Multi-class Part Parsing with Joint Boundary-Semantic Awareness

Yifan Zhao[1] Jia Li[1,3*] Yu Zhang[1*] Yonghong Tian[2,3]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]National Engineering Laboratory for Video Technology, School of EE&CS, Peking University
[3]Peng Cheng Laboratory, Shenzhen, China
{zhaoyf, jiali}@buaa.edu.cn, zhangyulb@gmail.com, yhtian@pku.edu.cn

Abstract

Object part parsing in the wild, which requires to simultaneously detect multiple object classes in the scene and accurately segments semantic parts within each class, is challenging for the joint presence of class-level and part-level ambiguities. Despite its importance, however, this problem is not sufficiently explored in existing works. In this paper, we propose a joint parsing framework with boundary and semantic awareness to address this challenging problem. To handle part-level ambiguity, a boundary awareness module is proposed to make mid-level features at multiple scales attend to part boundaries for accurate part localization, which are then fused with high-level features for effective part recognition. For class-level ambiguity, we further present a semantic awareness module that selects discriminative part features relevant to a category to prevent irrelevant features being merged together. The proposed modules are lightweight and implementation friendly, improving the performance substantially when plugged into various baseline architectures. Our full model sets new state-of-the-art results on the Pascal-Part dataset, in both multi-class and the conventional single-class setting, while running substantially faster than recent high-performance approaches.

1. Introduction

Semantic part parsing, which decomposes objects into semantic components, has became an increasingly attended topic in computer vision. With the proposals of large benchmarks [10, 20, 24, 16] and deep learning models [25, 7, 15], recent research has shown remarkable performances in accurate segmenting of one specific category, such as vehicles [30, 35], animals [32, 31] and human bodies [40, 18]. High-quality part parsing results would be of great use in further applications, such as object detection [2], pose estimation [12], fine-grained action detection [34] and categorization [41].

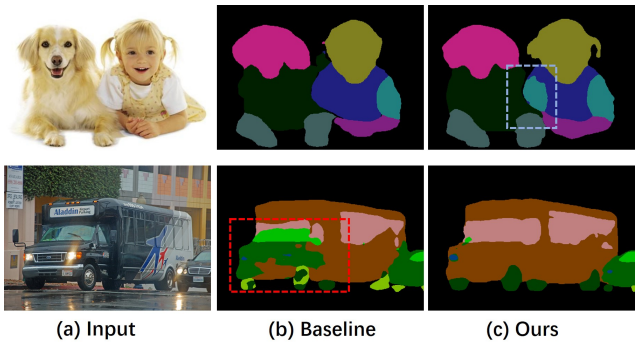

Figure 1. Motivation of the proposed approach. Simultaneously parsing the parts of multiple semantic classes an input scene (a) has its own challenges, including inaccurate boundary localization and inter-class appearance ambiguity (b). The proposed boundary-semantic awareness network effectively addresses these issues (c).

Various existing approaches have been proposed to address object part parsing, which could be roughly divided into two categories. The first category usually focuses on exploring the inner relationship [31, 32] and the structure information [18, 16] of object parts. For example, Liang *et al.* [16] proposed a self-supervised structure loss to maintain the parsing structure. Wang *et al.* [31] built a hierarchical tree structure to compose basic boundary landmarks into parts according to their spatial relations. Some other works [37, 38, 30] also resorted to additional structural information, *e.g.*, human pose and 3d information. The second category [9, 36, 7, 8, 42] focuses on improving the parsing resolutions in images or feature maps. For example, Chen *et al.* [9] proposed an attention model to fuse the parsing results of different image zooming scales. Xia *et al.* [36] proposed a two-stage network to fuse the global feature with detected local features.

Despite the effectiveness of existing models, they mainly address single-class setting, where the object is assumed to be well-localized beforehand. In this paper, we propose to

*Correspondence should be addressed to Jia Li and Yu Zhang. URL: http://cvteam.net

investigate the wilder multi-class object part parsing problem, which simultaneously handles all semantic classes and parts within each class in the scene. As shown in Fig. 1, in this novel setting, even the strong recent baseline [8] may face additional challenges. In particular, the cluttered appearance of multiple objects and the inter-class ambiguity may cause inaccurate boundary localization and severe classification error.

To address multi-class part parsing and handle the above issues, we propose a novel deep architecture with boundary-semantic awareness. At the core of the proposed approach, we develop two simple yet effective modules. The first one is a boundary-aware spatial selection module, which makes mid-level features attend to part-level boundaries at multiple scales. At each scale, features are promoted by a spatial attention block supervised with class-agnostic part boundaries, and passed to the next scale. With such a coarse-to-fine boundary refinement strategy, the network learns to focus on solving the ambiguity along part boundaries and produce finer parsing results. After that, the boundary-filtered mid-level features are then fused with the high-level features, to jointly preserve the shallow boundary information and the deep semantic context.

The final aggregated features are a mixing of part-class attributes, which may lead to confusions for parts with similar appearance across different classes, such as the bus example in Fig. 1. However, if the model understands the class distributions at each pixel, the head of the bus would not be mistakenly recognized as a car and classification errors could be fixed largely. Based on this observation, we further introduce a semantic-aware module to select the most beneficial features at each pixel to avoid such confusions. In this module, channel selection is performed at each location of the aggregated features, which is supervised by an additional branch predicting semantic classes. The proposed modules are lightweight and simple to implement, and can boost the performance significantly when plugged into various baseline architectures. Combining the proposed two lightweight modules into a unified parsing network, the proposed approach achieves new state-of-the-art results on the Pascal-Part dataset, in both the multi-class and single-class settings for semantic object parsing.

