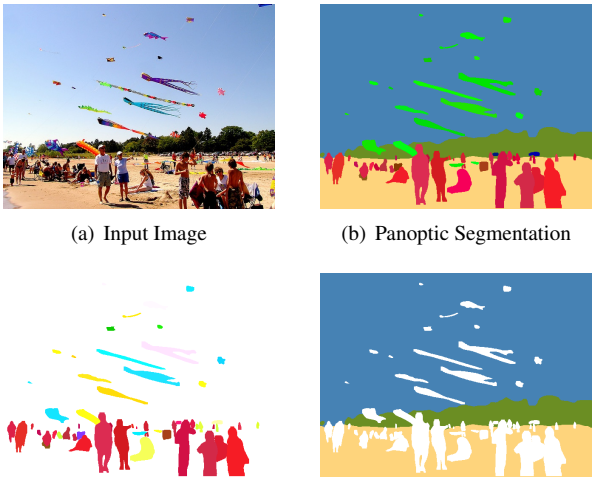
Attention-guided Unified Network for Panoptic Segmentation

注意力指导的联合网络做全景分割

Yanwei Li[1,2], Xinze Chen[3], Zheng Zhu[1,2], Lingxi Xie[4,5], 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]Johns Hopkins University [5]Noah's Ark Lab, Huawei Inc.
{liyanwei2017,zhuzheng2014,xingang.wang}@ia.ac.cn
{xinze.chen,guan.huang,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 a 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 both in the MS-COCO (46.5% PQ) and Cityscapes (59.0% PQ) benchmarks.



(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 produced by the proposed approach.**

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* [23], 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 [23] and MS-COCO panoptic challenge winners [1] dealt with this task by directly combining FG instance segmentation models [15] and BG scene parsing [45] algorithms, which ignores the underlying relationship and fails to borrow rich contextual cues between *things* and *stuff*.

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 a 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 in-

tegrating 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 two popular segmentation benchmarks, namely, the MS-COCO [28] and Cityscapes [8] datasets, and claim **the state-of-the-art performance** in terms of PQ, a standard metric integrating accuracies of both *things* and *stuff* [23]. 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 simple and unified framework for both FG and BG segmentation, which reaches the top performance in MS-COCO [28] and Cityscapes [8] datasets. 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 or background targets [15, 45]. Recently, the rapid progress in object detection [13, 14, 34] and instance segmentation [9, 15, 25, 31] made it possible to achieve object localization and segmentation at a finer level. Meanwhile, the development of semantic segmentation [5, 6, 33, 45] boosted the performance of scene parsing. Despite their effectiveness, the separation of these tasks caused the lack of contextual cues 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* [23], which aims at accomplishing both tasks (FG instance and BG semantic segmentation) simultaneously.

Panoptic Segmentation: In [23], the author gave a benchmark of panopic segmentation by combining instance and semantic segmentation models. Later, a weakly-supervised method [24] was proposed on top of initialized semantic results, and an end-to-end approach [11] was designed to combine both FG and BG cues. However, their performance is far from the benchmark [23]. Different from them, 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. Most recently works include [22, 29, 40].

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 [9], FCIS [25], Mask R-CNN [15] and PANet [31]. Moreover, segmentation-based methods aggregated pixel-level cues to compose instances combined with semantic segmentation [2, 26, 32] or depth ordering [44] results.

Semantic Segmentation: With the development of so-called encoding-decoding networks such as FCN [33], rapid progress has been made in semantic segmentation [5, 6, 45]. In segmentation, capturing contextual information plays a vital role, for which various approaches were proposed including ASPP used in DeepLab [5, 6] for multi-scale contexts, DenseASPP [41] for global contexts, and PSPNet [45] which collected contextual priors. There were also efforts to use attention modules for spatial feature selection, such as [12, 42, 43], which will be detailed discussed next.

Attention-based Modules: Attention-based modules have been widely applied in visual tasks, including image processing, video understanding, and object tracking [7, 19, 37, 46, 47]. In particular, SENet [19] formulated channel-wise relationships via an attention-and-gating mechanism, non-

