Attention-guided Unified Network for Panoptic Segmentation

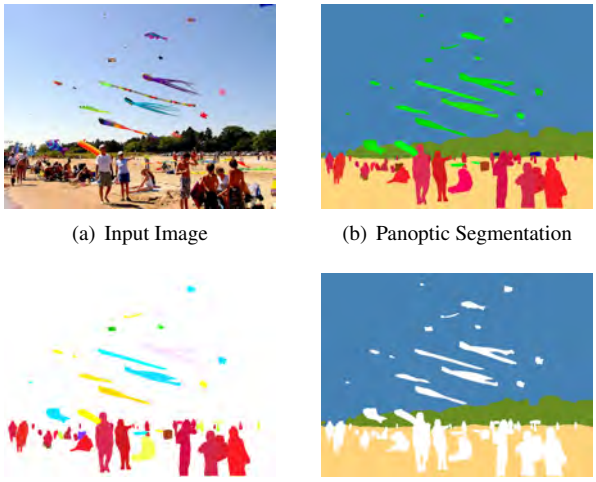
Yanwei Li[1,2], Xinze Chen[3], Zheng Zhu[1,2], Lingxi Xie[4,5], Guan Huang[3],
Dalong Du[3], Xingang Wang[1]
[1]Institute of Automation, CAS [2]University of Chinese Academy of Sciences
[3]Horizon Robotics, Inc. [4]Johns Hopkins University [5]Noah's Ark Lab, Huawei Inc.
{liyanwei2017,zhuzheng2014,xingang.wang}@ia.ac.cn
{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.

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*.



(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.**

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-

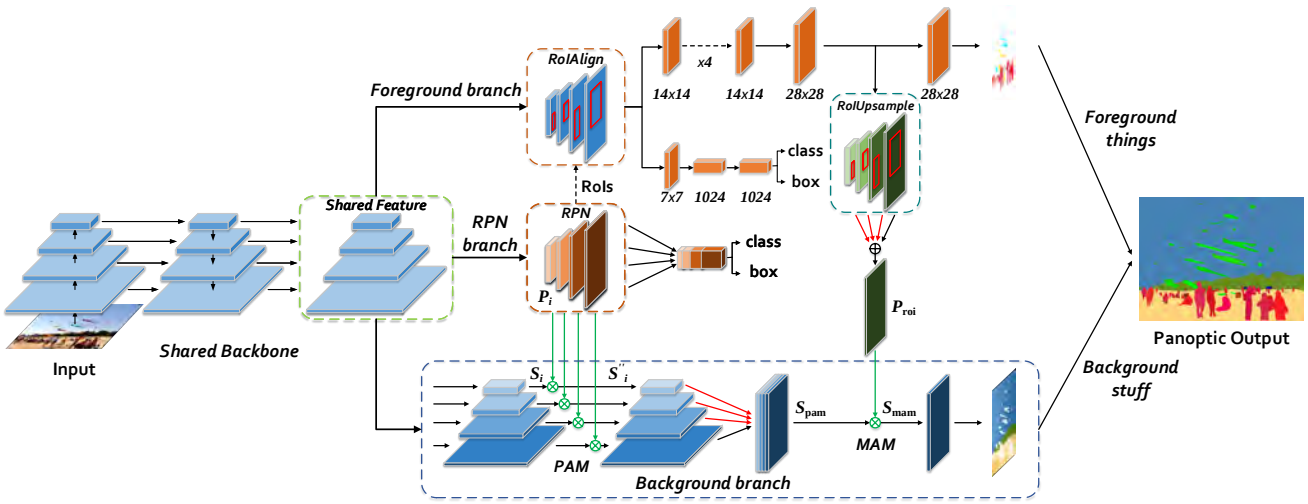


Figure 2. The proposed network structure. We adopt FPN as our backbone and share features with three parallel branches, namely *foreground branch*, *background branch*, and *RPN branch*. In the training stage, the network is optimized in an end-to-end manner. In the inference stage, panoptic results 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

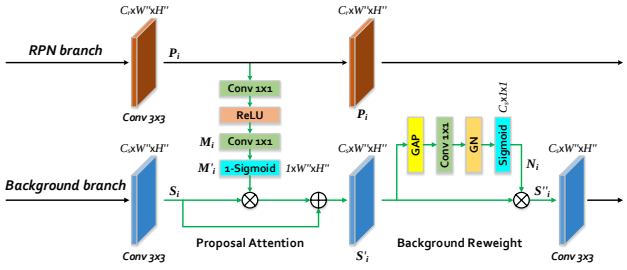


Figure 3. The designed proposal attention module (PAM) for complementary relationship establishment. We adopt this block in each scale of shared features, *i.e.*, W'' and H'' changes in each scale. Here, "$\otimes$" denotes spatial element-wise multiplication and "$\oplus$" denotes element-wise sum. 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:

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

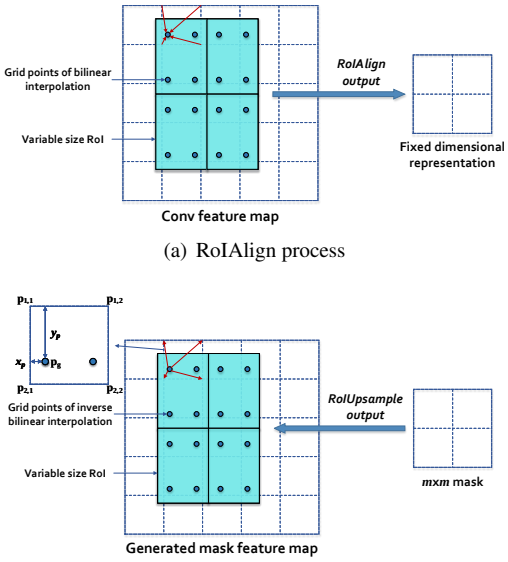
where $R(\mathrm{p}_{j,k})$ denotes the result of point $\mathrm{p}_{j,k}$ after inverse bilinear interpolation, $R(\mathrm{p}_g)$ here equals to one quarter of the corresponding value in the input mask, and normalized weights value_x, value_y are defined as:

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

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.

