arXiv:1812.03904v1 [cs.CV] 10 Dec 2018

Attention-guided Unified Network for Panoptic Segmentation

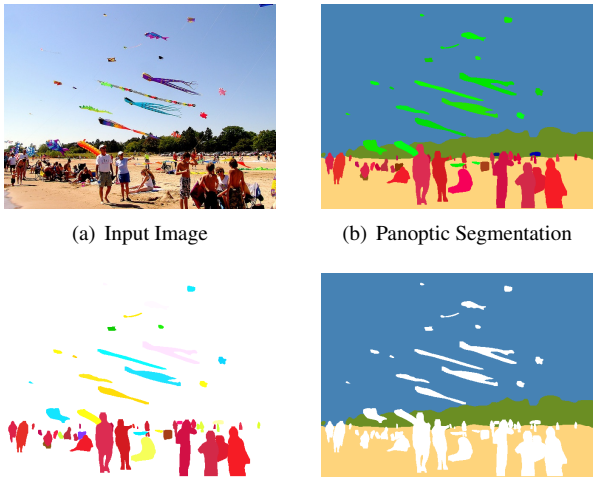
Yanwei Li[1,2], Xinze Chen[3], Zheng Zhu[1,2], Lingxi Xie[4], Guan Huang[3],
Dalong Du[3], Xingang Wang[1]
[1] Institute of Automation, CAS [2] University of Chinese Academy of Sciences
[3] Horizon Robotics, Inc. [4] The Johns Hopkins University
{liyanwei2017,zhuzheng2014,xingang.wang}@ia.ac.cn
{guan.huang,xinze.chen,dalong.du}@horizon.ai {198808xc}@gmail.com

Abstract

This paper studies panoptic segmentation, a recently proposed task which segments foreground (FG) objects at the instance level as well as background (BG) contents at the semantic level. Existing methods mostly dealt with these two problems separately, but in this paper, we reveal the underlying relationship between them, in particular, FG objects provide complementary cues to assist BG understanding. Our approach, named the Attention-guided Unified Network (AUNet), is an unified framework with two branches for FG and BG segmentation simultaneously. Two sources of attentions are added to the BG branch, namely, RPN and FG segmentation mask to provide object-level and pixel-level attentions, respectively. Our approach is generalized to different backbones with consistent accuracy gain in both FG and BG segmentation, and also sets new state-of-the-arts in the MS-COCO (46.5% PQ) benchmarks.

1. Introduction

Scene understanding is a fundamental yet challenging task in computer vision, which has a great impact on other applications such as autonomous driving and robotics. Classic tasks for scene understanding mainly include object detection, instance segmentation and semantic segmentation. This paper considers a recently proposed task named *panoptic segmentation* [19], which aims at finding all foreground (FG) objects (named *things*, mainly including countable targets such as *people*, *animals*, *tools*, *etc.*) at the instance level, meanwhile parsing the background (BG) contents (named *stuff*, mainly including amorphous regions of similar texture and/or material such as *grass*, *sky*, *road*, *etc.*) at the semantic level. The benchmark algorithm [19] and MS-COCO panoptic challenge winners [1] dealt with this task by directly combining FG instance segmentation models [12] and BG scene parsing [38] algorithms, which ignores the underlying relationship and fails to borrow rich contextual cues between *things* and *stuff*, *e.g.*, *people* are more likely to stand on *grass*, instead of *trees*, although they often have similar appearances.



(a) Input Image (b) Panoptic Segmentation (c) Foreground: *things* (d) Background: *stuff*

Figure 1. Given an image 1(a), the goal of panoptic segmentation 1(b) is to find FG *things* at the instance level 1(c) and BG *stuff* at the semantic level 1(d). The *things* of the same class share the same color family but appear in different intensities. **All these results are generated from the proposed approach.**

In this paper, we present a conceptually simple and unified framework for panoptic segmentation. To facilitate information flow between FG *things* and BG *stuff*, we combine conventional instance segmentation and semantic segmentation networks, leading to an unified network with two branches. This strategy brings an immediate improvement in segmentation accuracy as well as higher efficiency in computation (because the network backbone can be shared). This implies that panoptic segmentation benefits from complementary information provided by FG objects and BG contents, which lays the foundation of our approach.

Going one step further, we explore the possibility of integrating higher-level visual cues (*i.e.*, beyond the features extracted from the end of the backbone) towards the more accurate segmentation. This is achieved via two attention-based modules working at the object level and the pixel level, respectively. For the first module, we refer to the regional proposals, each of which indicates a possible FG *thing*, and adjusts the probability of the corresponding region to be considered as FG *things* and BG *stuff*. For the second module, we take out the FG segmentation mask, and use it to refine the boundary between FG *things* and BG *stuff*. In the context of deep networks, these two modules, named the Proposal Attention Module (PAM) and Mask Attention Module (MAM), respectively, are implemented as additional connections across FG and BG branches. Within MAM, a new layer named *RoIUpsample* is designed to define an accurate mapping function between pixels in the fixed-shape FG mask and the corresponding feature map. In practice, all additional connections go from the FG branch to the BG branch, mainly due to the observation that FG segmentation is often more accurate[1]. Furthermore, BG *stuff*, while being refined by FG *things*, also gives feedback via gradients. Consequently, both FG and BG segmentation accuracies are considerably improved.

The overall approach, named Attention-guided Unified Network (**AUNet**), can be easily instantiated to various network backbones, and optimized in an end-to-end manner. We evaluate AUNet in the popular segmentation benchmark, namely, the MS-COCO dataset [24], and claim **the state-of-the-art performance** in terms of PQ, a standard metric integrating accuracies of both *things* and *stuff* [19]. In addition, the benefits brought by joint optimization and two attention-based modules are verified through an extensive ablation study 4.2.

