Cascaded Context-Aware Instance Segmentation with Transformer-Encoder for Adverse Weather Condition

Segmentasi Instansi Berbasis Konteks Bertingkat dengan Transformer-Encoder untuk Kondisi Cuaca Buruk

Muhamad Amirul Haq[1*], Barkah Rizkananda[2], Le Nam Quoc Huy[3], Fiki Andrianto[4]

[1,2,4] *Computer Science Department, Faculty of Engineering, Universitas Muhammadiyah Surabaya, Indonesia*
[3]*Mechanical Engineering Department, National Taiwan University of Science and Technology, ROC Taiwan*
[*]Coresponding author
E-mail addresses: amirulhaq@ft.um-surabaya.ac.id[1], barkah.rizkananda@ft.um-surabaya.ac.id[2], d10903816@mail.ntust.edu.tw [3], fikiandrianto@um-surabaya.ac.id[4]

Abstract. *Localizing objects from an image has been a vital part in autonomous driving since object localization performance directly correlate with the safety of the passenger. Robust and accurate object localization that can adapt to any driving environment has always been improved to ensure a safe and reliable system. In this work, we propose CBNet, a two-stage instance segmentation network for an autonomous driving environment. The network leverages a powerful transformer network as the feature extractor to improve performance. In addition, our proposed network utilizes a cascade design for both the object proposal network and the region-of-interests classifier. The cascade design addresses the issue of degrading detections over a high detection threshold. Moreover, we implement shape and edge-aware losses for the segmentation mask and end-to-end knowledge distillation strategy during training to improve the robustness of the network in extreme conditions. Our proposed network achieves 6.5 AP and 5.7 mIoU improvement from the prior methods in Cityscapes driving dataset. Furthermore, we evaluate our network in Foggy Cityscapes dataset to ensure the robustness of our network in extreme conditions. CBNet is able to improve the performance of prior methods by 7.7 AP and 6.7 mIoU in Foggy Cityscapes dataset.*
Keywords: Object detection, Deep learning, Edge aware, Autonomous driving

Abstrak. *Melokalisasi objek dari gambar telah menjadi bagian penting dalam berkendara otonom karena kinerja lokalisasi objek berkorelasi langsung dengan keselamatan penumpang. Lokalisasi objek yang kuat dan akurat yang dapat beradaptasi dengan lingkungan berkendara apa pun selalu ditingkatkan untuk memastikan sistem yang aman dan andal. Dalam karya ini, kami mengusulkan CBNet, jaringan segmentasi instans dua tahap untuk lingkungan berkendara otonom. Jaringan tersebut memanfaatkan jaringan transformator yang kuat sebagai ekstraktor fitur untuk meningkatkan kinerja. Selain itu, jaringan yang kami usulkan menggunakan desain kaskade untuk jaringan proposal objek dan pengklasifikasi wilayah minat. Desain kaskade mengatasi masalah deteksi yang menurun di atas ambang deteksi yang tinggi. Selain itu, kami menerapkan kerugian bentuk dan tepi untuk masker segmentasi dan strategi penyulingan pengetahuan ujung ke ujung selama pelatihan untuk meningkatkan ketahanan jaringan dalam kondisi ekstrem. Jaringan yang kami usulkan mencapai peningkatan 6,5 AP dan 5,7 mIoU dari metode sebelumnya dalam dataset berkendara Cityscapes. Lebih jauh, kami mengevaluasi jaringan kami dalam dataset Foggy Cityscapes untuk memastikan ketahanan jaringan kami dalam kondisi ekstrem. CBNet mampu meningkatkan kinerja metode sebelumnya sebesar 7,7 AP dan 6,7 mIoU dalam dataset Foggy Cityscapes.*
Kata kunci: Deteksi objek, Pembelajaran mendalam, Sadar tepi, Mengemudi otonom

INTRODUCTION

Autonomous driving vehicles rely heavily on object localization to navigate through the road. The safety and robustness of an autonomous driving system can be determined by the accuracy of the localization algorithm and its robustness in any outdoor situation. Different autonomous driving systems utilize a combination of different modalities to locate objects and perceive their surrounding, but a camera-based system is almost irreplaceable [1]. In camera-based systems for an autonomous driving application, extreme lighting and weather conditions can produce noises and obstruct object detection. Therefore, it is compulsory to develop a localization algorithm that is capable of addressing these extreme conditions.