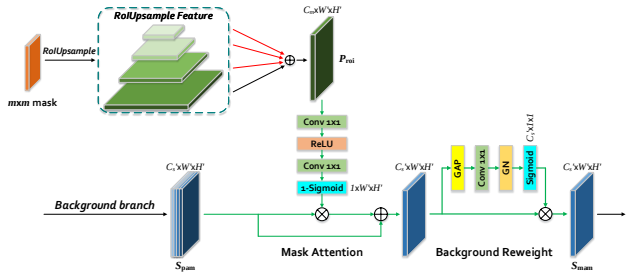


Figure 5. The proposed mask attention module (MAM) for a finer relationship modelling. Here, "$\otimes$" denotes spatial element-wise multiplication and "$\oplus$" denotes element-wise sum. The red and green arrows represent upsample and operations in MAM respectively. GAP and GN are identical with that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

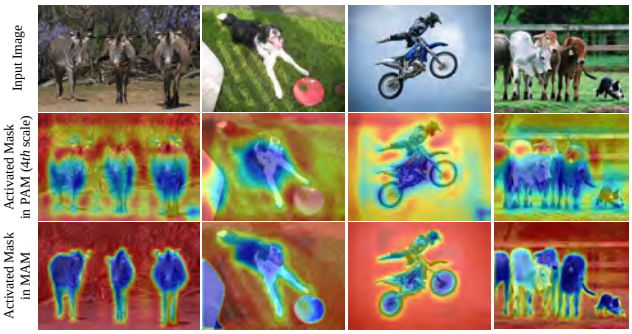


Figure 6. Heatmaps of the activated BG areas in PAM (the 4*th* scale, M_4') and MAM. The red regions are assigned more weights while the blue regions less weights in the *background branch*. All the input images are sampled from the MS-COCO *val* set.

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.

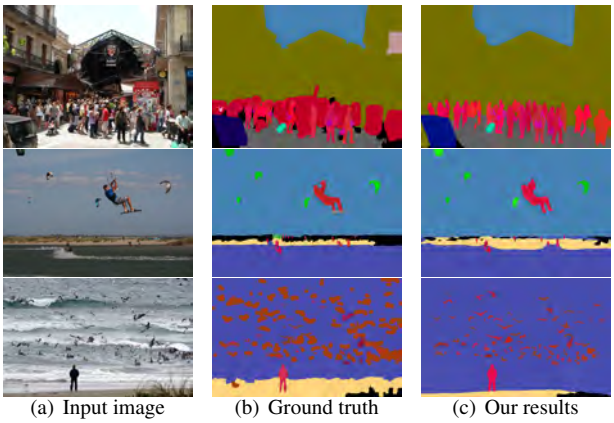
[3] We use the 64×4d variant of ResNeXt [39] with deformable conv [10] and non-local blocks [37].

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

CVPR2019 Paper Translation

姓名：廖恒逸

学号：2017300545

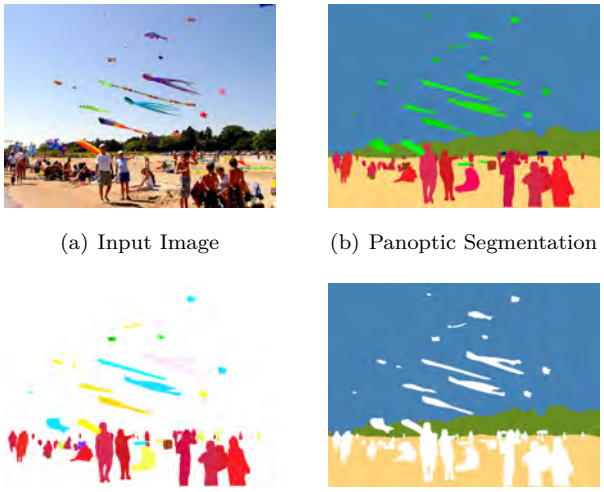
班号：10011705

针对全景分割的注意引导统一网络

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

摘要

这篇论文研究了一个最近被提出的问题——全景分割。全景分割不但需要对前景的物体进行实例级分割，而且需要对背景的内容进行语义级分割。现有的方法大多分别处理这两个问题，但是在这篇论文中，我们揭露了他们之间的根本联系，尤其是前景的物体提供了充足的线索来帮助我们识别背景。我们的方法被命名为注意引导统一网络（AUNet), 它是一个同时对前景和背景同时进行分割的统一结构。两个注意力资源被添加到了背景分支，即 RPN 和前景的分割掩码，分别提供对象级和像素级的注意力，我们的方法被推广到不同的主干网，在前景和背景的分割中取得了一致的准确度增益，并且同时也在 MS-COCO 和 Cityscapes 数据集上取得了最好的结果。



(a) Input Image (b) Panoptic Segmentation

(c) Foreground: *things* (d) Background: *stuff*

图 1. 给出图像 1(a)，全景分割 1(b)的目标是在实例级 1(c)识别前景事物并且在语义级 1(d)识别背景材料。同类事物共享同一种颜色系列，但显示为不同的色彩强度。**所有的结果都由我们提出的方法获得。**

1. 引言

场景理解是计算机视觉中一项基本但具有挑战性的任务，它对其他一些应用程序有很大的影响，比如自动驾驶和机器人。场景理解的经典任务主要包括对象检测，实例分割和语义分割。本文考虑了最近提出的一个名为全景分割 [21] 的任务，其目的是在实例层面识别所有前景（FG）对象（事物，主要包括可数的目标，比如人，动物，工具等），同时解析背景（BG）在语义层面上的内容（材料，主要包括具有相似纹理和/或材料的无定形区域，例如草地，天空，道路等）。基准算法 [21] 和 MS-COCO 全景挑战获胜者 [16] 通过直接结合前景的实例分割模型 [13] 和背景的场景解析算法 [42] 来处理这一任务，这些算法忽略了底层关系并且未能借用事物和材料间丰富的上下文线索。

在本文中，我们提出了一个概念上简单且统一的全景分割框架。为了促进前景事物和背景材料之间的信息流动，我们将传统的实例分割和语义分割网络结合起来，从而形成具有两个分支的统一网络。这种策略可以立即提高分割准确度，提高计算效率（因为网络骨干可以产生共享）。这意味着全景分割能被前景事物和背景内容提供的充足信息所优化，这也为我们的方法