The major contribution of this research is to present a conceptually simple and unified framework for both FG and BG segmentation, which reaches the top performance in MS-COCO dataset [24]. Furthermore, this work also investigate the complementary information delivered by FG objects and BG contents. While panoptic segmentation serves as a natural scenario of studying this topic, its application lies in a wider range of visual tasks. Our solution, AUNet, is a preliminary exploration in this field, yet we look forward to more efforts along this direction.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 elaborates the proposed AUNet, including two attention-based modules. After experiments are shown in Section 4, we conclude this work in Section 5.

[1]We find the *pixel accuracy* of *things* is much higher (6.7 % absolute gap) than that of *stuff*, when considering instance with the same semantic as one category, *e.g.*, all individuals are evaluated as *person* in testing. We evaluate them on the same MS-COCO semantic evaluation metric.

2. Related Work

Traditional deep learning based scene understanding researches often focused on foreground and background targets separately [12, 38]. Recently, the rapid progress in object detection [10, 11, 28] and instance segmentation [6, 12, 21, 25] made it possible to achieve object localization and segmentation at a finer level. Meanwhile, the development of semantic segmentation [4, 5, 27, 38] boosted the performance of scene parsing. Despite their effectiveness, the separation of these tasks caused the lack of contextual information in instance segmentation as well as the confusion brought by individuals in semantic segmentation. To bridge this gap, recently, researchers proposed a new task named *panoptic segmentation* [19], which aims at accomplishing both tasks (FG instance and BG semantic segmentation) simultaneously.

Panoptic Segmentation: In [19], the author gave a benchmark of panopic segmentation by combining instance and semantic segmentation models. Later, a weakly-supervised method [20] was proposed on top of initialized semantic results, and an end-to-end approach [8] was designed to combine both FG and BG cues. However, their performance is far from the benchmark [19]. Different from these works, our proposed AUNet achieves the top performance in an end-to-end framework. Furthermore, we also establish the bond between proposal-based instance and FCN based semantic segmentation.

Instance Segmentation: Instance segmentation aims at discriminating different instances of the same object. There are mainly two streams of methods to solve this task, namely, proposal-based methods and segmentation-based methods. Proposal-based methods, with the help of accurate regional proposals, often achieved higher performance. Recent examples include MNC [6], FCIS [21], Mask R-CNN [12] and PANet [25]. Moreover, segmentation-based methods aggregated pixel-level cues to compose instances based on pre-computed semantic segmentation [2, 22, 26] or depth ordering [37] results.

Semantic Segmentation: With the development of so-called encoding-decoding networks such as FCN [27], rapid progress has been made in semantic segmentation [4, 5, 38]. In segmentation, capturing contextual information plays a vital role, for which various approaches were proposed including ASPP used in DeepLab [4, 5] for multi-scale contexts, DenseASPP [34] for global contexts, and PSPNet [38] which collected contextual priors. There were also efforts to use attention modules for spatial feature selection, such as [9, 35, 36], which will be detailed discussed in the next paragraph.

Attention-based Modules: Attention-based module have been widely applied in visual tasks, including image processing and video understanding. In particular, SENet [16]

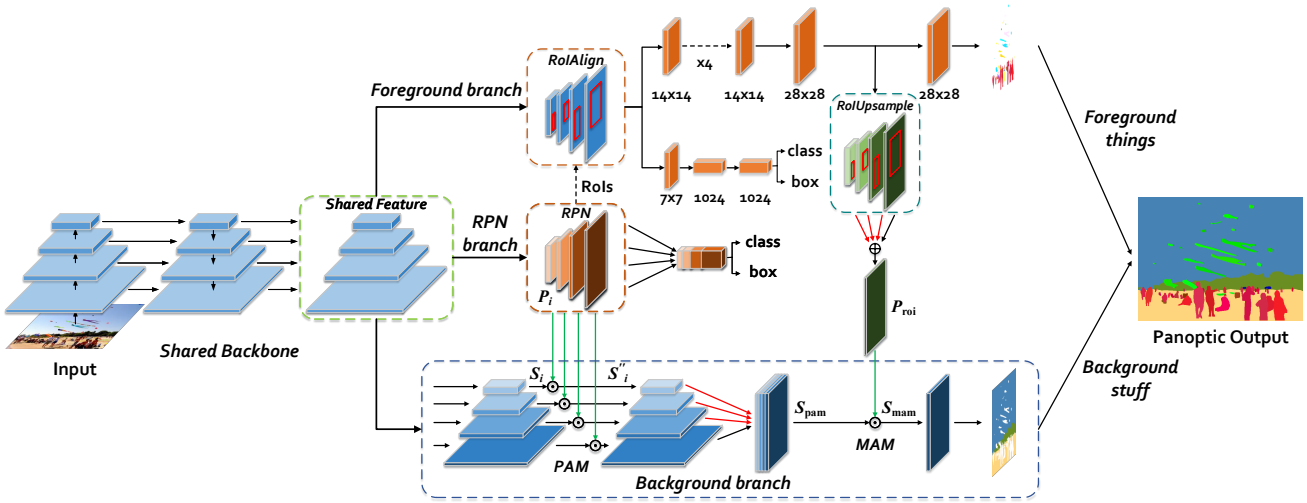


Figure 2. The proposed network structure. We adopt FPN as our backbone and share features with three parallel branches, namely *foreground branch*, *background branch*, and *RPN branch*. In training stage, the network is optimized in an end-to-end manner. In inference stage, panoptic results are generated by *things* and *stuff* results following the method in Section 3.4. "$\oplus$" denotes element-wise sum and "$\odot$" represents Proposal Attention Module (PAM) or Mask Attention Module (MAM) according to its position. PAM and MAM model the complementary relation between two branches. Details of PAM and MAM refer to Figure 3 and Figure 5. The red and green lines represent upsample and attention operations respectively.

formulated channel-wise relationships via an attention-and-gating mechanism, non-local network [31] bridged self-attention for machine translation [30] to video classifiaction using non-local filters. In the scope of scene understanding, [35] and [36] aggregated global contextual information as well as class-dependent features by channel-attention operations. More recently, self-attention and channel attention were adopted by [9] to model long-range contexts in the spatial and channel dimensions, respectively. In this work, we establish the relationship between foreground *things* and background *stuff* in panoptic segmentation with a series of coarse-to-fine attention blocks.

