26, 27, 6]. ADE20K-Part dataset contains 22210 images of different sizes, along with 544 part-level annotations of 150 object- and stuff-level classes as in [7]. We use 20210 images for training and 2000 images for testing as the original split in [25]. Also, we follow the same evaluation metrics of the state-of-the-art and other well-known part parsing methods [6, 7, 8, 10] by using the mean Intersection over Union (mIoU); and applying the same evaluation strategy.

Training details. We employ the DeepLab v3+ [21] network trained on the ImageNet dataset [24] as a backbone and follow the same training schemes as in [21, 28]. For PASCAL-Part, we train a model with a batch size of 16 for $80K$ iterations. For ADE20K-Part, we train our model with a batch size 6 for $100K$ iterations. To make the weight balance and enhance the part segmentation, we set $\lambda_e = 0.10$ and $\lambda_o = 0.03$ for the first $50K$ iterations, then we adjust $\lambda_e = 0.20$ and $\lambda_o = 0.0$ for the last iterations. More training details can be found in the supplementary material.

4.2. Ablation Study

We first evaluate the effectiveness of integrating part and boundary graphs, the boundary branch and the object branch. The experiments were carried out on two frameworks, DeepLab v3+ and AFPSNet, to demonstrate the general benefit of integrating these components.

DeepLab v3+. The first row in Table 1 shows that without any of the proposed components, using DeepLab v3+ achieved 57.60% in mIoU for part parsing. Adding the boundary constraint improves the performance to 59.90. The further addition of part and boundary graphs increases the performance from 59.90 to 59.99. Moreover, adding the object awareness further improves the performance to 60.08.

AFPSNet. When using AFPSNet, the baseline method achieved a mIoU of 58.68% for part parsing. Adding the boundary constraint improves the mIoU to 60.02. Integrating the part and boundary graphs increases the performance from 60.27 to 61.28. The further addition of the object segmentation achieves the best performance with 61.59% mIoU.

4.3. Comparisons with state-of-the-art methods

We compared our method with DeepLab v3+ [21] and the state-of-the-art multi-class part parsing methods [6, 7, 8, 9, 10]. The segmentation performance of these methods is compared based on the PASCAL-Part-58 benchmark. To compare

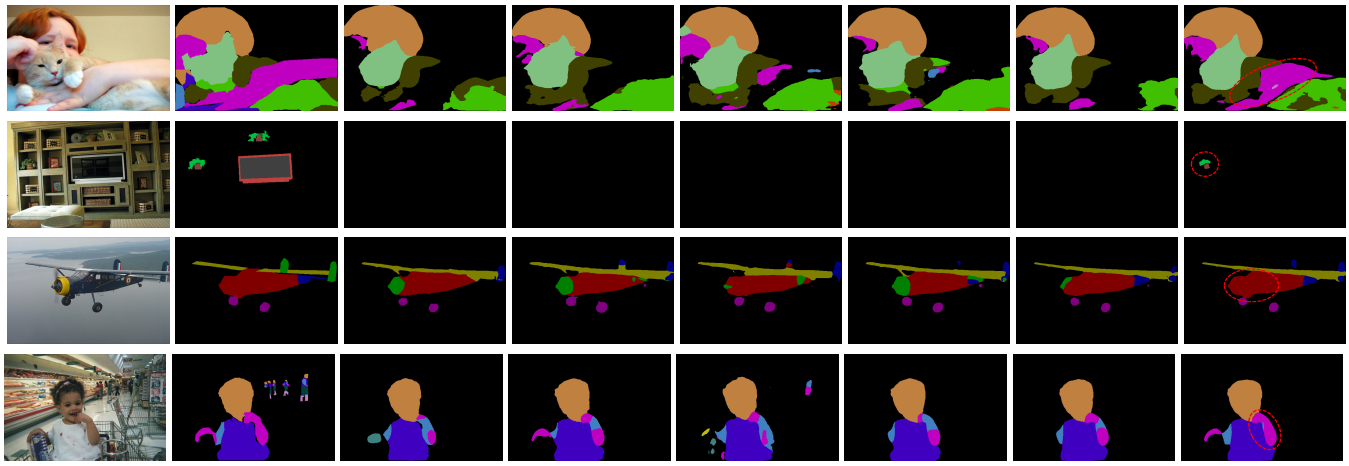


Fig. 4. Segmentation results on PASCAL-Part-58 dataset. Our model generates notable results by achieving better part localization and more accurate boundaries compared to the state-of-the-art models.