Contributions of this paper are summarized as follows: 1) We propose to address object part parsing in the less explored multi-class setting, and propose a unified network architecture to solve this important problem. 2) We introduce two lightweight yet effective modules, the boundary awareness and semantic awareness module, to address the part-level and class-level ambiguities in multi-class object part parsing. 3) The proposed approach is able to achieve the state-of-the-art results on both the multi-class and the conventional single-class settings, while running a magnitude faster than recent high-performance approaches.

2. Related Work

Structure-based part parsing. The structure-based methods [24, 19, 31, 32, 18, 16, 37, 38] mainly resort to the compositional or morphological model to regularize part parsing. For example, Wang *et al.* [31] build a compositional model under different viewpoints and poses to parse certain animal classes. In [33], a hierarchical poselets model is built to represent the composition of human bodies. Wang *et al.* [32] propose a joint deep model to explore the relationship between parts and body with fully-connected CRF. Some studies [38, 35] also use the structured tree model to organize the part in hierarchical ways. While Liang *et al.* [17] propose a structure-evolving LSTM to refine the parsing results by generated super-pixel maps.

For human parsing tasks, recent studies [27, 13, 16, 37, 19, 11, 14] usually explore relationships between segmentation and other tasks, especially pose estimation. For example, Liang *et al.* [19] propose a structure evolving LSTM model to learn a structure graph model for human parsing. Xia *et al.* [37] propose a joint model to refine the segmentation results by supervised pose estimation. Liang *et al.* [16] propose a self-supervised structure-sensitive network to simultaneously estimate human pose and part parsing. Fang *et al.* [13] propose a pose-guided model to make use of the dataset annotations as part priors. Yamaguchi *et al.* [40] propose a joint estimation to benefit clothing segmentation and human pose and then improve the garment retrieval methods [39]. Nie *et al.* [27] propose a mutual learning model to adapt pose estimation task to promote the part segmentation results. Besides, Song *et al.* [30] embed the extra 3D information into part segmentation models with a teacher-student architecture.

Scale aggregation. The scale aggregation techniques [7, 3, 21] in semantic segmentation have become a popular way to enhance the model representation. Badrinarayanan *et al.* [3] propose a short connection way to transfer the lower features into higher-level representations. Chen *et al.* [7] propose a novel atrous spatial pyramid pooling architecture to aggregate feature maps with different field of views. Amirul *et al.* [1] propose a feedback refinement network with gating strategy to aggregate features of multiple levels.

In the specific area of part segmentation, some representative works also make use of aggregation of different scale features. Xia *et al.* [36] propose an auto-zoom framework to fuse the global feature with the detected large local features. Chen *et al.* [9] present a attention-based fusion strategy to fuse the image features with different resolutions. Zhao *et al.* [42] propose a weight-sharing network with input image of different resolutions. Luo *et al.* [26] use generative adversarial networks in both high and low resolution features to enhance the semantic consistency. Moreover, there are also some works which focus on the accurate boundary matting [28] and self-attention mechanism [5].

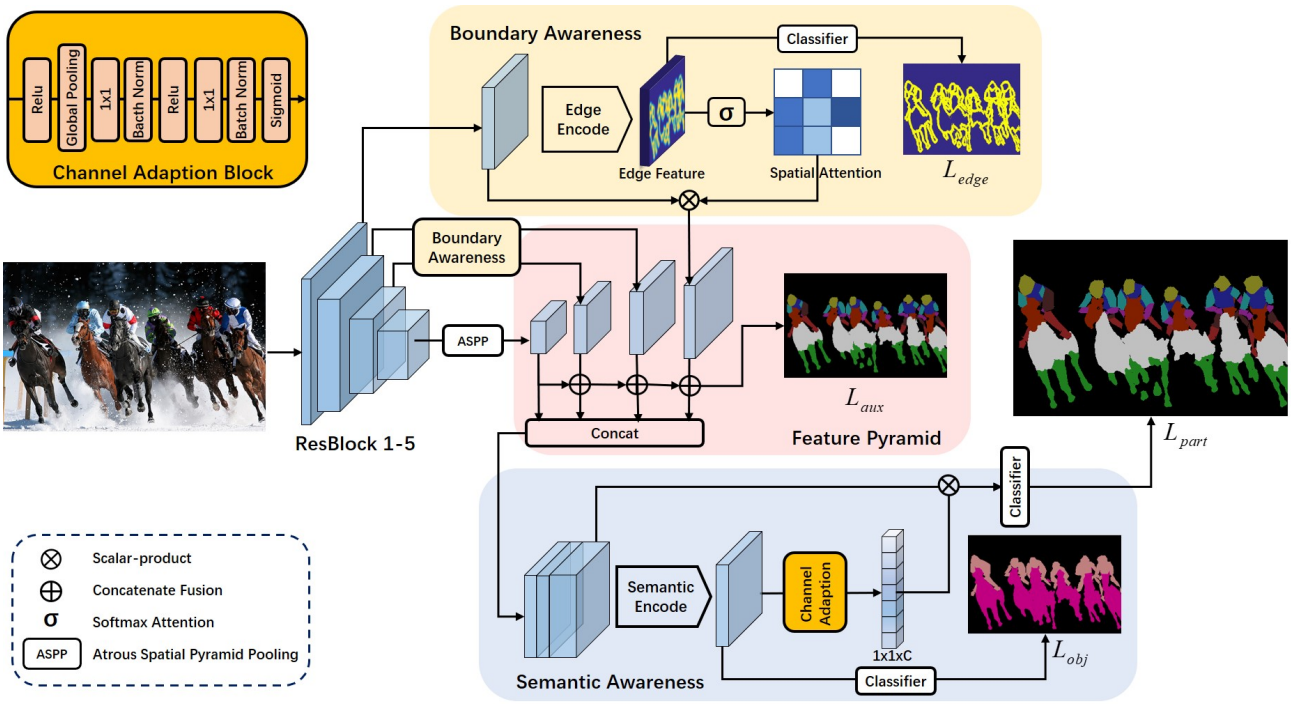

Figure 2. Our joint Boundary-Semantic Awareness Network (BSANet) framework, is mainly composed of a boundary aware spatial selection module and a semantic aware channel selection module. The boundary awareness module aims to aggregate the local features near boundaries in low-level and semantic context in high-level, which is supervised by an edge regression loss. Semantic awareness module aims to use the supervised semantic object context to enhance the expression of class-relevant feature channels.