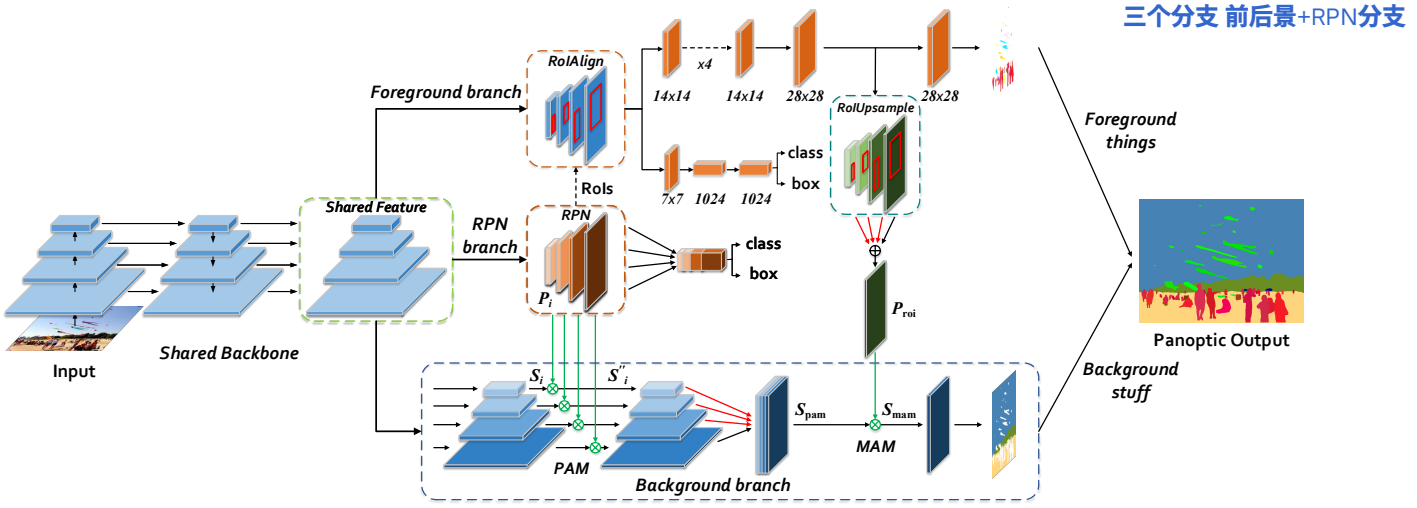


Figure 2. The proposed network structure. We adopt FPN as our backbone and share features with three parallel branches, namely *foreground branch*, *background branch*, and *RPN branch*. In the training stage, the network is optimized in an end-to-end manner. In the inference stage, panoptic results are generated by *things* and *stuff* results following the method described in Section 3.4. "$\oplus$" denotes element-wise sum and the green "$\otimes$" 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 are shown in Figure 3 and Figure 5. The red and green arrows represent upsample and attention operations, respectively.

local network [37] bridged self-attention for machine translation [36] to video classification using non-local filters. In the scope of scene understanding, [42] and [43] aggregated global contextual information as well as class-dependent features by channel-attention operations. More recently, self-attention and channel attention were adopted by [12] 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, 23] directly combined the instance and semantic results from separate models, such as Mask R-CNN [15] and PSPNet [45].

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 a 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) [23] 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 [27]) 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 a 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 $S_i \otimes P_i$, where "$\otimes$" 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 $S_{\text{pam}} \otimes P_{\text{roi}}$. 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) [34] 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

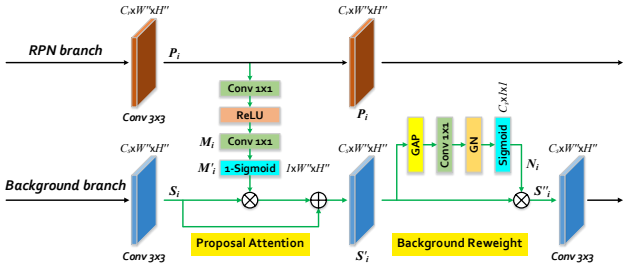


Figure 3. The designed proposal attention module (PAM) for complementary relationship establishment. We adopt this block in each scale of shared features, *i.e.*, W'' and H'' changes in each scale. Here, "$\otimes$" denotes spatial element-wise multiplication and "$\oplus$" denotes element-wise sum. The green arrows represent operations in PAM. GAP and GN indicate Global Average Pooling and Group Normalization [38], 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 weighted map 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 weighted map, 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 weighted map 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 $S_i \in \mathbb{R}^{C_s \times W'' \times H''}$.