3. Attention-guided Unified Network

3.1. Problem and Baselines

Panoptic segmentation task aims at understanding everything visible in one view, which means each pixel of an image must be assigned a semantic label and an instance ID. To address this issue, the existing top algorithms [1, 19] directly combined the instance and semantic results from separate models, such as Mask R-CNN [12] and PSPNet [38].

We formulate the problem of panoptic segmentation as recognizing and segmenting all FG *things* and understanding all BG *stuff*. In this way, we solve the problem from two aspects, namely foreground branch and background branch in an unified network (Figure 2). In detail, given an input image $\boldsymbol{X}$, our goal is to generate FG *things* result $\boldsymbol{Y}_{\mathrm{Th}}$ and BG *stuff* result $\boldsymbol{Y}_{\mathrm{St}}$ simultaneously. Thus, the panoptic result $\boldsymbol{Y}_{\mathrm{Pa}}$ can be generated from $\boldsymbol{Y}_{\mathrm{Th}}$ and $\boldsymbol{Y}_{\mathrm{St}}$ directly using the fusion method in Section 3.4. The performance of panoptic results is evaluated by panoptic quality (PQ) [19] as described in Section 4.1. For this purpose, we firstly introduce our unified framework for panoptic segmentation in this section. Then, key elements in our designed attention-guided modules are elaborated, including proposal attention module (PAM) and mask attention module (MAM). Finally, we give our implementation details.

In this work, we view the method, in which *things* and *stuff* are generated from separate models, as our baseline. Specifically, the baseline method gives the result of *things* $\boldsymbol{Y}_{\mathrm{Th}}$ and *stuff* $\boldsymbol{Y}_{\mathrm{St}}$ from separate models $\mathbb{M}_{\mathrm{Th}}$ and $\mathbb{M}_{\mathrm{St}}$ respectively. And the FG model $\mathbb{M}_{\mathrm{Th}}$ and BG model $\mathbb{M}_{\mathrm{St}}$ are given the similar backbones (*e.g.*, FPN [23]) for the following unified framework.

3.2. Unified Framework

In order to bridge the gap between FG *things* with BG *stuff*, we propose the *Attention-guided Unified Network* (AUNet). Comparing with the baseline approach, the proposed AUNet fuses two models ($\mathbb{M}_{\mathrm{Th}}$ and $\mathbb{M}_{\mathrm{St}}$) together by sharing the same backbone and generates $\boldsymbol{Y}_{\mathrm{Th}}$ and $\boldsymbol{Y}_{\mathrm{St}}$ from parallel branches. As clearly illustrated in Figure 2, the AUNet is conceptually simple: FPN is adopted as the backbone to extract discriminative features from different scales and shared by all the branches.

Different from traditional approaches, which directly combine results from $\mathbb{M}_{\mathrm{Th}}$ and $\mathbb{M}_{\mathrm{St}}$, the proposed AUNet optimizes them using a joint loss function $\mathcal{L}$ (defined in Section 3.4) and facilitates both tasks in an unified framework.

In detail, we adopt a proposal-based instance segmentation module to generate finer masks M in *foreground branch*. And for *background branch*, light heads are designed to aggregate scene information from shared multi-scale features. In this way, the shared backbone is supervised by FG *things* and BG *stuff* simultaneously, which promotes the connection between two branches in feature space. In order to build up the bond between FG objects and BG contents more explicitly, two sources of attention modules are added. We consider the coarse attention operation between the i-th scale BG feature map with the corresponding RPN feature map, denoted by S_i and P_i respectively. The attention module can be formulated as $S_i \odot P_i$, where "$\odot$" denotes attention operations, as illustrated in Figure 2. Furthermore, the finer relationship is established by the attention between the processed feature map S_{pam} and the generated FG segmentation mask P_{roi}, which can be formulated as $S_{\text{pam}} \odot P_{\text{roi}}$. Details will be investigated in the following section.

3.3. Attention-guided Modules

Considering the complementary relationship between FG *things* and BG *stuff*, we introduce features from *foreground branch* to *background branch* for more contextual cues. From another perspective, the attention operation connecting two branches also establishes a bond between proposal-based method and FCN-based method segmentation. To this end, two spatial attention modules are proposed, namely proposal attention module (PAM) and mask attention module (MAM).

3.3.1 Proposal Attention Module

In classic two-stage detection frameworks, region proposal network (RPN) [28] is introduced to give predicted binary class labels (foreground and background) and bounding-box coordinates. This means RPN features contain rich background information which can only be obtained from

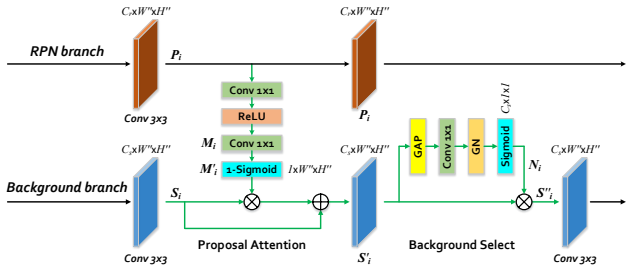


Figure 3. The designed proposal attention module (PAM) for complementary relationship establishment. We adopt this block in each scale of shared features, where W'' and H'' changes in each scale. Here, "$\otimes$" denotes spatial element-wise multiplication and "$\oplus$" denotes element-wise sum. The green lines represent operations in PAM. GAP and GN indicate Global Average Pooling and Group Normalization [32], respectively.

stuff annotations in background branch. Therefore, we propose a new approach to establish the complementary relationship between FG elements and BG contents, called Proposal Attention Module (PAM). As shown in Figure 3, we utilize contextual cues from RPN branch for attention operation. Here, we give a detailed formulation for this process. Given an input feature map $P_i \in \mathbb{R}^{C_r \times W'' \times H''}$ from the i-th scale RPN branch, the FG mask M_i before sigmoid activation can be formulated as:

$$M_i = f(\sigma(f(P_i, w_{i,1})), w_{i,2}) \tag{1}$$

where $f(\cdot,\cdot)$ denotes a convolution function, σ represents the ReLU activation function, $M_i \in \mathbb{R}^{1 \times W'' \times H''}$ means the generated FG mask, both $w_{i,1} \in \mathbb{R}^{C_r' \times C_r \times 1 \times 1}$ and $w_{i,2} \in \mathbb{R}^{1 \times C_r' \times 1 \times 1}$ indicate convolutional parameters.