3. Method

3.1. Overview

In this section, we propose a novel joint Boundary-Semantic Awareness Network (BSANet) for multi-class object part parsing, which is composed of two modules, *i.e.*, a boundary-aware spatial selection module and a semantic-aware channel selection module (see Fig. 2). In the first module, we adopt a boundary-aware spatial attention to enhance the features near boundaries, which are usually ambiguous in the downsampled high-level features. Features from each level pass through the boundary module to construct a cascade feature pyramid. These features are stepwise fused to predict initial segmentation results. We then concatenate these features with 1×1 convs and pass them into the semantic selection module, which emphasizes the class-correlated features and suppresses the irrelevant ones.

Given a picture $\mathcal{I}$ with extracted feature $\mathbf{P}^{W \times H \times C}$, our joint Boundary-Semantic Awareness Network can be formulated as follows:

$$\phi(\mathcal{I}) = \mathbb{C}(\mathbb{S}(\mathbf{P} \odot W_s) \odot W_c), \tag{1}$$

where $\mathbb{S}$ is the boundary supervised spatial selection module and $\mathbb{C}$ is the semantic supervised channel selection module. W_s and W_c are the attention weights for $\mathbb{S}$ and $\mathbb{C}$, respectively. ϕ is the final segmentation model and $\odot$ represents the dot product operation.

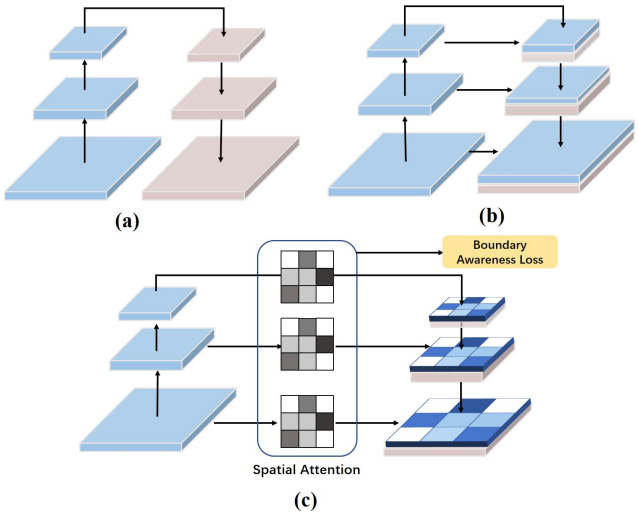

Figure 3. Differences of three pyramid decoders. **(a):** Top-down pyramid decoder. **(b):** Top-down pyramid decoder with feature transfer. **(c):** Spatial aware feature pyramid.

3.2. Boundary Aware Spatial Selection

For object part parsing, boundary ambiguities exist universally as there is often no apparent image evidence that

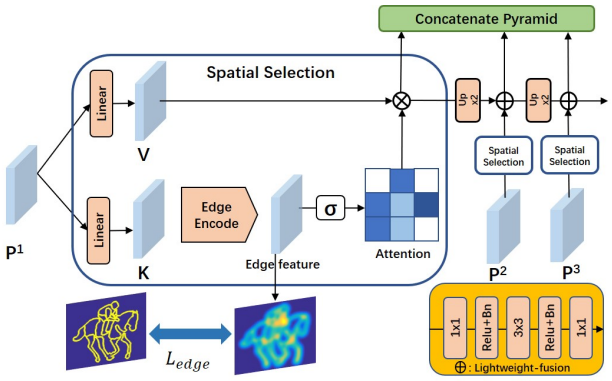

Figure 4. Illustration of boundary-aware spatial selection mechanism. $\mathbf{P}^i$ represents the pyramid feature from different encode blocks. $\oplus$ represents the light-weight fusion block (view in yellow). These features are finally concatenated to get the final output (view in green).

implies the transition of parts. To address this issue, several existing approaches [6, 4] propose to add an additional branch to learn accurate boundary predictions, which may introduce large computational burdens. In this section, we propose a lightweight boundary attention module.

The basic idea is that in the task of part parsing, low-level and mid-level features should take more responsibility along part boundaries as they provide more detailed localization cues, while being suppressed in the inner region of parts due to their limited discriminative power. Keeping this idea in mind, we propose to detect class-agnostic part boundaries at early stages of feature extraction, which is possible as no semantic information needs to be inferred. The predicted soft boundaries are then used in an attention mechanism to emphasize the features along boundaries and suppress others for mid-level and low-level stages. Such attention is performed at multiple feature scales to detect part transitions at various levels. Finally, the boundary-attended early-stage features then serve as compensations to high-level features to preserve both classification and boundary localization accuracy.