Motivated by [19], a simple background reweight function is designed to downweight useless background layers after attention operation. We believe it could be improved, but it is beyond the scope of this work. The reweighted 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 [38] respectively, $N_i \in \mathbb{R}^{C_s \times 1 \times 1}$ means reweighting 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 reweight 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 [15], 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 [21], in which near points are given more contributions, an operation for *inverse* bilinear interpolation is formulated:

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

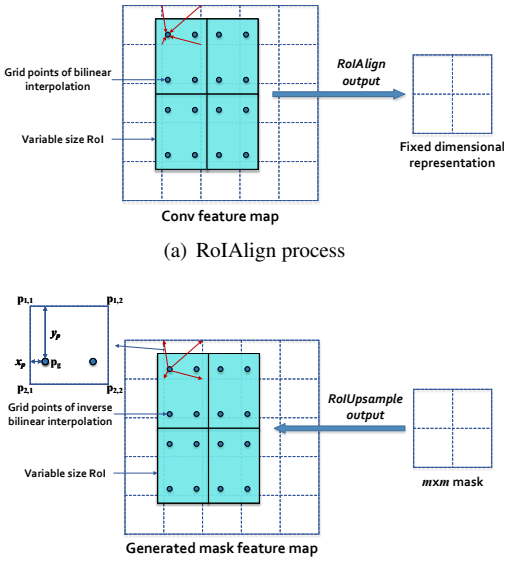
where $R(\mathrm{p}_{j,k})$ denotes the result of point $\mathrm{p}_{j,k}$ after inverse bilinear interpolation, $R(\mathrm{p}_g)$ here equals to one quarter of the corresponding value in the input mask, and normalized weights value_x, value_y are defined as:

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

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 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 [27]. 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 reweight function is adopted to aggregate useful highlighted background



(a) RoIAlign process

(b) RoIUpsample process

Figure 4. Comparison between RoIAlign [15] 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 show an example of RoIAlign output and RoIUpsample input with $m = 2$ for an intuitive illustration.

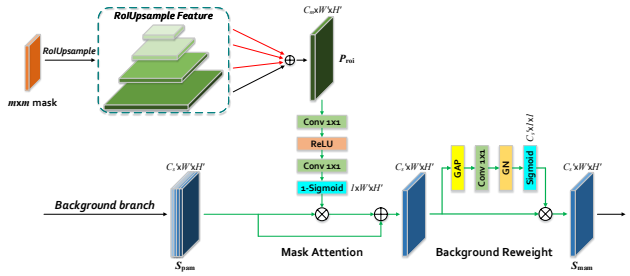


Figure 5. The proposed mask attention module (MAM) for a finer relationship modelling. Here, "$\otimes$" denotes spatial element-wise multiplication and "$\oplus$" denotes element-wise sum. The red and green arrows represent upsample and operations in MAM respectively. GAP and GN are identical with that 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 a unified framework. The whole network is optimized via a joint loss function $\mathcal{L}$ during training stage:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{RPN}} + \lambda_2 \mathcal{L}_{\mathrm{RCNN}} + \lambda_3 \mathcal{L}_{\mathrm{Mask}} + \lambda_4 \mathcal{L}_{\mathrm{Seg}} \quad (7)$$

where $\mathcal{L}_{\text{RPN}}$, $\mathcal{L}_{\text{RCNN}}$, $\mathcal{L}_{\text{Mask}}$, and $\mathcal{L}_{\text{Seg}}$ denotes the loss function of RPN, RCNN, instance segmentation, and semantic segmentation, respectively. Specifically, hyperparameters are designed to balance training processes, where λ_1 to λ_4 are set to $\{1, 1, 1, 0.3\}$ for MS-COCO and $\{1, 0.75, 1, 1\}$ for Cityscapes.

In details, we adopt ResNet-FPN [17, 27] as our backbone. And the hyperparameters in the foreground branch are set following Mask R-CNN [15]. The backbone is pretrained on ImageNet [35], and the remaining parameters are initialized following [16]. As standard practice [10, 17, 27], 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 several epochs (18 for MS-COCO and 100 for Cityscapes) using mini-batch stochastic gradient descent (SGD) with a weight decay of 4e-5 and a momentum of 0.9. Batch Normalization [20] in the backbone is fixed and Group Normalization [38] is added to all of the branches in our final results. For **MS-COCO** [28], the learning rate is initialized with 0.02 for the first 13 epochs and divided by 10 at 15-th and 18-th epoch respectively. Input images are horizontally flipped and reshaped to the scale with a 600 pixels short edge during training. Multi-scale testing is adopted for final results 4.3. For **Cityscapes** [8], the learning rate is initialized with 0.01 and divided by 10 at 68-th and 88-th epoch respectively. We construct each mini-batch for training from 16 random 512×1024 image crops (2 crops per GPU) after randomly flipping and scaling each image by 0.5 to 2.0×. Multi-scale testing is dropped in 4.3.