奠定了基础。

更进一步，我们探索了将更高级的视觉线索（即，从主干末端提取的特征之外）整合来完成更准确的分割的可能性。这是通过分别在对象级和像素级工作的两个基于注意力的模块来实现的。对于在对象级工作的模块，我们参考区域建议，每个建议都表示一个可能的前景事物，并调整相应区域被视为前景事物和背景材料的概率。对于在像素级工作的模块，我们取出前景的分割掩膜，并用它来优化前景事物和背景材料之间的界限。在深度网络环境中，这两个模块分别称为建议关注模块（PAM）和掩膜关注模块（MAM），实现为跨越前景和背景分支的额外连接。在 MAM 模块中，一个名为 *RoIUpsample* 的新层被用于定义固定形状的掩模和相应特征图中的像素之间的精确映射函数。在实践中，所有额外连接都是从前景分支到背景分支，主要是由于观察到前景分割通常更准确[1]。此外，背景材料虽然由前景事物重新定义，但也通过梯度提供反馈。因此，前景和背景的分割准确性都得到了显著提高。

名为注意力引导统一网络（**AUNet**）的整体方法可以轻松地实现为各种网络框架，以端到端的方式进行优化。我们在两个流行的目标检测数据集中评估 AUNet，即 MS-COCO [27] 和 Cityscapes [6] 数据集，并在 PQ 方面具有**最佳性能**，PQ 是一种整合了事物和区域精确度的标准度量 [21]。此外，联合优化和两个基于注意力的模块带来的好处通过大量的模型简化测试进行了验证 4.2。

这项研究的主要贡献是为前景和背景的分割提供了一个简单且统一的框架，在 MSCOCO [27] 和 Cityscapes [6] 数据集中达到了最佳性能。此外，本工作还研究了前景对象和背景内容提供的互补信息。虽然全景分割可以作为研究这一主题的原始场景，但其应用在于更广泛的视觉任务。我们的解决方案 AUNet 是这个领域的初步探索，但我们期待在这个方向上做出更多努力。

本文的其余部分安排如下。第 2节简要的回顾了相关工作。第 3节详细描述了 AUNet，包括两个基于注意力的模块。第 4节中展示了后续的实验，第 5节中是对这项工作的总结

2. 相关工作

基于传统深度学习的场景理解研究往往侧重于前景或背景目标 [13,42]。最近，物体检测 [11,12,32] 和实例分割 [7,13,24,29] 的快速发展使得有可能在更高的水平上实现物体定位和分割。同时，语义分割 [3,4,31,42] 的发展提高了场景解析的性能。尽管它们具有有效性，但这些手段的分离导致了在实例分割中缺乏上下文联系，同时语言分割中的个体也给我们带来了困惑。为了弥合这一差距，最近，研究人员提出了一项名为全景分割 [21] 的新任务，旨在同时完成这两项任务（前景实例和背景语义分割）。

全景分割: 在 [21] 中，作者通过结合实例和语义分割模型给出了全景分割的基准。后来，在初始化的语义结果之上提出了一种弱监督方法 [23]，并且设计了一种端到端的方法 [9] 来组合前景和背景的线索。然而，他们的表现远非基准 [21]。与它们不同，我们提出的 AUNet 在端到端框架中实现了最佳性能。此外，我们还建立了基于建议的实例和基于 FCN 的语义分段之间的联系。最近的作品包括 [22,28,37]。

实例分割: 实例分割旨在区分相同对象的不同实例。主要有两种方法来解决这个问题，即基于建议的方法和基于分割的方法。基于建议的方法，在准确的区域建议的帮助下，往往能实现更高的性能。最近的例子包括 MNC [7]，FCISFCIS [24]，Mask RCNN [13] 和 PANet [29]。此外，基于分割的方法整合像素级线索来识别由语义分割 [1,25,30] 或深度排序 [41] 结果组合而成的实例。

语义分割: 随着所谓的编码-解码网络比如 FCN [31] 的发展，语义分割 [3,4,42] 已经取得了快速的进步。在语义分割中，捕获上下文信息起着至关重要的作用，为此提出了各种方法，包括在 DeepLab [3,4] 中用于多尺度上下文的 ASPP，用于全局上下文的 DenseASPP [38]，以及收集上下文先验的 PSPNet [42]。还有一些研究将注意力模块用于空间特征选择，例如 [10,39,40]，在后文将详细讨论。

基于注意的模块: 基于注意的模块已广泛应用于各种视觉任务，包括图像处理，视频理解和目标跟踪 [5,18,35,43,44]。特别是 SENet [18] 通过注意力，门控机制

[1]当考虑到具有相同语义的同类事物时，比如测试中的所有个体都被评估为人时，我们发现事物的像素准确度远远高于材料 (6.7% 的绝对差距)。我们根据相同的 MS-COCO 语义评估指标对它们进行评估。

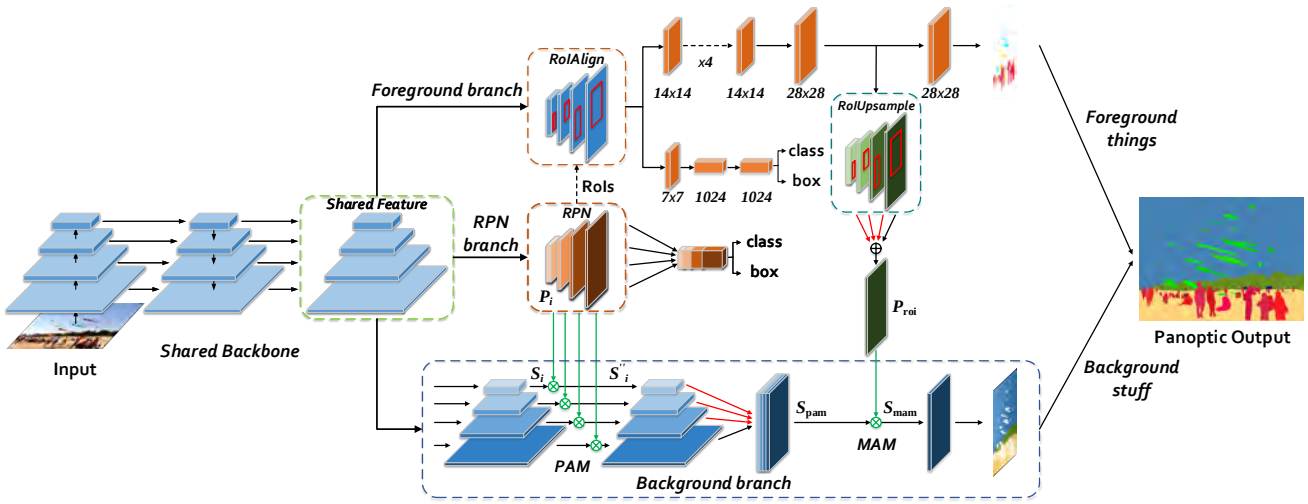


图 2. 提出的网络的结构。我们采用 FPN 作为我们的框架并在三个并行分支中共享特征，即前景分支，背景分支，和 *RPN* 分支。在训练阶段，网络通过端到端的方式进行优化。在推断阶段，事物和材料结果遵循3.4节描述的方法，生成全景结果。“⊕”表示元素求和，绿色的“⊗”根据其所在位置表示建议关注模块 (PAM) 或掩膜关注模块 (MAM)。PAM 和 MAM 模拟了两个分支间的互补关系。PAM 和 MAM 的细节在图 3和图5中展示。红色和绿色的箭头分别表示采样和关注操作。