As illustrated in Fig. 3, the classical top-down pathway decodes the high level feature by up-sampling or trans-convolutions to enlarge the parsing resolution, which provides strongly semantical features but coarser in spatial. While in the state-of-the-art models [22, 3], lateral connections are adopted to fuse the low-level and high-level features (see Fig. 3 (b)). However, in this way, the low-level and high-level features are treated equally in spatial, which may suppress the semantical features. To this end, we propose a novel boundary awareness feature pyramid which emphasizes the features near boundaries at low level to provide finer spatial predictions, and maintain the high-level features to provide the semantical predictions.

As it stands, the boundary awareness module takes a recurrent multi-scale structure. In Fig. 4, we show the detailed architecture of the first scale, which for other scales share the similar structures. Given the low-level features $\mathbf{P}^1$, two linear mappings are adopted to produce two transformed feature maps $\mathbf{V}^{N\times C}$ and $\mathbf{K}^{N\times C}$, where $N = W \times H$. The feature map $\mathbf{K}$ is then passed through an edge encoder $\varphi(\mathbf{K})$ to get the part boundary features. With these features, we pass them to a softmax attention function σ to generate the attention map, and point-wise multiply with the feature map $\mathbf{V}$ to yield boundary-enhanced features. The above operations can be formally represented as:

$$\begin{aligned}\mathbb{S}(\mathbf{P}^s) &= \frac{\varphi(\mathbf{K}^s_{i,j})}{\sum_i \varphi(\mathbf{K}^s_{i,j})} \odot \mathbf{V}^s_{i,j}, \\ \mathbf{V}^s_{i,j} &= tanh(\mathbf{w}_v\mathbf{P}^s + \mathbf{b}_v), \\ \mathbf{K}^s_{i,j} &= tanh(\mathbf{w}_k\mathbf{P}^s + \mathbf{b}_k),\end{aligned} \tag{2}$$

where $\mathbf{w}_k, \mathbf{b}_k, \mathbf{w}_v, \mathbf{b}_v$ are the learnable parameters and $\odot$ is the scalar-product operation.

Note that instead of self-supervised attention, the edge features $\varphi(\mathbf{K})$ is directly regularized by class-agnostic part boundaries. One typical edge decoder is composed of 1×1, 3×3, 1×1 convolutional blocks with batchnorm and ReLU. To this end, we pass $\varphi(\mathbf{K})$ through a binary cross-entropy classifier, which generates soft part boundaries and compared with the ground-truth. Such supervision is adopted on all three scales. Thus, the total loss of the auxiliary part boundary detection task is given by

$$\begin{aligned}L_{edge} = -\sum_{k=1}^{S}\sum_{i\in\Omega_\mathcal{I}} &\gamma^s_k\gamma^c(1-\mathbf{y}_{ik})\log(1-\mathbf{p}_{ik}) \\ &+\gamma^s_k(1-\gamma^c)\mathbf{y}_{ik}\log(\mathbf{p}_{ik}),\end{aligned} \tag{3}$$

where $\mathbf{p}_{ik}$ is the prediction of ground-truth $\mathbf{y}_{ik}$ and $\Omega_\mathcal{I}$ is the lattice domain of image $\mathcal{I}$. γ^s_k is the balanced weight of each scale and γ^c is the class balance weight. By regularizing the feature attention by the edge loss, each pyramid feature $\mathbf{P}^s$ of scale s is filtered by the spatial attention $\mathbb{S}$. These features are stepwise fused with light-weighted fuse operations (denoted as $\oplus$ in Fig. 4) to get $\mathbf{Q}^s$. Then these features are concatenated to fuse to the final output $\mathcal{F}$, which can be formulated in Eqn. (4)

$$\begin{aligned}\mathbf{Q}^k &= \mathbb{S}(\mathbf{P}^1) \oplus \mathbb{S}(\mathbf{P}^2) \oplus \ldots \mathbb{S}(\mathbf{P}^k), k = 1...S, \\ \mathcal{F} &= \sum_{i=1}^{S} \mathbf{Q}^i \cdot \mathbf{w}^i_q,\end{aligned} \tag{4}$$

where S is the downsample scale in the spatial feature pyramid. Especially, for the high-level feature of the last layer, we adopt the ASPP architecture without spatial selection to emphasize its semantic comprehending.

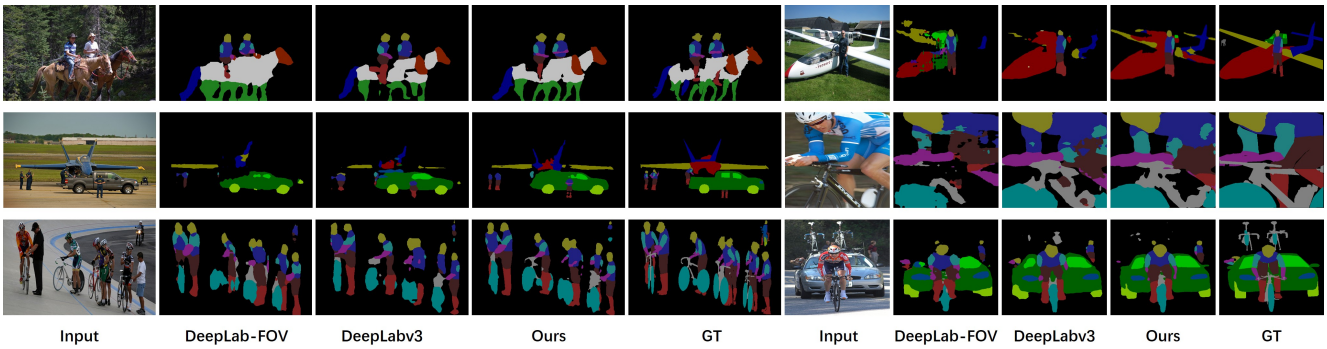

Figure 5. Qualitative comparisons on PASCAL-Part dataset. Our model generates superior results with finer local details and semantic understanding comparing to the-state-of-the-art models.