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 [23]. 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 [28] and Cityscapes [8] datasets. We first give description of the datasets 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** [28] is the most suitable and challenging one for the new panoptic segmentation task, for the detailed annotations 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. **Cityscapes** [8] dataset is adopted to further illustrate the effectiveness of the proposed method. In detail, it contains 2975 images for training, 500 images for validation and 1525 images for testing with *fine* annotations. It has another 20k *coarse* annotations for training, which are not used in our experiment. We report our results on *val* set with 19 semantic label and 8 annotated instance categories.

Evaluation Metrics: We adopt the evaluation metrics introduced by [23], 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:

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

where $\text{IoU}(p, g)$ means the intersection-over-union between predicted object p and ground truth g, true positives (TP) denotes matched pairs of segments ($\text{IoU}(p, g) > 0.5$), 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 [23] with no trick or multi-scale data augmentation in training and inference stage when doing component-wise analyses. As presented in Table 1, our proposed AUNet achieve an absolute improvement of $\mathbf{2}.\mathbf{4}\%$ 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 PQ^{St} and PQ^{Th}, which brings $\mathbf{1}.\mathbf{1}\%$ absolute improvements in PQ. This can be attributed to the

Table 1. Comparison among different settings of panoptic quality (%) on the MS-COCO dataset. "rewt" means using background reweight function in PAM and MAM. PQ^{Th} and PQ^{St} indicates PQ for *things* and *stuff* respectively.

Method	PAM	MAM	rewt	PQ	PQ^{Th}	PQ^{St}	AP	mIoU
sep	✗	✗	✗	37.2	47.1	22.8	33.4	44.5
e2e	✗	✗	✗	38.3	47.9	23.9	33.7	44.8
PAM	✓	✗	✗	39	48.5	24.5	34.2	45.1
PAM_r	✓	✗	✓	39.4	48.9	25.2	34.4	**45.3**
MAM	✗	✓	✗	38.9	48.6	24.2	34.3	45.2
MAM_r	✗	✓	✓	39.2	48.6	24.9	34.3	45.3
AUNet	✓	✓	✓	**39.6**	**49.1**	**25.2**	**34.7**	45.1

shared backbone and joint optimization, with which the network is supervised to focus on more discriminative features for both *things* and *stuff*. 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. 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 reweight 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. As shown in Figure 6, the activated feature map M_4' emphasize the background areas with context cues. It is worth noting that we have tried other fusion methods for FG and BG feature fusion, such as concatenation and direct summary after feature transformation. But these strategies have minor contributions, which means the attention is more appropriate for relationship establishment.

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 and Figure 6. 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 reweight function, MAM_r achieves 39.2% in PQ.

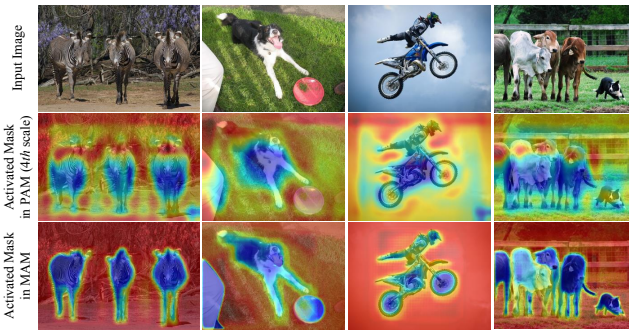


Figure 6. Heatmaps of the activated BG areas in PAM (the 4*th* scale, M_4') and MAM. The red regions are assigned more weights while the blue regions less weights in the *background branch*. All the input images are sampled from the MS-COCO *val* set.

4.3. Comparison to State-of-the-arts

We compare our proposed network with other state-of-the-art methods on MS-COCO [28] *test-dev* and Cityscapes [8] *val* set.

MS-COCO: 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 [4] is adopted for *things* and extra blocks or label bank [18] are added for *stuff* as well. Different from them, the proposed AUNet achieves the top performance in a unified framework with no extra data or additional network enhancements 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.

Filtering out the improvement bring by model ensemble, we compare the 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^{St}, for about $\mathbf{4.9}\%$ absolute improvements. Furthermore, the AUNet overpasses the former end-to-end method, namely JSIS-Net [11], 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 be-

[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.

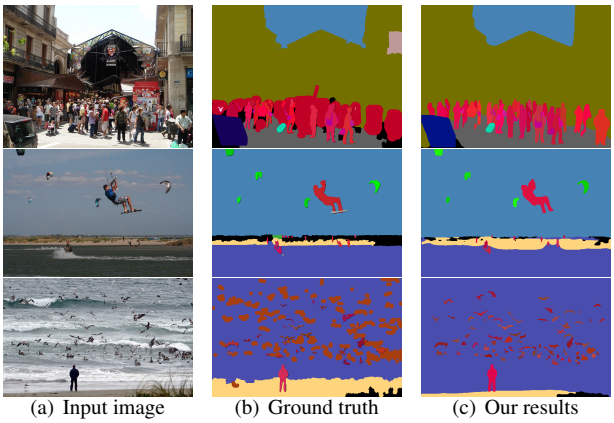
[3] We use the 64×4d variant of ResNeXt [39] with deformable conv [10] and non-local blocks [37].