To emphasize the background contents, we formulate the attention mask M_i' as $1 - \text{sigmoid}(M_i)$. Then, the i-th scale activated feature map $S_i' \in \mathbb{R}^{C_s \times W'' \times H''}$ can be presented as:

$$S'_{i,j} = S_{i,j} \otimes M_i' \oplus S_{i,j} \tag{2}$$

where $\otimes$ and $\oplus$ denotes element-wise multiplication and sum respectively, $S_{i,j}$ means the j-th layer of semantic feature map $S_i \in \mathbb{R}^{C_s \times W'' \times H''}$.

Furthermore, background select function is designed to filter out useless background layers after attention operation. The selected feature map $S_i'' \in \mathbb{R}^{C_s \times W'' \times H''}$ can be generated as:

$$N_i = \text{sigmoid}(\text{GN}(f(\text{G}(S_i'), w_{i,3}))) \tag{3}$$

$$S''_{i,k} = S'_{i,k} \otimes N_i \tag{4}$$

where G and GN denotes global average pooling and group norm [32] respectively, $N_i \in \mathbb{R}^{C_s \times 1 \times 1}$ means selecting operator, $w_{i,3} \in \mathbb{R}^{C_s \times C_s \times 1 \times 1}$ represents convolutional parameter, and $S'_{i,k}$ indicates the k-th pixel channel in S'_i.

Based on the above formulation of PAM, we highlight the background regions in the shared feature maps via attention operation and background select function. It also facilitates the learning of *things* in turn by enhancing the weights of activated foreground regions during backpropagation (see Section 4.2).

3.3.2 Mask Attention Module

With the introduction of contextual cues by PAM, background branch is encouraged to focus more on the regions of *stuff*. However, the predicted coarse areas from RPN branch lack enough cues for precise BG representations. Unlike RPN features, the $m \times m$ fixed-shape masks generated from foreground branch encode finer FG layouts. Thus, we propose Mask Attention Module (MAM) to further model the

relationship, as illustrated in Figure 5. Consequently, the $1 \times W' \times H'$ shape FG segmentation mask is needed for similar attention operations as before. Now, the problem is: how to reproduce the $W' \times H'$ shape FG feature map from $m \times m$ masks?

RoIUpsample: In order to solve the size mismatching problem, we propose a new differentiable layer called *RoIUpsample*. Specifically, RoIUpsample is designed similar to the inverse process of RoIAlign [12], as clearly illustrated in Figure 4. In the RoIUpsample layer, the $m \times m$ mask (m equals to 14 or 28 in Mask R-CNN) is firstly reshaped to the same size of RoIs (generated from RPN). Then we utilize the designed inverse bilinear interpolation to compute values of the output features at four regularly sampled locations (same with RoIAlign) in each mask bin, and then sum up the final results as the generated mask feature map. To meet the requirement of bilinear interpolation [18], in which near points are given more contributions, an operation for *inverse* bilinear interpolation is formulated:

$$\begin{cases} R(\mathrm{p}_{1,1}) = \frac{(1-x_p)(1-y_p)}{\mathrm{value}_x \times \mathrm{value}_y} R(\mathrm{p}_g) \\ R(\mathrm{p}_{1,2}) = \frac{(1-x_p)y_p}{\mathrm{value}_x \times \mathrm{value}_y} R(\mathrm{p}_g) \\ R(\mathrm{p}_{2,1}) = \frac{x_p(1-y_p)}{\mathrm{value}_x \times \mathrm{value}_y} R(\mathrm{p}_g) \\ R(\mathrm{p}_{2,2}) = \frac{x_p y_p}{\mathrm{value}_x \times \mathrm{value}_y} R(\mathrm{p}_g) \end{cases} \quad (5)$$

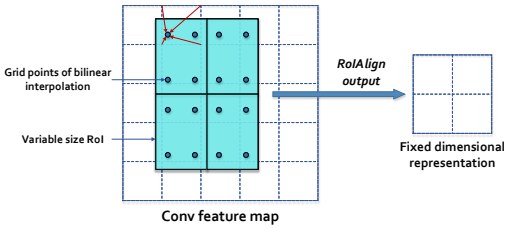
where $R(\mathrm{p}_{j,k})$ denotes the result of point $\mathrm{p}_{j,k}$ after inverse bilinear interpolation, $R(\mathrm{p}_g)$ here equals to one quarter of the corresponding value in the input mask, and normalized weights value_x, value_y are defined as:

$$\mathrm{value}_x = x_p^2 + (1 - x_p^2), \mathrm{value}_y = y_p^2 + (1 - y_p^2) \quad (6)$$

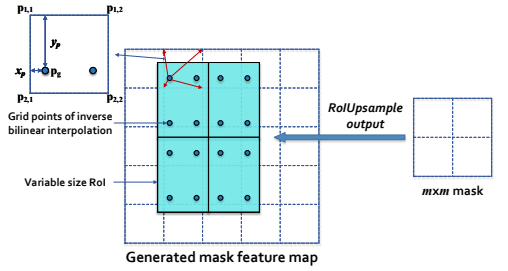
in which x_p and y_p indicate the distance between grid point p_g and generated $\mathrm{p}_{1,1}$ in two axes respectively, as presented in Figure 4(b). Note that with the designed equation 5 and 6, the $m \times m$ mask can also be reverted from the generated $W' \times H'$ feature map with the *forward* bilinear interpolation.