和非局部网络 [35] 将用于机器翻译 [34] 的注意力桥接到使用非本地过滤器的视频分类，形成了通道关系。在场景理解中，[39] 和 [40] 通过频道注意操作整合全局上下文信息以及类依赖特征。最近，[10] 采用了自我关注和频道注意来分别对空间和信道维度中的长距离情境进行建模。在这项工作中，我们通过一系列粗到细的注意块建立了全景分割中的前景事物和背景材料之间的关系。

3. 注意引导统一网络

3.1. 问题和基线

全景分割的任务旨在理解一个视图中所有可见的东西，这意味着必须为图像的每个像素分配一个语义标签和一个实例 ID。为了解决这个问题，现有的顶级算法 [16, 21] 直接从分离的模型组合了实例和语义结果，如 Mask R-CNN [13] 和 PSPNet [42]。

我们将全景分割的问题制定为识别和分割所有前景事物并理解所有背景材料。在这种情况下，我们解决了来自两个方面的问题，即统一网络中的前景分支和背景分支（图2）。详细地说，给定输入图像 $\boldsymbol{X}$，我们的目标是同时生成前景事物结果 $\boldsymbol{Y}_{\mathrm{Th}}$ 和背景材料结果 $\boldsymbol{Y}_{\mathrm{St}}$。因此，可以使用 3.4节中的融合方法直接从 $\boldsymbol{Y}_{\mathrm{Th}}$ 和 $\boldsymbol{Y}_{\mathrm{St}}$ 生成全景结果 $\boldsymbol{Y}_{\mathrm{Pa}}$。如第4.1节所述，全景结果的性能通过全景质量 (PQ) [21] 评估。为此，我们首先在本节介绍了用于全景分割的统一框架。然后，我们对关注度模块中的关键元素进行了详细阐述，包括候选关注模块 (PAM) 和掩模关注模块 (MAM)。最后，我们给出了实现细节。

在这项工作中，我们将从单独的模型生成事物和材料的方法作为我们的基线。具体而言，基线方法分别从单独的模型 $\mathbb{M}_{\mathrm{Th}}$ 和 $\mathbb{M}_{\mathrm{St}}$ 给出事物$\boldsymbol{Y}_{\mathrm{Th}}$ 和材料$\boldsymbol{Y}_{\mathrm{St}}$ 的结果。并且前景模型 $\mathbb{M}_{\mathrm{Th}}$ 和背景模型 $\mathbb{M}_{\mathrm{St}}$ 给出了类似的骨干网（例如，FPN [26]）以及后面的统一框架。

3.2. 统一框架

为了弥合前景事物与背景材料之间的差距，我们提出了注意引导统一网络 (AUNet)。与基线方法相比，所提出的 AUNet 通过共享相同的骨干将两个模型（$\mathbb{M}_{\mathrm{Th}}$ 和 $\mathbb{M}_{\mathrm{St}}$）融合在一起，并从并行分支生成 $\boldsymbol{Y}_{\mathrm{Th}}$ 和 $\boldsymbol{Y}_{\mathrm{St}}$。如图 2清楚所示，AUNet 在概念上很简单：采用 FPN 作为主干，从不同范围里提取判别特征并由所有分支共享。

与传统方法不同，传统方法直接整合了 $\mathbb{M}_{\mathrm{Th}}$ 和 $\mathbb{M}_{\mathrm{St}}$ 的结果，提出的 AUNet 使用一种联合损耗函数 $\mathcal{L}$（在第 3.4节中定义）对其进行了优化，并在统一框架中促进这两项任务。详细地说，我们采用基于建议的实例分割模块来在前景分支中生成更精确的掩模 M。对于

背景分支，*Light Heads* 用于整合来自共享的多尺度特征的场景信息。通过这种方式，共享骨干网由前景事物和背景材料同时监督，这可以促进这两个分支在特征空间中的连接。为了更明确地建立前景对象与背景内容之间的联系，我们增加了两个注意模块资源。我们考虑第 i 个比例背景特征图与相应的 RPN 特征图之间的粗略注意操作，分别用 S_i 和 P_i 表示。如图 2所示，“$\otimes$”表示注意操作，那么注意模型就可以表示为 $S_i \otimes P_i$。此外，更优化的关系能通过已处理的特征图 S_{pam} 和已生成的前景分割掩膜 P_{roi} 确定，可以表示为 $S_{\text{pam}} \otimes P_{\text{roi}}$。详情将在下面的部分进行探讨。

3.3. 注意引导模块

考虑到前景事物和背景材料之间的互补关系，我们引入了从前景分支到背景分支的特征，以获得更多的上下文线索。从另一个角度来看，连接两个分支的注意操作也建立了基于建议方法和基于 FCN 方法分割之间的联系。为此，提出了两个空间关注模块，即建议关注模块 (PAM) 和掩模关注模块 (MAM)。

3.3.1 建议关注模块

在经典的二阶检测框架中，引入了候选区域网络 (RPN) [32] 来给出预测的二进制类标签（前景和背景）和边界框坐标。这意味着 RPN 包含丰富的背景信息，这些信息只能从背景分支中的注释获得。因此，我们提出了一种建立前景事物和背景内容之间互补关系的新方法，称为候选关注模块（PAM）。如图3所示，我们利用来自 RPN 分支的上下文线索进行注意操作。在这里，