3.3. Semantic Aware Channel Selection

As discussed in Section 3.1, each channel map has a certain mapping relation to corresponding semantic categories, while the unrelated features would be confusions to the final prediction. In order to address this limitation, we use auxiliary semantic information to supervise the attention weights on related features. As a result, it actually makes multi-class part segmentation a cascade process, in which understating part semantics requires understanding the classes as a prior.

As shown in Fig. 2, we obtain features $\mathcal{F} \in \mathbb{R}^{W' \times H' \times C'}$ of image $\mathcal{I}$ from the spatial selection module. The semantic encoder ξ shares similar structures with edge encoder, which is composed of several $3 \times 3, 1 \times 1$ convolutional blocks with batchnorm and ReLU. To obtain the predicted semantic features $\xi(\mathcal{F}^c_{i,j})$, we use per-pixel semantic labels to regularize this feature map with a softmax cross-entropy loss L_{obj}. In this way, features relevant to specific object categories are emphasized by the supervision, which can further eliminate the inter-class confusions.

We resort to soft semantic label $\mathbf{V}$ to encode the channel information of semantic feature $\xi(\mathcal{F}^c_{i,j})$, which is generated by global pooling operations to reduce the size $W' \times H' \times C'$ to $1 \times 1 \times C'$. The cth value of $\mathbf{V}$ can be represented as:

$$V^c = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} \xi(\mathcal{F}^c_{i,j}), c = 1 \dots C', \quad (5)$$

The final channel attention operation $\mathbf{A}$ is learned by two fully connected layers, which emphasizes the object-relevant channels in feature $\mathcal{F}$:

$$\begin{aligned} \mathbf{A} &= \sigma(\tau(\mathbf{w}_1 \mathbf{V} + \mathbf{b}_1)\mathbf{w}_2 + \mathbf{b}_2), \\ \mathcal{G}^c &= \mathbf{A}^c \cdot \mathcal{F}^c, c = 1 \dots C', \end{aligned} \quad (6)$$

where τ is the Rectified Linear Unit(ReLU) and σ is the sigmoid function. $\mathbf{w}_{\{1,2\}}, \mathbf{b}_{\{1,2\}}$ are learned parameters of fully connected layers. $\mathcal{G}$ is the final feature map after channel selection. This fused feature map passes through a simple classifier with channel reductions to get the final prediction.

3.4. Joint Boundary-Semantic Awareness

With the proposal of two selection modules, we further build our joint Boundary-Semantic Awareness Network (B-SANet) based on DeeplabV3 [9] encoder, which is the state-of-the-art semantic segmentation network. The boundary aware module and semantic aware module are conducted sequentially to get the final prediction. Our final framework is a boundary-semantic joint solving procedure, which conducts different losses in corresponding stages.

At the end of spatial feature pyramids in Fig. 2, we further add an auxiliary loss L_{aux} for the last block in feature pyramid to accelerate the training procedure. The final part prediction loss L_{part} and L_{aux} are standard cross-entropy loss defined on part categories. The final loss function L_{sum} of our framework is calculated as a balanced sum of 4 terms,

$$L_{sum} = \lambda_e \cdot L_{edge} + \lambda_o \cdot L_{obj} + \lambda_a \cdot L_{aux} + L_{part}, \quad (7)$$

where $\lambda_{\{e,o,a\}}$ are balanced terms. Moreover, the boundary and object semantic maps are automatically generated from the semantic part labels, which do not requires additional annotations.

4. Experiments

4.1. Experiment Settings

Dataset. PASCAL-Part dataset [10] is the largest dataset to date for multi-class object part parsing with pixel-level part annotations. It contains 10103 images with pixel-level part annotation of 20 semantic objects collected from PASCAL-VOC2010 challenge. Specially, the dataset contains very detailed part annotations, including eyes, noses and mouths of humans and animals. We follow the merging rules of [31, 32] for animals, [30] for vehicles, and [9, 16] for human bodies. This dataset yields 58 part classes in total. We use 4998 images in the $trainset$ for training and 5105 images in $valset$ for testing, which is divided by [10].

For single-class part parsing, PASCAL-Person-Part is a widely used benchmark with dozens of models, which is also a sub-dataset generated from PASCAL-Part [10]. We

Table 1. Segmentation Performance of mIoU on PASCAL-Part Benchmark. Avg.: the average per-object-class mIoU. mIoU: per-part class mIoU. †: use pretrained model on MS-COCO dataset.

Method	backg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	potted	sheep	sofa	train	tv	mIoU	Avg.
SegNet [3]	85.4	13.7	40.7	11.3	21.7	10.7	36.7	26.3	28.5	16.6	8.9	16.6	24.2	18.8	44.7	35.4	16.1	17.3	15.7	41.3	26.1	24.4	26.5
FCN [25]	87.0	33.9	51.5	37.7	47.0	45.3	50.8	39.1	45.2	29.4	31.2	32.5	42.4	42.2	58.2	40.3	38.3	43.4	35.7	66.7	44.2	42.3	44.9
Deeplab-Fov† [7]	89.8	40.7	58.1	43.8	53.9	44.5	62.1	45.1	52.3	**36.6**	41.9	38.7	49.5	53.9	66.1	49.0	45.3	45.3	40.5	76.8	56.5	49.9	51.9
Deeplabv3 [8]	90.8	44.8	60.9	46.7	56.8	47.9	65.9	50.0	60.4	35.7	50.5	42.1	55.9	60.6	69.3	54.5	52.0	48.7	**43.8**	79.8	56.8	54.4	55.9
Baseline	90.6	45.7	60.7	48.5	55.7	46.8	66.9	50.2	59.4	33.1	48.9	38.3	55.0	58.7	68.6	54.3	50.3	46.5	42.6	78.1	56.4	54.0	55.0
BSANet-101	**91.6**	**50.0**	**65.7**	**54.8**	**60.2**	**49.2**	**70.1**	**53.5**	**63.8**	36.5	**52.8**	**43.7**	**60.8**	**66.0**	**73.3**	**58.4**	**55.0**	**49.6**	43.1	**82.2**	**61.4**	**58.2**	**59.1**

Table 2. Segmentation Performance of mIoU on Pascal-Person-Part Benchmark. $*$: re-trained on the proposed dataset. Pose An.: learning with auxiliary pose annotation.