Then, the generated feature map is assigned to four different scales according to the size of RoIs, which is similar with that in FPN [23]. Consequently, the generated FG feature map is achieved for the following operations.

Attention Operation: Different from traditional instance segmentation tasks, the predicted FG masks are utilized to give background branch more contextual guidance in pixel-level. We firstly aggregate them together to the $C_m \times W' \times H'$ feature map using RoIUpsample, as presented in Figure 5. Then, the finer $1 \times W' \times H'$ activated BG regions can be produced, similar with that in PAM. With the introduction of attention, the FG masks is also supervised by semantic loss function, which enables a further improvement in scene understanding (both for *things* and *stuff*), as discussed in Section 4.2. A similar background select function is adopted to aggregate useful highlighted background



(a) RoIAlign process



(b) RoIUpsample process

Figure 4. Comparison between RoIAlign [12] and our proposed RoIUpsample. The designed RoIUpsample, which can be viewed as an *inverse* operation of RoIAlign, reverts the feature map from FG masks according to their accurate spatial locations. Here, we give an example of RoIAlign output and RoIUpsample input when $m = 2$ for intuitive illustration.

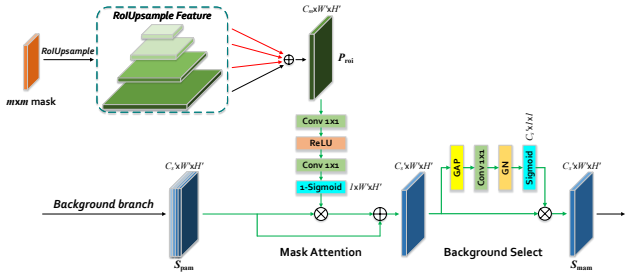


Figure 5. The proposed mask attention module (MAM) for a finer relationship modelling. Here, "$\otimes$" denotes spatial element-wise multiplication and "$\oplus$" denotes element-wise sum. The red and green lines represent upsample and operations in MAM respectively. GAP and GN are identical with that in PAM.

features. Consequently, we model the complementary relationship between FG *things* and BG *stuff* with the proposed PAM and MAM.

3.4. Implementation Details

In this section, we give more implementation details on the training and inference stage of our proposed AUNet.

Training: As well elaborated in Section 3.2, all of our proposed methods are trained in an unified framework. The whole network is optimized via a joint loss function $\mathcal{L}$ during training stage:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{box} + \lambda_3 \mathcal{L}_{mask} + \lambda_4 \mathcal{L}_{seg} \quad (7)$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{box}$, $\mathcal{L}_{mask}$, and $\mathcal{L}_{seg}$ denotes the loss function of object classification, bonding-box regression, instance segmentation, and semantic segmentation, respectively. In our experiments, λ_1 to λ_4 are designed to balance training processes, where λ_1 to λ_3 are set to 1 and λ_4 is set to 0.3 for the best performance.

In details, we adopt ResNet-FPN [14, 23] as our backbone. And the hyper-parameters in the foreground branch are set following Mask R-CNN [12]. The backbone is pre-trained on ImageNet [29], and the remaining parameters are initialized following [13]. As standard practice [7, 14, 23], 8 GPUs are used to train all the models. Each mini-batch has 2 images per GPU for ResNet-50 and ResNet-101 based networks and 1 image per GPU for the others. The networks are optimized for 18 epochs using mini-batch stochastic gradient descent (SGD) with a weight decay of 0.00004 and a momentum of 0.9. As for learning rate, it is initialized with 0.02 for the first 13 epochs and divided by 10 at 15-th and 18-th epoch respectively. Batch Normalization [17] in the backbone is fixed and Group Normalization [32] is added to all of the branches in our final results. For data augmentation, input images are reshaped to the scale with a 600 pixels short edge for MS-COCO dataset [24]. And the input images are also horizontally flipped in random.

Inference: The panoptic results are produced in inference stage by fusing the results of FG *things* and BG *stuff* in a similar way with that in [19]. In this stage, the overlaps of *things* are first resolved in a NMS-like procedure which predicts the segments with higher confidence scores. And the relationships among categories are also considered during this procedure. For example, *ties* should not be overlapped by *person* in the final result. Then, the non-overlapping instance segments are combined with *stuff* results by assigning instance label first in favor of the *things*.

4. Experiments

In this section, our approach is evaluated on Microsoft COCO [24] dataset. We first give description of the dataset as well as the evaluation metrics. Then we evaluate our method and give detailed analyses. Comparison with the state-of-the-art methods in panoptic segmentation are presented at last.

4.1. Dataset and Metrics

Dataset: Due to the novelty of panoptic task itself, there are few datasets with detailed panoptic annotations as well as public evaluation metrics. Microsoft COCO [24] is the most suitable and challenging one for the new panoptic segmentation task, for its high annotation quality and high data complexity. It consists of 115k images for training and 5k images for validation, as well as 20k images for *test-dev* and 20k images for *test-challenge*. MS-COCO panoptic annotations includes 80 *thing* categories and 53 *stuff* categories. We train our models on *train* set with no extra data and reports results on *val* set and *test-dev* set for comparison.

Evaluation Metrics: We adopt the evaluation metrics introduced by [19], which computes *panoptic quality* (PQ) metric for evaluation. PQ can be explained as the multiplication of a *segmentation quality* (SQ) and a *recognition quality* (RQ) term:

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}\,(p,g)}{|TP|}}_{\text{segmentation quality(SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality(RQ)}} \tag{8}$$

where $\mathrm{IoU}(p,g)$ means the intersection-over-union between predicted object p and ground truth g, true positives (TP) denotes matched pairs of segments ($\mathrm{IoU}(p,g) > 0.5$), false positives (FP) represents unmatched predicted segments, and false negatives (FN) means unmatched ground truth segments. PQ, SQ, and RQ of both *thing* and *stuff* are also reported in our results.