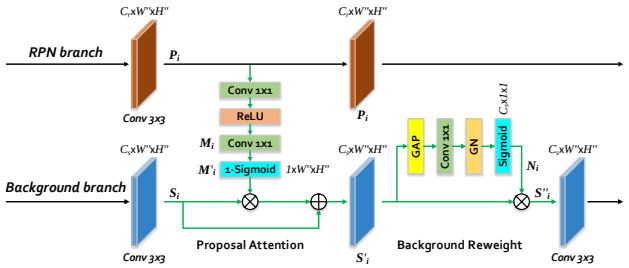


图 3. 为建立互补关系而设计的建议关注模块 (PAM)。我们在每个共享特征的尺度内采用此块, 比如 W'' 和 H'' 在每个尺度下都会改变。在这里，“$\otimes$”表示空间元素乘法，“$\oplus$”表示元素加法。绿色箭头代表 PAM 中的操作。GPA 和 GN 分别表示全局池化层和分组归一化 [36]。

我们给出了一个详细的公式说明这个过程。给定来自第 i 个区域的 RPN 分支的输入特征图 $P_i \in \mathbb{R}^{C_r \times W'' \times H''}$，在 sigmoid 之前的前景加权映射 M_i 可以表示为：

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

其中 $f(\cdot,\cdot)$ 表示卷积函数，σ 表示 ReLU 激活函数，$M_i \in \mathbb{R}^{1 \times W'' \times H''}$ 表示生成的 FG 加权映射，$w_{i,1} \in \mathbb{R}^{C'_r \times C_r \times 1 \times 1}$ 和 $w_{i,2} \in \mathbb{R}^{1 \times C'_r \times 1 \times 1}$ 都表示卷积参数。

为了关注背景内容，我们将注意加权映射 M'_i 表示为 $1 - \text{sigmoid}(M_i)$。然后，第 i 个尺度激活特征映射 $S'_i \in \mathbb{R}^{C_s \times W'' \times H''}$ 可以表示为：

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

其中 $\otimes$ 和 $\oplus$ 分别表示元素点乘法和元素求和，$S_{i,j}$ 表示第 j 层语义特征映射 $S_i \in \mathbb{R}^{C_s \times W'' \times H''}$。

受 [18] 的启发，一个简单的再加权函数被设计为在注意操作后减少无用的背景层。我们相信它可以做得更好，但改进它已在本文工作以外。再加权特征映射 $S''_i \in \mathbb{R}^{C_s \times W'' \times H''}$ 可以表示为：

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

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

其中 G 和 GN 分别表示全局池化层和分组归一化 [36]，$N_i \in \mathbb{R}^{C_s \times 1 \times 1}$ 表示再加权算子，$w_{i,3} \in \mathbb{R}^{C_s \times C_s \times 1 \times 1}$ 表示卷积参数，$S'_{i,k}$ 表示 S'_i 中的第 k 个像素通道。

基于上述 PAM 公式，我们通过注意操作和背景再加权函数突出显示了共享特征映射中的背景区域。它还通过在反向传播算法期间提高已激活前景区域的权重来促进识别独立事物的学习（参见第4.2节）。

3.3.2 掩膜关注模块

随着 PAM 语境线索的引入，背景分支能更加专注于材料区域。然而，来自 RPN 分支的预测区域缺乏足够的用于精确表示背景线索。与 RPN 特征不同，$m \times m$ 固定形状掩码由前景分支的编码布局生成。因此，我们提出掩模注意模块（MAM）来进一步建立关系，如图5所示。因此，与之前类似的注意操作需要 $1 \times W' \times H'$ 形状的前景分割掩模。现在的问题是：如何从 $m \times m$ 掩模重现 $W' \times H'$ 形状前景特征图？

RoIUpsample: 为了解决尺寸不匹配问题，我们提出了一种名为 *RoIUpsample* 的可区分层。具体而言，RoIUpsample 与 RIAIAlign 的逆过程 [13] 相似，如图4所示。在 RoIUpsample 层中，$m \times m$ 掩模（m 等于掩模 R-CNN 中的 14 或 28）首先被重新整形为相同大小的 RoIs（由 RPN 生成）。然后我们利用设计的反双线性插值来计算每个掩模箱中四个定期采样位置（与 RoIAlign 相同）的输出特征值，然后将最终结果整合得到生成的掩模特征图。为了满足双线性插值 [20] 的要求，近似点给与了更多的贡献，反双线性插值的公式如下：

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

其中 $R(\mathrm{p}_{j,k})$ 表示经过反双线性插值后的点 $\mathrm{p}_{j,k}$ 的结果，此处的 $R(\mathrm{p}_g)$ 等于输入掩码中相应值的四分之一，归一化权重 value_x，value_y 定义为：

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

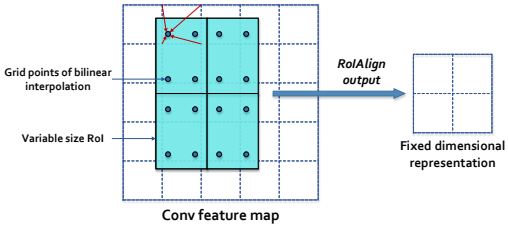
其中 x_p 和 y_p 分别表示点 p_g 和生成的 $\mathrm{p}_{1,1}$ 之间的间距，如图4(b)所示。注意，对于方程5和6，$m \times m$ 掩模也可以通过前向双线性插值从生成的 $W' \times H'$ 特征映射中恢复。

然后，根据 RoI 的大小将生成的特征映射分配给四个不同的尺度，这与 FPN [26] 类似。因此，生成的 FG 特征图用于以下操作。

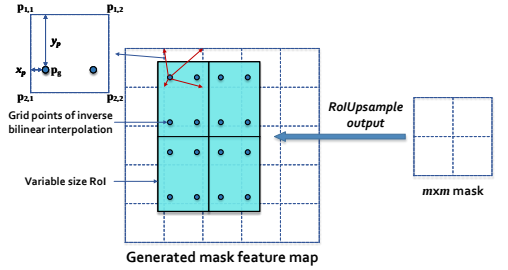
注意操作: 与传统的实例分割任务不同，预测的 FG 掩模用于在像素级上为背景分支提供更多的上下文线索。我们首先使用 RoIUpsample 将他们一起放到 $C_m \times W' \times H'$ 特征图中，如图5所示。然后，类似于 PAM，可以生成 $1 \times W' \times H'$ 激活的背景区域。通过引入注意，前景掩膜由语义损失函数监督，它可以进一步改善场景理解（包括事物和材料），如第4.2节所述。我们采用了一个类似的背景再加权函数来聚合有用的被标记背景特征。因此，我们用提出的 PAM 和 MAM 模拟了前景事物和背景材料之间的互补关系。