Method	Pose An.	head	torso	u-arms	l-arms	u-legs	l-legs	bkg	Avg.
HAZN [36]		80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [9]		81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
LG-LSTM [19]		82.72	60.99	45.40	47.76	42.33	37.96	88.63	57.97
SS-JPPNet [16]		83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
Graph-LSTM [19]		82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
SS-NAN [42]		86.43	67.28	51.09	48.07	44.82	42.15	**97.23**	62.44
Deeplabv3* [8]		84.06	66.96	54.26	52.80	48.08	43.59	94.79	63.50
Str.-LSTM [17]		82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint [37]	✓	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
MuLA [27]	✓	-	-	-	-	-	-	-	65.10
Baseline		82.94	66.18	53.90	52.71	46.54	43.02	94.51	62.83
BSANet-101		86.49	70.20	59.31	58.72	51.91	49.32	95.62	67.37
BSANet-152		**86.98**	**71.35**	**61.36**	**60.26**	**53.28**	**49.95**	95.79	**68.43**

follow the annotations of [9, 37], which is composed of 3533 images (1716 images for training and 1817 images for testing) of 7 classes, *i.e.*, Background, Head, Torso, Upper/Lower Arms and Upper/Lower Legs. This challenging dataset contains images of multi-person in various scales.

Training details. We refer to the same training schemes in prior works [7, 8]. These images are randomly left-right flipped and resized from 0.5 to 2.0 times in our experiments. We train our model with the start learning rate $7e-3$ with a weight decay for all these datasets. Our super parameters are simply set without bells and whistles. It takes us about 15 hours to train a model with 50K iterations for PASCAL-Part dataset on one NVIDIA 1080Ti GPU. For the PASCAL-Person benchmark, we only trained our model for 30K iterations to prevent possible overfitting. The inference time is within 0.2 seconds per 512×512 image. The atrous rate of ASPP follows the prior work [7], which is set as $(6, 12, 18)$. We set the downsample $stride = 16$ in all our models and we use the Resblock $2 \sim 5$ to build our pyramid decoder considering the memory and computation cost. We set $\lambda_e = 0.10$, $\lambda_a = 0.20$ and $\lambda_o = 0.40$ to make the weight balance and enhance the part classification regularization. The γ_k^s is simply set as 1 when upsampled to the same scale and $\gamma^c = 0.1$ to emphasize the boundaries.

Baselines and evaluations. To validate our first attempt on the multi-object class part parsing challenge, we compare our framework with four state-of-the-art representative works [3, 25, 7, 8]. To make a fair comparison with these models, we carefully tuned the sup-parameters following the training schemes in original papers. For [3, 25], we finetune the official model of with ImageNet pretrained VGG-16 [29] backbone. For [7], we adopt the ResNet-101 model pretrained on MS-COCO dataset [23], which is provided by the author. For [8, 7] and our model, we use the ResNet-101 as backbone for a fair comparison. Notably, our model is trained without any additional datasets like MS-COCO.

We reproduce the Deeplabv3 model [8] in PyTorch as our baseline, which performs a bit lower performance in the part segmentation benchmark, as shown in Tab. 1. In this paper, we choose the mean Intersection over Union (mIoU) as evaluation criteria for all experiments owing that pixel accuracy is not sensitive to the segmentation of small parts.

4.2. Comparisons with the state-of-the-art

PASCAL-Part Benchmark. As shown in Tab. 1, we compare our model and four state-of-the-art methods with two criterions, *i.e.*, per-object-class mIoU and per-part-class mIoU. FCN [25], which is the fundamental work of semantic segmentation, achieves 42.3% in per-part mIoU of 58 classes. With the improvement of larger field of views, [7] improves a lot by achieving 49.9%, which is finetuned on the COCO-pretrained model. By reproducing Deeplab-v3 [8] as our baseline, [8] and our reproduction obtain similar results as 54.4% and 54.0%, respectively. Starting

Table 3. Performances of Ablation Experiments on Pascal-Person-Part Dataset. BA-1: boundary awareness module with only one pyramid block. w/o sup: model without auxiliary supervision.

Method	head	torso	uarm	larm	uleg	lleg	bg	Avg.
baseline	82.94	66.18	53.90	52.71	46.54	43.02	94.51	62.83
lateral-all	84.98	68.04	56.30	55.12	49.82	45.90	95.24	65.06
BA-1	84.92	67.64	55.07	54.78	49.07	45.49	95.18	64.59
BA-all	**86.53**	69.76	58.64	57.57	51.38	47.96	95.55	66.77
BA-all(w/o sup)	84.97	68.17	55.69	54.50	49.78	46.24	95.22	64.94
BSANet-final	86.49	**70.20**	**59.31**	**58.72**	**51.91**	**49.32**	**95.62**	**67.37**

from a strong baseline, our final model combining with the boundary selection and semantic selection, achieves a higher result on multi-class object part segmentation: 58.2% on mIoU, outperforms state-of-the-art models by a margin.