4.2. Component-wise Analysis and Diagnosis

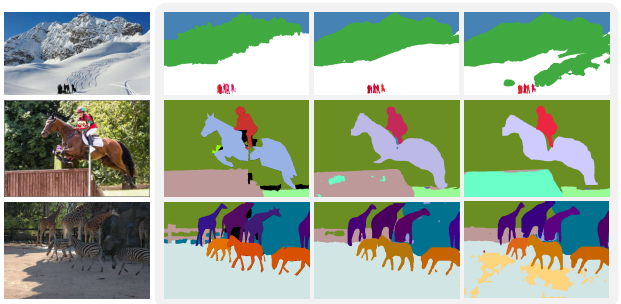
In this section, we will decompose our approach step-by-step to reveal the effect of each component. All experiments in this section are trained and evaluated on MS-COCO dataset in a single model with no extra data. Here, we adopt ResNet-50-FPN as our backbone. For fair comparison, we strictly follow the merging method in [19] with no trick in training and inference stage when doing component-wise analyses. As presented in Table 1, our proposed AUNet achieve an absolute improvement of $\mathbf{2.4}\%$ in PQ when compared with separate training method.

4.2.1 Unified Framework

As elaborated in Section 3.2, our proposed unified framework deals with FG *things* and BG *stuff* in parallel branches. As shown in Table 1, the unified framework boosts up the performance both in $\mathrm{PQ}^{\mathrm{St}}$ and $\mathrm{PQ}^{\mathrm{Th}}$, which brings $\mathbf{1.1}\%$

Table 1. Comparison among different settings of panoptic quality (%) on the MS-COCO dataset (based on ResNet-50-FPN). "e2e" denotes using unified framework for training and inference. "select" means using background select function in PAM and MAM. $\mathrm{PQ}^{\mathrm{St}}$ and $\mathrm{PQ}^{\mathrm{Th}}$ indicates PQ result for *stuff* and *thing* respectively.

Method	e2e	PAM	MAM	select	PQ	$\mathrm{PQ}^{\mathrm{St}}$	$\mathrm{PQ}^{\mathrm{Th}}$
sep	✗	✗	✗	✗	37.2	22.8	47.1
e2e	✓	✗	✗	✗	38.3	23.9	47.9
PAM	✓	✓	✗	✗	39	24.5	48.5
$\mathrm{PAM}_{\mathrm{s}}$	✓	✓	✗	✓	39.4	25.2	48.9
MAM	✓	✗	✓	✗	38.9	24.2	48.6
$\mathrm{MAM}_{\mathrm{s}}$	✓	✗	✓	✓	39.2	24.9	48.6
All	✓	✓	✓	✓	**39.6**	**25.2**	**49.1**



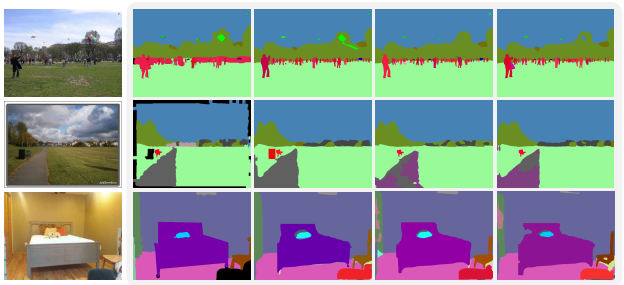
(a) Input Image (b) Ground Truth (c) Unified Net (d) Separate Net

Figure 6. Comparison between the proposed unified framework and separate models in MS-COCO *val* set. The unified framework produces more accurate predictions. For example, in the first row, the *snow* areas (white pixels) surrounding persons are misclassified to the *mountain* (green pixels) when generating results from separate models, but this is corrected in the unified framework. The same conclusion can be drawn from other rows.

absolute improvements in PQ. Given an intuitive presentation in Figure 6, the proposed method considers the image in a more holistic view. This can be attributed to the shared backbone and joint optimization, with which the network is supervised to focus on more discriminative features for both *things* and *stuff*. Due to the lack of holistic consideration, the results from separate models have massive confusion areas, especially in *stuff*, as shown in Figure 6. With the shared backbone, the misclassification in *stuff* are effectively reduced and the *things* are given more details.

4.2.2 Proposal Attention Module

The proposed PAM builds the complementary relationship between *things* and *stuff* from different scales. By this way, the binary-classified RPN branch is optimized under the supervision of semantic labels. The optimization of RPN branch enables a further enhancement in FG classification, as presented in the first row of Figure 7 where more FG objects are found out using PAM. With the bond between *stuff* and *things* established, the network performs consistent gain in PQ^{St} and PQ^{Th}, as presented in Table 1. The background select function proves its effectiveness in PQ^{St}. This can be resulted from the global contextual features introduced by global average pooling in Equation 3, which means it chooses to aggregate highlighted BG features under the guidance of global context. The second row images in Figure 7 confirm this, in which PAM_s eliminates the confusion brought by PAM (especially in gray *pavement* region). It is worth noting that we have tried other fusion methods for FG and BG feature fusion, such as concatenation and direct summary after transformation. But these strategies have minor contributions, which means the attention is more appropriate for relationship establishment.



(a) Input Image (b) Ground Truth (c) PAM_s (d) PAM (e) Unified Net

Figure 7. Comparison between the proposed PAM and the raw unified framework in MS-COCO *val* set. PAM_s denotes using background selection in PAM. PAM_s and PAM effecitvely improve the performance on both *things* and *stuff*. The performance of PAM_s even surpasses human annotation in the first row, *e.g.*, *persons* (pixels in red color family) and *kites* (pixels in light green color family) are much finer than ground truth.