Table 1. Detailed performance comparison of each component in our proposed GRPSNet approach.

Method	Object branch	Boundary branch	Boundary&Part graphs	mIoU(%)
DeepLab v3+				57.60
DeepLab v3+		✓		59.90
DeepLab v3+		✓	✓	59.99
DeepLab v3+	✓	✓	✓	**60.08**
AFPSNet				58.68
AFPSNet		✓		60.27
AFPSNet		✓	✓	61.28
AFPSNet	✓	✓	✓	**61.59**

Table 2. Segmentation performance of mIoU on the benchmark datasets. mIoU: per-part-class mIoU. Avg.: average per-object-class mIoU.

Method	PASCAL-Part-58		PASCAL-Part-108		Pascal-Person-Par	ADE20K-Part	
	mIoU	Avg.	mIoU	Avg.	mIoU	mIoU	Avg.
DeepLab v3 [28]	54.4	55.9	41.3	43.7	63.5	8.9	17.6
DeepLab v3+ [21]	57.6	59.1	46.5	48.9	68.1	15.9	27.5
BSANet [6]	58.2	58.9	42.9	46.3	67.4	9.7	19.6
GMNet [7]	59.0	61.8	45.8	50.5	67.5	10.6	21.3
GMENet [9]	59.6	**62.2**	46.3	51.2	68.4	12.9	23.6
CSR [8]	60.7	60.6	-	-	-	-	-
AFPSNet [10]	61.3	62.0	49.2	51.2	69.6	18.2	29.1
GRPSNet	**61.6**	**62.2**	**50.5**	**52.4**	**70.1**	**18.7**	**29.4**

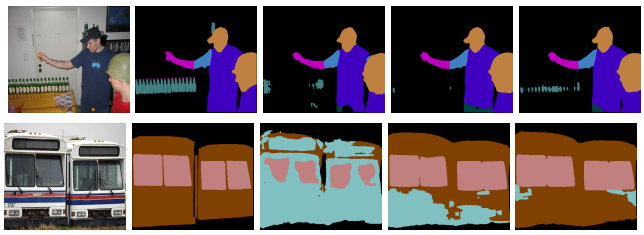
the performances of these methods quantitatively, two metrics were used: mIoU, which computes the mean per-part IoU on the 58 part classes, and Avg., which computes the average per-object mIoU on the 21 object classes (including background). Table 2 shows that DeepLab v3+ achieved 57.6% in per-part mIoU, while the first work addressing part-based semantic segmentation, BSANet, achieved 58.2%. GMNet improved its performance, achieving 59.0%. Then, GMENet enhanced the performance to 59.6%. While CSR improved the performance of part segmentation to 60.7%. The current state-of-the-art method, AFPSNet, achieved 61.3%. The proposed GRPSNet achieved the highest per-part mIoU of 61.6% compared to the above methods, outperforming the current state-of-the-art methods. The same is observed for the average per-object-class mIoU as well.

Table 2 shows the overall results of these methods on PASCAL-Part and ADE20K-Part benchmarks. Our method achieved the highest per-part mIoU for all the benchmarks, outperforming the state-of-the-art method.

Fig. 4 illustrates a qualitative comparison of the segmentation results from these methods. Our model shows overall better segmentation results in better localization of parts and more accurate boundaries. For example, the recognition of the plant in the second row is difficult to detect by the other methods, while GRPSNet can successfully localize and segment it. Moreover, GRPSNet shows superior performance in detecting parts even in a very crowded scene. For example, the human arm in the first row, and the girl's arms in the last row. Also, GRPSNet can better predict the boundaries of the airplane body in the third row. These segmentation results demonstrate that GRPSNet can effectively enhance the recognition and localization of parts. Further experimental results can be found in the supplementary material.