3.4. 实现细节

在本节中，我们将提供 AUNet 的训练和推理阶段的更多实现细节。



(a) RoIAlign process



(b) RoIUpsample process

图 4. RoIAlign [13] 和我们提出的 RoIUpsample 的比较。我们设计的 RoIUpsample, 可以被看作一种 RoIAlign 的逆操作, 能根据前景掩膜的准确空间位置恢复特征图。在这里，我们展示了一组 $m = 2$ 的 RoIAlign 输出，RoIUpsample 输入的例图。

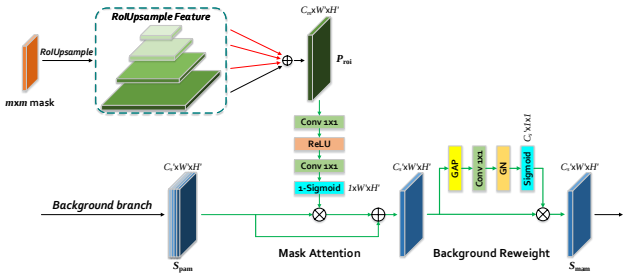


图 5. 为实现更精细的关系建模提出的掩膜关注模块。这里，“⊗”表示空间元素乘法，“⊕”表示元素加法。红色和绿色的箭头分别代表 MAM 中的采样和操作。GAP 和 GN 的含义和 PAM 中相同。

训练: 如3.2节所述，我们提出的所有方法都在一个统一的框架中进行训练。整个网络通过联合损失函数 $\mathcal{L}$ 在训练阶段进行优化：

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

其中 $\mathcal{L}_{\mathrm{RPN}}$，$\mathcal{L}_{\mathrm{RCNN}}$，$\mathcal{L}_{\mathrm{Mask}}$ 和 $\mathcal{L}_{\mathrm{Seg}}$ 分别表示 RPN，RCNN，实例分割和语义分割的损失函数。具体而言，超参数被设计用于平衡训练过程，其中 λ_1 至 λ_4 针对 MS-COCO 数据集设置为 {1, 1, 1, 0.3}，针对 Cityscapes

数据集设置为 {1, 0.75, 1, 1}。

详细而言，我们采用 ResNet-FPN [15, 26] 作为我们的框架。前景分支中的超参数是学习 Mask R-CNN [13] 设置的。框架在 ImageNet [33] 上预先训练，其后的参数学习 [14] 进行初始化。8 个 GPU 作为标准练习 [8, 15, 26] 用于训练所有模型。在每个小批量中，每 2 个基于 ResNet-50 和 ResNet-101 网络的图像都有一个 GPU，其他网络每个 GPU 有 1 个图像。我们的网络使用小批量随机梯度下降法 (SGD)，其权重衰减为 4e-5，动量为 0.9，网络在多次训练中进行了优化（MS-COCO 为 18，Cityscapes 为 100）。骨干中的批量归一化 [19] 是固定的，组归一化 [36] 被添加到我们最终结果中的所有分支中。对于 **MS-COCO** 数据集 [27]，学习率在前 13 个训练周期初始化为 0.02，并且分别在第 15 次和第 18 次训练周期各降低 10 倍。在训练期间，输入的图像被水平翻转和重塑为边长为 600 个像素点的方形区域。最终结果采用多区域测试4.3。对于 **Cityscapes** 数据集 [6]，学习率初始化为 0.01，并且分别在第 68 次和第 88 次训练周期各降低 10 倍。我们构建小批量从 16 个随机 512×1024 图像裁剪（每个 GPU 裁剪 2 份）进行训练，随机翻转并将每个图像缩放 0.5 到 2.0 倍。多区域测试在4.3中被放弃。

推理: 通过以与 [21] 中类似的方式整合前景事物和背景材料的结果，在推理阶段产生全景结果。在这个阶段，事物的重叠问题首先在类似于 NMS 的过程中被解决，这些分割预测具有较高可信度。在此过程中还会考虑类别之间的关系。例如，在最终结果中，绳不应该与人物重叠。然后，通过首先分配实例标签以区分事物，将非重叠实例分割和材料结果组合。

4. 实验

在本节中，我们的方法在 Microsoft COCO [27] 和 Cityscapes 数据集 [6] 上进行评估。我们首先介绍数据集以及评估指标。然后我们评估我们的方法并给出详细的分析。最后给出了与全景分割中最先进方法的比较。

4.1. 数据集和指标

数据集: 由于全景任务本身的新颖性，很少有数据集具有详细的全景注释以及公共评估指标。对于详细的注释和数据高复杂性，**Microsoft COCO** [27] 数据集是最适合和最具挑战性的。它包括用于训练的 115k 张图像和用于验证的 5k 张图像，以及用于 *test-dev* 的 20k 张图像和用于 *test-challenge* 的 20k 张图像。MS-COCO 全景注释包括 80 个事物类别和 53 个材料类别。我们在 *train* 数据集上训练我们的模型，没有额外的数据，并比较 *val* 集和 *test-dev* 集的结果。采用 **Cityscapes** [6] 数据集是为了进一步说明该方法的有效性。详细地说，它包含 2975 张用于训练的图像，500 张用于验证的图像和用于使用精细注释进行测试的 1525 张图像。它还有 20k 张带粗略注释的图像用于训练，但在我们的实验中没有使用它们。我们使用 19 个语义标签和 8 个带注释的实例类别报告我们在 *val* 集上的结果。

评估指标: 我们采用 [21] 引入的评估指标，该指标计算了全景质量 (PQ)。PQ 可以解释为分割质量 (SQ) 项和识别质量 (RQ) 项的乘法：

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

其中 $\mathrm{IoU}(p,g)$ 表示模型产生的目标窗口和原来标记窗口的交叠率，真正 (TP) 表示匹配正确的分割 $(\mathrm{IoU}(p,g) > 0.5)$，假正 (FP) 表示不匹配的预测分割，假负 (FP) 表示被模型预测为负的正样本。我们的结果中也报告了事物和材料的 PQ，SQ 和 RQ。