4.2.3 Mask Attention Module

While the PAM establishes the bond between FG objects and BG contents, the MAM gives background finer representations, as elaborated in Section 3.3.2. As that in PAM, MAM also achieves better performance over the raw method in both PQ^{St} and PQ^{Th}. However, the contribution of MAM is slightly lower than PAM. We guess this is caused by the lack of contextual cues in the generated FG segmentation mask.[2] In fact, we also evaluate the performance when adopting different resolution masks for RoIUpsample, namely the 14×14 mask and the 28×28 one. The result shows the high resolution mask features bring a further gain (0.1% absolute improvement in PQ) over the smaller one. This is reasonable, because RoIUpsample layer generates finer layouts if given higher resolution masks. With the help of background select function, MAM_s achieves 39.2% in PQ. More intuitive results can be found in Figure 8.

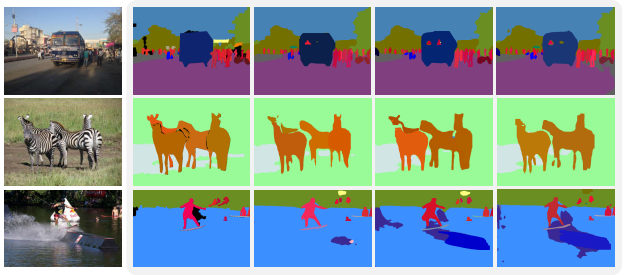
4.3. Comparison to State-of-the-Arts

We compare our proposed network with other state-of-the-art methods on MS-COCO [24] *test-dev* subset. As shown in Table 2, the proposed AUNet achieves the leading PQ performance $\mathbf{46.5}\%$ in MS-COCO dataset without bells-and-whistles. In details, winners of COCO2018 panoptic challenge [1] adopt numerous additional network enhancements during training and inference stage, *e.g.*, abundant extra data (110k external annotated MS-COCO images), multi-scale training, model ensemble. Moreover, considering the network enhancements adopted by the winner teams, cascade R-CNN [3], additional SE blocks [16] are adopted to solve *things*, and extra blocks or label

[2] We adopt zero padding for vacant areas in RoIUpsample layer, resulting in blank BG context. This needs to be investigated in the future works.

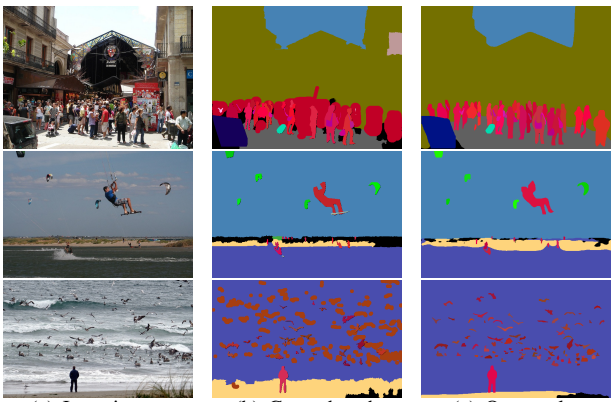
Table 2. Panoptic quality (%) on MS-COCO 2018 *test-dev*. "extra data" here denotes using extra dataset for training, "e2e" represents using an unified framework for *things* and *stuff* prediction, and "enhance$_{\text{Th}}$" and "enhance$_{\text{St}}$" indicates using additional enhancement techniques in network heads for *things* and *stuff* respectively. PQ$^{\text{Th}}$ and PQ$^{\text{St}}$ means PQ result for *things* and *stuff* respectively. We report our results in one single model with *no* extra data and network enhancements.

Method	backbone	extra data	e2e	enhance$_{\text{Th}}$	enhance$_{\text{St}}$	PQ	SQ	RQ	PQ$^{\text{Th}}$	SQ$^{\text{Th}}$	RQ$^{\text{Th}}$	PQ$^{\text{St}}$	SQ$^{\text{St}}$	RQ$^{\text{St}}$
Megvii (Face++)	ensemble model	✓	✗	✓	✓	53.2	83.2	62.9	62.2	85.5	72.5	39.5	79.7	48.5
Caribbean	ensemble model	✗	✗	✓	✓	46.8	80.5	57.1	54.3	81.8	65.9	35.5	78.5	43.8
PKU_360	ResNeXt-152-FPN	✗	✗	✓	✓	46.3	79.6	56.1	58.6	83.7	69.6	27.6	73.6	35.6
JSIS-Net [8]	ResNet-50	✗	✓	✗	✗	27.2	71.9	35.9	29.6	71.6	39.4	23.4	72.3	30.6
Ours	ResNet-101-FPN	✗	✓	✗	✗	45.2	80.6	54.7	54.4	83.3	64.8	31.3	76.6	39.4
Ours	ResNet-152-FPN	✗	✓	✗	✗	45.5	80.8	55.0	54.7	83.4	65.2	31.6	76.9	39.7
Ours	ResNeXt-152-FPN	✗	✓	✗	✗	**46.5**	**81.0**	**56.1**	**55.9**	**83.7**	**66.3**	**32.5**	**77.0**	**40.7**



(a) Input Image (b) Ground Truth (c) MAM$_s$ (d) MAM (e) Unified Net

Figure 8. Comparison between the proposed MAM and the raw unified framework in MS-COCO *val* set. MAM$_s$ represent adopting background select function. MAM$_s$ and MAM give finer FG and BG layouts than the raw unified framework, *e.g.*, in the second row, MAM$_s$ and MAM characterize four and three *zebras* (pixels in brown color family) respectively, while the raw framework only finds two. The finer BG representations are reflected in other rows.



(a) Input image (b) Ground truth (c) Our results

Figure 9. Example results of AUNet on MS-COCO *val* set. Our performance on *things* 9(c) are even better than human annotation 9(b). The *things* of the same class share the same color family but appear in different intensities.

bank [15] are added for *stuff* as well. Different from them, the proposed AUNet achieves the top performance in an unified framework with no extra data or additional network enhancement for both *things* and *stuff*. To be more specific, only one single model based on the ResNeXt-152-FPN[3] is adopted in the AUNet.