Moreover, our model shows superior performance on objects with more part components or detailed information, *e.g.*, bird by 8.1% and horse by 5.4%. This verifies the effectiveness of our model which focuses on detailed information by aggregating multi-level features. For some special categories, such as chair and sofa, which are not composed of multiple parts, our model also generates comparable results to other methods. The qualitative comparisons of parsing results are shown in Fig. 5. Comparing to those methods in the visualized results, our model generates clear boundaries and is sensitive to pixels in finer scales. Moreover, our method generates superior results on both single-class part parsing and multi-class occlusion scenarios.

PASCAL-Person-Part. To validate the performance on single-class parsing tasks, we conduct experiments on the widely-used human parsing benchmark. We compare our methods with 10 state-of-the-art models with the reported performances, as shown in Tab. 2. Especially, in [37, 27], human pose annotation is used as auxiliary information to facilitate the part parsing task. The Deeplabv3 model [8] achieves results of 63.50% mIoU with only pixel-level part annotations. While [17] generates a little higher results by conducting refinement with superpixels.

Moreover, our baseline model shows a slightly lower performance than Deeplabv3 [8], but still surpass most of the state-of-the-art models. Benefiting from the boundary-semantic awareness framework, our model reaches a huge performance boost of 68.43%, improved about 5.6% in mIoU. With class-agnostic part-level boundary guidance, our model shows superior results on parts with confused outlines, *e.g.*, human arms and legs. Our model improves 6.0% mIoU on lower-arms and 6.6% on upper-arms, comparing to the best model [37] with additional annotations.

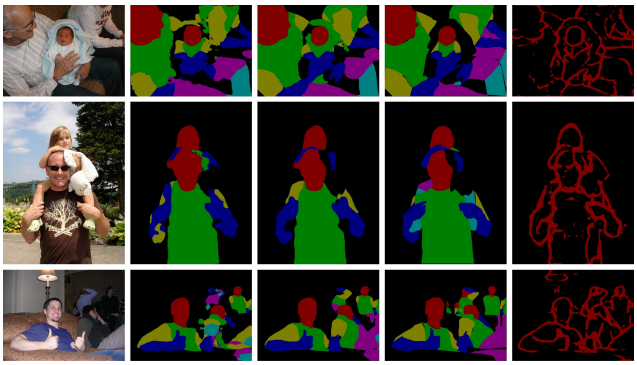

Figure 6. Qualitative results of boundary module on PASCAL-Person-Part. Our model obviously improves the detailed local information, especially in the area near boundary attention.

4.3. Performance Analysis

In this section, we evaluate the effectiveness of our two proposed modules of BSANet, which is boundary-aware spatial selection module and semantic-aware channel selection module, respectively. We further analysis the performances of extensions on other similar backbones.

Boundary-aware spatial selection. To evaluate the effectiveness of our spatial selection module, we reconstruct our model with different ablation factors. Tab. 3 shows the ablations on PASCAL-Person dataset. In the first row, We first build our model with lateral connections between low-level and high-level feature maps, shown in Fig. 3 (b). This effective operation can improve the baseline by 2.2 mIoU in total, while our boundary aware spatial selection (BA) module further improves the baseline from 62.83% to 67.37%.

We further conduct experiments with different spatial blocks, while BA-1 is the model which only concatenates the block-2 feature of ResNet backbone. By introducing more blocks into our module, the performance improves steadily from 64.59% to 66.77%. The forth line in Tab. 3 shows the model without boundary regularization loss, which is similar to a self-attention procedure. The performance drops steadily without the auxiliary loss and our semantic supervised selection block can further boost the performance to 67.37%, finally outperforms 4.5% by the high baseline. From the visualized results in Fig. 5, our model improves notably on details near the boundary attention area, which is visualized in the fifth column calculated by softmax operations.

Semantic-aware channel selection. To further evaluate the semantic-aware module, we conduct our experiments on PASCAL-Part dataset of 20 VOC semantic object categories. As illustrated in Fig. 7, the model without the semantic selection would be confused in near-duplicated local features, *e.g.*, the boats in the second column are mistaken as cars.

Table 4. Module Ablation Experiments on PASCAL-Part Dataset. (Both baseline and BSANet adopt the same Res-101 backbone.)

Method	Boundary Sel.	Semantic Sel.	Loss Reg.	mIoU
baseline				54.03
BSANet	✓		✓	56.97
BSANet		✓	✓	55.87
BSANet	✓	✓		54.59
BSANet	✓	✓	✓	**58.18**

Figure 7. Qualitative results on PASCAL-Part Dataset. The second column without semantic selection is hard to distinguish confuse classes, while our model in the third column shows better performance.

Quantitative results can be found in Tab. 4. The boundary selection module achieves nearly 3% mIoU boost compared to the baseline in the first row. Our single semantic selection module is also effective in many situations, which promote the results from 54.03% to 55.87%. Finally combining these two modules reaches the best performance, which also has a bit higher computation cost. Our final model achieves a huge performance boost by 4.15% in mIoU of all 58 part categories. By adopting these models without supervised selection, which is a self-attention procedure, reaches a small performance boost of 54.59%.

Models with different backbones. We further explore on more backbone with different depths of [8], as shown in Tab. 5. Our model is easy to extend on several encoder-decoder architectures, which promotes the baseline by a large margin. While the shallower network shows comparable performance with the IoU of 62.29%, drops by 6.1%. The deeper network has a strong ability in handling small and confusing parts like legs and arms.