4.4. Failure cases

Our method does have limitations. As shown in the first row of Fig. 5, despite the ability of the model to detect the bottles, the model may be confused if several objects occupy a small region in the scene. One avenue for enhancement could be by further improving the boundary awareness branch. Additionally, our model might produce localized segmentation errors if the objects have similar appearances and their surroundings are not visible in the input image, such as the bus in the second row of Fig. 5, which will need further investigation.



Input GT DeepLabv3+ AFPSNet GRPSNet

Fig. 5. Two typical failure cases. Our model may be confused with several objects occupying a small region in the scene. The invisible surrounding of the object can sometimes confuse our model.

5. CONCLUSION

In this paper, we proposed GRPSNet, which utilizes part relationships and the multi-task framework to address the multi-class part parsing challenges. The ablation study demonstrates the effectiveness of our method, and the experiments show that GRPSNet achieves state-of-the-art performance on multi-class part parsing on both PASCAL-Part and ADE20K-Part benchmark datasets. In the future, we will consider enhancing boundary awareness to further improve the boundary localization of parts. Moreover, further investigations of the localized segmentation errors will also be carried out.

6. REFERENCES

[1] Hossein Azizpour and Ivan Laptev, "Object detection using strongly-supervised deformable part models," in *European Conference on Computer Vision*. Springer, 2012, pp. 836–849.

[2] Yang Wang, Duan Tran, Zicheng Liao, and David Forsyth, "Discriminative hierarchical part-based models for human parsing and action recognition.," *Journal of Machine Learning Research*, vol. 13, no. 10, 2012.

[3] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan, "Towards unified human parsing and pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 843–850.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1971–1978.

[5] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, "Part-based r-cnns for fine-grained category detection," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 834–849.

[6] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian, "Multi-class part parsing with joint boundary-semantic awareness," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9177–9186.

[7] Umberto Michieli, Edoardo Borsato, Luca Rossi, and Pietro Zanuttigh, "Gmnet: Graph matching network for large scale part semantic segmentation in the wild," in *European Conference on Computer Vision*. Springer, 2020, pp. 397–414.

[8] Xin Tan, Jiachen Xu, Zhou Ye, Jinkun Hao, and Lizhuang Ma, "Confident semantic ranking loss for part parsing," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[9] Umberto Michieli and Pietro Zanuttigh, "Edge-aware graph matching network for part-based semantic segmentation," *International Journal of Computer Vision*, vol. 130, no. 11, pp. 2797–2821, 2022.

[10] Njuod Alsudays, Jing Wu, Yu-Kun Lai, and Ze Ji, "Afpsnet: Multi-class part parsing based on scaled attention and feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4033–4042.

[11] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing, "Symbolic graph reasoning meets convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[12] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis, "Graph-based global reasoning networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 433–442.

[13] Yin Li and Abhinav Gupta, "Beyond grids: Learning graph representations for visual recognition," *Advances in neural information processing systems*, vol. 31, 2018.

[14] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *2012 IEEE Conference on Computer vision and pattern recognition*. IEEE, 2012, pp. 3570–3577.

[15] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan, "Deep human parsing with active template regression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 12, pp. 2402–2414, 2015.

[16] Yafei Song, Xiaowu Chen, Jia Li, and Qinping Zhao, "Embedding 3d geometric features for rigid object part segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 580–588.

[17] Wenhao Lu, Xiaochen Lian, and Alan Yuille, "Parsing semantic parts of cars using graphical models and segment appearance consistency," *arXiv preprint arXiv:1406.2375*, 2014.

[18] Jianyu Wang and Alan L Yuille, "Semantic part segmentation using compositional model combining shape and appearance," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1788–1797.

[19] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille, "Joint object and part segmentation using deep learned potentials," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1573–1581.

[20] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[21] 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)*, 2018, pp. 801–818.

[22] Qimai Li, Zhichao Han, and Xiao-Ming Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[25] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[26] 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*, 2016, pp. 3640–3649.

[27] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6769–6778.

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