4.2. 分量分析和诊断

在本节中，我们将逐步分解我们的方法，以揭示每个组件的作用。本节中的所有实验都在 MSCOCO 数据集上进行训练和评估，并且没有额外的数据。在这里，我们采用 ResNet-50-FPN 作为我们的框架。为了公平比较，我们严格遵循 [21] 中的合并方法，在进行分量分析时，在训练和推理阶段没有欺骗或多尺度数据增加。如表 1所示，与单独的训练方法相比，我们提出的 AUNet 实现了 **2.4**% 的 PQ 改善。

4.2.1 统一框架

正如 3.2节所阐述的那样，我们提出的统一框架在并行的分支处理前景事物和背景材料。如表 1所示，统一框架提高了 $\mathrm{PQ}^{\mathrm{St}}$ 和 $\mathrm{PQ}^{\mathrm{Th}}$ 的性能，为 PQ 带来 **1.1**% 的改进。这可以归因于共享框架和联合优化，网络被监督使其能专注于事物和材料的辨别特征。通过

表 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

共享框架，可以有效地减少对材料的错误分类，并为事物提供更多细节。

4.2.2 建议注意模块

提出的 PAM 建立了不同尺度的事物和材料之间的互补关系。通过这种方式，二进制分类 RPN 分支在语义标签监督下得到优化。通过建立事物和材料之间的联系，网络在 PQ^{St} 和 PQ^{Th} 中获得了持续的优化，如表1所示。背景再加权函数证明了它在 PQ^{St} 中的有效性。这可能是由公式 3中全局池化层引入的全局上下文特征造成的，这意味着它选择在全局上下文的指导下高亮背景特征。如图 6所示，激活的特征图 M_4' 强调了具有上下文线索的背景区域。值得注意的是，我们已经尝试了其他聚合方法用于前景和背景的特征聚合，例如特征转换后的级联和直接汇总。但这些策略的贡献很小，这意味着注意力更适合建立关系。

4.2.3 掩膜关注模块

虽然 PAM 建立了前景对象与背景内容之间的联系，但 MAM 给出了背景细节表示，如第 3.3.2 节和图 6所述。在 PAM 中，MAM 在 PQ^{St} 和 PQ^{Th} 中也比原始方法具有更好的性能。但是，MAM 的效果略逊于 PAM。我们猜测这是由于生成的前景分割掩模中缺少上下文线索造成的。[2]实际上，我们还评估了 RoIUpsample 在采集不同分辨率掩模时的性能，即 14×14 掩模和 28×28 掩模。结果表明高分辨率掩模特征比低分辨掩模特征带来了更好的增益（PQ 提高 0.1%）。这是合理的，因为如果给定更高分辨率的掩模，

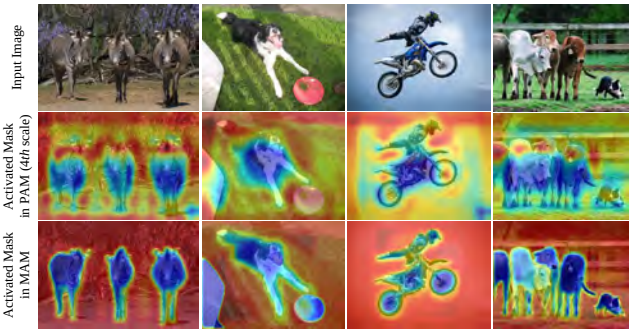


图 6. PAM 中背景区域的热像图 (第四个尺度, M_4') 和 MAM。在背景分支中，红色区域被赋予较多权重，而蓝色区域被赋予较少权重。所有输入的图像都来自 MS-COCO 的 *val* 集。

RoIUpsample 层会生成更精细的图层。在背景再加权函数的帮助下，MAM_r 在 PQ 中达到 39.2%。

4.3. 与现有技术比较

我们将我们提出的网络在 MS-COCO [27] 的 *test-dev* 数据集和 Cityscapes [6] 的 *val* 数据集上与其他最先进的方法进行比较。

MS-COCO: 如表 2所示，提出的未附加额外功能的 AUNet 在 MS-COCO 数据集中实现了**46.5**% 的领先性能。详细地说，COCO2018 全景挑战的获胜者 [16] 在训练和推理阶段采用了许多额外的网络增强，例如，丰富的额外数据（110k 张外部注释的 MS-COCO 图像），多尺度训练，模型集成。此外，考虑到获胜者团队采用的网络增强功能，级联 R-CNN [**?**] 被用于事物识别，额外的块或标签库 [17] 也被添加。与它们不同的是，提出的 AUNet 在统一的框架中实现了最佳的性能，没有额外的数据或者对于事物和材料的额外网络增强。更具体的是，AUNet 中仅仅采用了一个基于 ResNeXt-152-FPN[3]的单一模型。

过滤掉模型集成带来的增益，我们将 AUNet 与"PKU 360"团队进行了比较，他们采用了类似的框架但具有额外的功能。结果表明，我们的算法表现优于它们，特别是在 PQ^{St} 中，绝对改进率为**4.9**%。此外，AUNet 超越了以前的端到端方法，即 JSIS-Net [9]，具有**19.3**% 的绝对改善，证明了该方法的有效性。在表 2中，很明显 AUNet 在识别事物和材料之间有很好的平衡，即使在与挑战获胜者（没有额外数据）进行比

[2]我们对 RoIUpsample 层中的空白区域采用零填充，导致了背景上下文的空白。这种情况需要在未来进行观察。

[3]我们使用带有变形转换 [8] 和非本地区块 [35] 的 $64 \times 4d$ 变种 [**?**]。

表 2. 在 MS-COCO 2018 *test-dev* 中的全景质量 (%)。这里“额外数据”表示在训练中使用的额外数据集，“e2e”表示为事物和材料预测使用统一框架，“$\text{enhance}_{\text{Th}}$”和“$\text{enhance}_{\text{St}}$”分别表示在网络中为事物和材料使用了其他增强技术。PQ^{Th} 和 PQ^{St} 分别表示事物和材料的 PQ 结果。我们的网络没有额外数据或网络增强.