Inference Time. Comparing to [37] with 6.0s and [17] with 1.3s per image in inference phase, our model takes less than 200ms per 512×512 image on a single consumer 1080Ti GPU, which only adds 9.8% inference time comparing to the baseline [8]. While other fast model [42] with 500ms inference time generates much lower performance.

Failure modes. As the first row in Fig. 8, our model mistakes the dogs as horses and introduces more errors that the heads and legs are recognized as horse heads and legs. While dogs in Deeplabv3 [8] with large class confusions show higher performance in mIoU. For the second case with severely occlusions and viewpoint variances, both baseline and our model still face great challenges.

Table 5. Performances of different backbones on PASCAL-Person-Part Dataset.

Method	head	torso	uarm	larm	uleg	lleg	bg	Avg.
Baseline(Res50)	82.07	63.00	49.46	48.22	43.57	39.35	94.09	59.96
BSANet-50	84.11	65.92	53.23	51.79	45.07	41.33	94.56	62.29
Baseline(Res101)	82.94	66.18	53.90	52.71	46.54	43.02	94.51	62.83
BSANet-101	86.49	70.20	59.31	58.72	51.91	49.32	95.62	67.37
Baseline(Res152)	83.60	66.23	54.46	52.76	47.02	42.76	94.65	63.07
BSANet-152	**86.98**	**71.35**	**61.36**	**60.26**	**53.28**	**49.95**	**95.79**	**68.43**

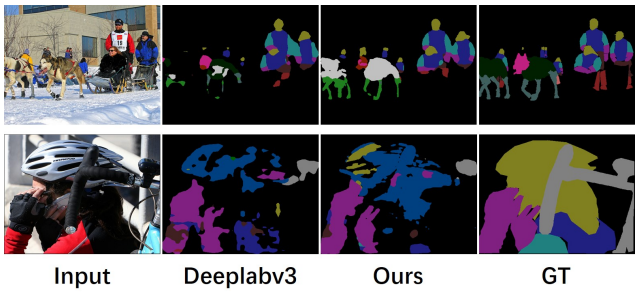

Figure 8. Two typical failure modes. Input image, DeepLabv3, our results and ground truth masks. Our model can be confused in complex images which are easily annotated by human.

5. Conclusion

In this paper, we make an attempt on the less explored multi-class object part parsing task and propose a unified framework to handle its two main challenges, *i.e.*, inaccurate boundary localization and inter-class appearance ambiguity. For the first challenge, we resort to semantic boundary information generated from part labels to regularize a spatial selector, which aims to aggregate low-level features with more local details and high-level feature with semantic comprehending. For the second challenge, we propose a semantic supervised channel selector to choose the object-relevant feature maps. By conducting these two modules sequentially, our framework outperforms the-state-of-the-art models in both single-class and multi-class parsing tasks.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under grant 2017YFB1002400, the National Natural Science Foundation of China (61672072, U1611461 and 61825101), and Beijing Nova Program under Grant Z181100006218063.

References

[1] Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *CVPR*, pages 3751–3759, 2017. 2

[2] Hossein Azizpour and Ivan Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, pages 836–849, 2012. 1

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE T-PAMI*, (12):2481–2495, 2017. 2, 4, 6

[4] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *CVPR*, pages 3602–3610, 2016. 4

[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, pages 5659–5667, 2017. 2

[6] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, pages 4545–4554, 2016. 4

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE T-PAMI*, 40(4):834–848, 2018. 1, 2, 6

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 1, 2, 6, 7, 8

[9] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016. 1, 2, 5, 6

[10] 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 *CVPR*, pages 1971–1978, 2014. 1, 5

[11] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Structured feature learning for pose estimation. In *CVPR*, pages 4715–4723, 2016. 2

[12] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *CVPR*, pages 843–850, 2014. 1

[13] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, pages 70–78, 2018. 2

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, pages 770–785, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[16] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE T-PAMI*, 2018. 1, 2, 5, 6

[17] Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, and Eric P Xing. Interpretable structure-evolving lstm. In *CVPR*, pages 1010–1019, 2017. 2, 6, 7, 8

[18] 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 T-PAMI*, 12(37):2402–2414, 2015. 1, 2

[19] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, pages 125–143. Springer, 2016. 2, 6

[20] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, December 2015. 1

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017. 2

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[24] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1):253–265, 2014. 1, 2

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 6

[26] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *ECCV*, pages 418–434, 2018. 2

[27] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, pages 502–517, 2018. 2, 6, 7

[28] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, pages 92–107. Springer, 2016. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[30] Yafei Song, Xiaowu Chen, Jia Li, and Qinping Zhao. Embedding 3d geometric features for rigid object part segmentation. In *ICCV*, pages 580–588, 2017. 1, 2, 5

[31] Jianyu Wang and Alan L Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *CVPR*, pages 1788–1797, 2015. 1, 2, 5

[32] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *ICCV*, pages 1573–1581, 2015. 1, 2, 5

[33] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR*, pages 1705–1712. IEEE, 2011. 2

[34] 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*, 13(Oct):3075–3102, 2012. 1

[35] Alan Yuille Wenhao Lu, Xiaochen Lian. Parsing semantic parts of cars using graphical models and segment appearance consistency. In *BMVC*, 2014. 1, 2

[36] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, pages 648–663, 2016. 1, 2, 6

[37] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, pages 6769–6778, 2017. 1, 2, 6, 7, 8

[38] Fangting Xia, Jun Zhu, Peng Wang, and Alan L Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016. 1, 2

[39] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, pages 3519–3526, 2013. 2

[40] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012. 1, 2

[41] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849, 2014. 1

[42] Jian Zhao, Jianshu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. Self-supervised neural aggregation networks for human parsing. In *CVPR Workshops*, pages 7–15, 2017. 1, 2, 6, 8