When it comes to the performance on *things* and *stuff*, we achieve the $\mathbf{55.9}\%$ on PQ$^{\text{Th}}$ and $\mathbf{32.5}\%$ on PQ$^{\text{St}}$, which means FG *things* perform better than BG *stuff* especially in recognition quality (RQ). Comparing with other top algorithms without extra data, we achieve the best performance in SQ and SQ$^{\text{Th}}$ without network enhancement skills added. Filtering out the improvement bring by model ensemble, we compare our AUNet with "PKU_360" team who adopted a similar backbone but with additional skills. The result shows that our algorithm perform better than them especially in PQ$^{\text{St}}$, for about $\mathbf{5}\%$ absolute improvements. Furthermore, the AUNet overpasses the former end-to-end method, namely JSIS-Net [8], with a $\mathbf{19.3}\%$ absolute gap, which proves the effectiveness of the proposed method. In Table 2, it is clear that the AUNet have a great balance between *things* and *stuff*, even when comparing with the challenge winners (no extra data). This is due to the introduction of unified framework and attention-guided modules for complementary relationship establishment, as well elaborated in Section 4.2. Figure 9 gives intuitive presentations of the top performance using our proposed AUNet.

5. Conclusions

This paper presents AUNet, an unified framework for panoptic segmentation. The key difference from prior approaches lies in that we unify FG (instance-level) and BG (semantic-level) segmentation into one model, so that the FG branch, often being better optimized, can assist the BG branch via two sources of attentions (*i.e.*, proposal attention module and mask attention module), which offer object-level and pixel-level guidance, respectively. In experiments, we observe consistent accuracy gain in MS-COCO, based on which new state-of-the-arts are achieved (*no extra data*).

Our research delivers an important message: in visual tasks, it is often beneficial to partition targets into a few

[3]We use the 64×4d variant of ResNeXt [33] with deformable conv [7] and non-local blocks [31].

subclasses according to their properties, so that complementary information can be propagated across subclasses to assist scene understanding. Panoptic segmentation, being a new task, offers a natural partition between FG *things* and BG *stuff*, yet more possibilities remain unexplored and to be studied in the future.

Appendix A. Timing

We evaluate the proposed AUNet (based on ResNet-50-FPN) on an Nvidia Titan X GPU. Our model reaches **10 fps** in the inference stage, which claims potential applications in real-time panoptic segmentation.

Appendix B. Results Visualization

We provide additional results on MS-COCO *val* set. As presented in Figure 10, the proposed AUNet achieves impressive performance in both *things* and *stuff*.

Appendix C. Activated Mask Visualization

As we have elaborated in Section 3.3 of the main article, the Proposal Attention Module (PAM) and Mask Attention Module (MAM) bring *contextual* cues to *background branch* using the attention mechanism. Here, we provide intuitive visualization of the *activated background masks* in PAM and MAM. For clear presentation, the heatmaps with the highest resolution (the 4*th* scale, M_4') in PAM are shown in the second row of Figure 11. In addition, the finer activated background regions (the output of the $1 - \text{sigmoid}$ operation) are presented to illustrate the effectiveness of the proposed MAM. Equivalent to the analyses in Section 3.3 of the main article, the activated masks M_4' in PAM focus more on the background regions, which bring contextual cues to *background branch*. As for the activated masks in MAM, all background areas are given the same weight, which is caused by the zero-padding in *RoIUpsample* layer. Con-

Figure 10. Example results of AUNet on MS-COCO *val* set. The *things* of the same class share the same color family, but appear in different intensities.

Figure 11. Heatmaps of activated background mask in PAM (the 4*th* scale, M_4') and MAM. The red regions are assigned more weights while the blue regions less weights in the *background branch*. All the input images are sampled from the MS-COCO *val* set.

sequently, PAM produces *coarse* background regions with abundant contextual cues while MAM complements *finer* background layouts. Thus, they should be combined in the attention module, as designed in the AUNet (see Figure 2).

References

[1] COCO: Panoptic Leaderboard. http://cocodataset.org/#panoptic-leaderboard. 1, 3, 7

[2] A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 2017. 2

[3] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 7

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 2018. 2

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 2

[6] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 2

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, 2017. 6, 8

[8] D. de Geus, P. Meletis, and G. Dubbelman. Panoptic segmentation with a joint semantic and instance segmentation network. *arXiv:1809.02110*, 2018. 2, 8

[9] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. *arXiv:1809.02983*, 2018. 2, 3

[10] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 3, 5, 6

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 6

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[15] H. Hu, Z. Deng, G.-T. Zhou, F. Sha, and G. Mori. Labelbank: Revisiting global perspectives for semantic segmentation. *arXiv:1703.09891*, 2017. 7

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 7

[17] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6

[18] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 5

[19] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. *arXiv:1801.00868*, 2018. 1, 2, 3, 6

[20] Q. Li, A. Arnab, and P. H. Torr. Weakly-and semi-supervised panoptic segmentation. In *ECCV*, 2018. 2

[21] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. *arXiv:1611.07709*, 2016. 2

[22] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. Proposal-free network for instance-level object segmentation. *arXiv:1509.02636*, 2015. 2

[23] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3, 5, 6

[24] 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 *ECCV*, 2014. 2, 6, 7

[25] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 2

[26] Y. Liu, S. Yang, B. Li, W. Zhou, J. Xu, H. Li, and Y. Lu. Affinity derivation and graph merge for instance segmentation. In *ECCV*, 2018. 2

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 4

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[31] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 3, 8

[32] Y. Wu and K. He. Group normalization. In *ECCV*, 2018. 4, 6

[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 8

[34] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2

[35] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. *arXiv:1804.09337*, 2018. 2, 3

[36] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 2, 3

[37] Z. Zhang, S. Fidler, and R. Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *CVPR*, 2016. 2

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2, 3