Table 2. Panoptic quality (%) on MS-COCO 2018 *test-dev*. "extra data" here denotes using extra dataset for training, "e2e" represents using a 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 single model results with *no* extra data or network enhancement.

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 [11]	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**	**55.8**	**83.7**	**66.3**	**32.5**	**77.0**	**40.7**

Table 3. Panoptic quality (%) on the Cityscapes *val* set. PQ$^{\text{Th}}$ and PQ$^{\text{St}}$ denotes PQ result for *things* and *stuff* respectively. We compare our results with the bottom-up methods (the first row). Ours$_{\text{equ}}$ indicates all *things* are considered as one category in the background branch during training.

Method	backbone	PQ	PQ$^{\text{Th}}$	PQ$^{\text{St}}$	AP	mIoU
DWT [3]	VGG16	-	-	-	21.2	-
SGN [30]	VGG16	-	-	-	29.2	-
Li *et. al.* [24]	ResNet-101	53.8	42.5	62.1	28.6	-
Mask R-CNN [15]	ResNet-50	-	-	-	31.5	-
Ours$_{\text{equ}}$	ResNet-50-FPN	55.0	51.2	57.8	32.2	-
Ours	ResNet-50-FPN	56.4	52.7	59.0	33.6	73.6
Ours	ResNet-101-FPN	**59.0**	**54.8**	**62.1**	**34.4**	**75.6**



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

Figure 7. Example results of AUNet on the MS-COCO *val* set. Our performance on *things* 7(c) is even better than human annotations 7(b). The *things* of the same class share the same color family but appear in different intensities.

tween *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 7 gives intuitive presentations of the top performance using our proposed AUNet.

Cityscapes: We compare our proposed method with the leading bottom-up methods and Mask R-CNN in Table 3. Firstly, we adopt the same training strategy with that in MS-COCO, which means all *things* are considered as *one* category in background branch, denoted as **Ours**$_{\text{equ}}$. However, the strategy is inferior to that when using all 19 semantic labels, as illustrated in Table 3. Additionally, the MAM, which is proved to decrease the PQ in Cityscapes, is disabled in the final results. We guess the decline is caused by the inconsistency with prior information 1, which means the relatively worse *things* prediction may give wrong cues to *stuff*. Overall, the proposed method surpass previous state-of-the-art [24], with a 5.2% absolute gap.

5. Conclusions

This paper presents AUNet, a 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 attention (*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.

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

Acknowledgement

We would like to thank Jiagang Zhu and Yiming Hu for valuable discussions. This work was supported by the National Key Research and Development Program of China No. 2018YFD0400902 and National Natural Science Foundation of China under Grant 61573349.

References

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

[2] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 2017. 2

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017. 8

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 7

[5] 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. *PAMI*, 2018. 2

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

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 2

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 6, 7

[9] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 2

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 6, 7

[11] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Panoptic segmentation with a joint semantic and instance segmentation network. *arXiv:1809.02110*, 2018. 2, 7, 8

[12] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv:1809.02983*, 2018. 2, 3

[13] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 3, 5, 6, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[18] Hexiang Hu, Zhiwei Deng, Guang-Tong Zhou, Fei Sha, and Greg Mori. Labelbank: Revisiting global perspectives for semantic segmentation. *arXiv:1703.09891*, 2017. 7

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 4

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

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 5

[22] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *arXiv:1901.02446*, 2019. 2

[23] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *arXiv:1801.00868*, 2018. 1, 2, 3, 6

[24] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *ECCV*, 2018. 2, 8

[25] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. *arXiv:1611.07709*, 2016. 2

[26] Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *arXiv:1509.02636*, 2015. 2

[27] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3, 5, 6

[28] 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*, 2014. 2, 6, 7

[29] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. *arXiv:1903.05027*, 2019. 2

[30] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *ICCV*, 2017. 8

[31] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 2

[32] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *ECCV*, 2018. 2

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 4

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3, 7

[38] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 4, 6

[39] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 7

[40] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. *arXiv:1901.03784*, 2019. 2

[41] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *arXiv:1804.09337*, 2018. 2, 3

[43] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 2, 3

[44] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *CVPR*, 2016. 2

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2, 3

[46] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 2

[47] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *CVPR*, 2018. 2