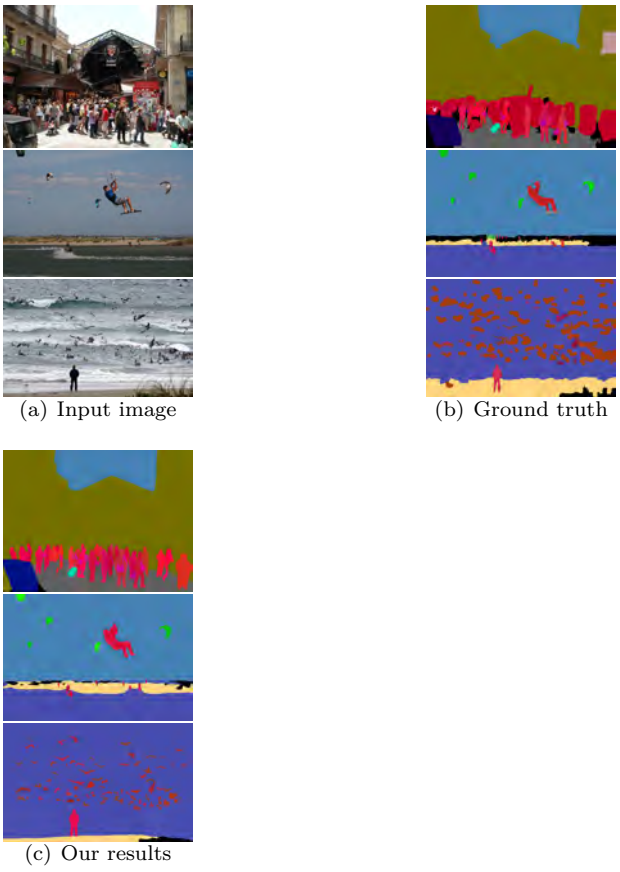
Method	backbone	extra data	e2e	$\text{enhance}_{\text{Th}}$	$\text{enhance}_{\text{St}}$	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
Megvii (Face++)	ensemble model	✓	✗	✓	✓	53.2	83.2	62.9	62.2	85.5	72.5	39.5	79.7	48.5
Caribbean	ensemble model	✗	✗	✓	✓	46.8	80.5	57.1	54.3	81.8	65.9	35.5	78.5	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 [9]	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**

表 3. 在 Cityscapes *val* 集中的全景质量 (%)。PQ^{Th} 和 PQ^{St} 分别表示事物和材料的 PQ 结果。我们结果与自底向上法（第一排）进行了比较。我们的 $_{\text{equ}}$ 表明所有事物训练时在背景分支中都被看作一类。

Method	backbone	PQ	PQ^{Th}	PQ^{St}	AP	mIoU
DWT [2]	VGG16	-	-	-	21.2	-
SGN [?]	VGG16	-	-	-	29.2	-
Li *et. al.* [23]	ResNet-101	53.8	42.5	62.1	28.6	-
Mask R-CNN [13]	ResNet-50	-	-	-	31.5	-
$\textbf{Ours}_{\text{equ}}$	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**

较时也是如此。这是由于统一框架的引入和注意引导模块建立了互补关系，第 4.2节对此进行了详细阐述。图 7给出了使用我们提出的 AUNet 达到最佳性能的直观展示。

Cityscapes: 我们将我们提出的方法与领先的自底向上法和表 3中的 Mask R-CNN 进行比较。首先，我们使用了和在 MS-COCO 数据集中相同的训练策略，这意味着所有的事物在背景分支中都被看作一个类，表示为 $\textbf{Ours}_{\text{equ}}$。然而，这种策略不如使用所有 19 个语义标签时的策略，如表 3所示。此外，MAM（被证明会降低 Cityscapes 中的 PQ）在最终结果中被禁用。我们认为这种性能下降是由于信息不一致1导致的，这意味着错误的事物预测可能会给材料预测提供错误的线索。总体而言，所提出的方法超过了先前的先进技术 [23]，绝对差距为 5.2%。



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

图 7. AUNet 在 MS-COCO *val* 集上的实例结果。其识别事物的表现甚至优于人为注释。同类的事物使用同一种色系但显示为不同的色彩强度。

5. 结论

本文介绍了 AUNet，一个统一的全景分割框架。与先前方法的主要区别在于我们将前景（实例级）和背景（语义级）分割统一到一个模型中，因此前景分支（通常能被更好的优化）能通过两个注意资源（比如建议关注模块和掩膜关注模块）协助背景分支，他们分别提供了对象级和像素级的指导。在实验过程中，我们在 MS-COCO 数据集中实现了最高水平的准确度增益。

我们的研究提供了一个重要信息：在视觉任务中，根据特性将目标划分为几个子类通常是有益的，因此可以跨子类传播互补信息以帮助场景理解。作为一项

新任务的全景分割提供了前景事物和背景材料之间的自然划分，但更多的可能性仍未被探索，并将在未来进行研究。

6. 致谢

我们衷心感谢 Jiagang Zhu 和 Yiming Hu 提出的宝贵建议。这项工作得到了中国国家重点研发计划第 2018YFD0400902 号和国家自然科学基金项目第 61573349 号的资助。

参考文献

[1] Anurag Arnab and Philip H S Torr. Pixelwise instance segmentation with a dynamically instantiated network. pages 879–888, 2017.

[2] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. pages 2858–2866, 2017.

[3] Liangchieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[4] Liangchieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[5] Liangchieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[6] 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. 2016.

[7] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Honglin Hu, and Yichen Wei. Deformable convolutional networks. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[9] Daan De Geus, Panagiotis Meletis, and Gijs Dubbelman. Panoptic segmentation with a joint semantic and instance segmentation network. *arXiv: Computer Vision and Pattern Recognition*, 2018.

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2018.

[11] Ross B Girshick. Fast r-cnn. pages 1440–1448, 2015.

[12] Ross B Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. pages 580–587, 2014.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. pages 1026–1034, 2015.

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

[16] http://cocodataset.org/panoptic leaderboard. Coco: Panoptic leaderboard. 2018.

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

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv: Learning*, 2015.

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv: Computer Vision and Pattern Recognition*, 2015.

[21] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2018.

[22] Alexander Kirillov, Kaiming He, Ross B Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2019.

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

[24] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[25] Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin, and Shuicheng Yan. Proposal-free

network for instance-level object segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2015.

[26] Tsungyi Lin, Piotr Dollar, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. pages 936–944, 2017.

[27] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. pages 740–755, 2014.

[28] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[29] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. pages 8759–8768, 2018.

[30] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. pages 708–724, 2018.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015.

[33] 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. *International journal of computer vision*, 115(3):211–252, 2015.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

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

[36] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

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

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

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018.

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

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

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

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

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