In recent years, deep-learning-based object detection algorithms have become the preferred method due to their exceptional performance. The early object detection algorithm employs a sliding window to generate millions of candidate bounding boxes in every possible

size and location on an image [2]. Afterward, all of the bounding boxes will be classified using a classifier [3]. However, classifying a large number of candidate bounding boxes is time-consuming and inefficient. Thus, subsequent researches are dedicated to develop an object proposal algorithm that can reduce the number of candidate bounding boxes into a reasonable number [2]. The algorithm aimed at reducing the number of candidate bounding boxes needs to have a high-recall rate to not miss any object of interest. Needless to say, the more candidate bounding boxes passed to the classification network, the more accurate the overall detection quality will be. Therefore, object proposal algorithms have to consider the trade-off between high detection quality and detection speed. We can see it in Figure 1.

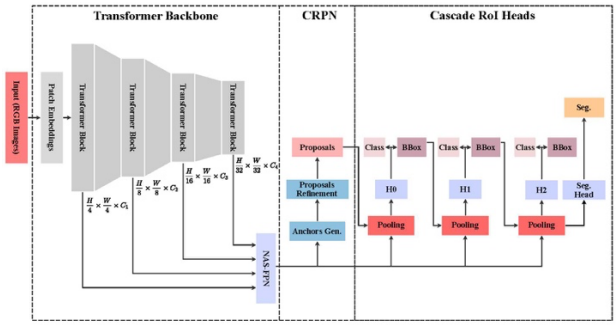

Figure 1. Our proposed network utilizes two-stage approach. Mix-Vision Transformer (MiT) is utilized as the feature extractor, whereas the region proposal network utilizes center-point detection to generate anchor-free proposals that will be classified by the region of interest (RoI) heads.

Prior works utilize focal loss (FL) [4], feature pyramid network (FPN) [5], [6], You Only Look Once (YOLO) [7], Single-shot Detector (SSD) [8], and visual cues-based objectness evidences [9], [10], [11] to obtain high quality candidate bounding boxes. FL, FPN, YOLO, and SSD utilize CNN's powerful representation ability to extract features from an image. These methods are also known as the top-down classification approach. Nevertheless, the top-down approach needs a large amount of training data and tends to generalize poorly on unseen object [9]. In contrast, objectness evidences method measures objectness based on low-level visual cues such as saliency [9], superpixels straddling [10], and edge density [11]. These methods are also known as the bottom-up approach. The bottom-up approach assumes that all objects of interest share common visual properties that distinguish them from the background. Consequently, these rule-based bottom-up algorithms can distinguish object boundaries with less training data and generalize better than the top-down approach. However, prior bottom-up approaches do not benefit from the abundance of data since they cannot learn new features as a deep-learning model does. Additionally, our study has found that edge density provides the best time-quality ratio among other bottom-up approaches.

In accordance with the above observation, we propose an shape and edge-aware instance segmentation algorithm that incorporates additional training strategies, such as shape-aware, edge-aware losses, and end-to-end knowledge distillation strategy. Our proposed network leverages the powerful transformer-based feature extraction as the backbone followed Neural Architecture Search FPN (NAS-FPN) [6] as the adapter for the subsequent cascaded region proposal network. The detection head consists of cascaded classification networks with two kinds of outputs, one for producing bounding boxes and the other to produce segmentation maps. The network will be tested in the Foggy Cityscapes dataset to measure its robustness in extreme weather conditions [12]. Our experiments show that the proposed network is able to consistently achieve state-of-the-art performance in both normal and extreme weather conditions. Lastly, the contribution of this work can be summarized in the following points: (a) This work proposed a robust object detection algorithm with state-of-the-art performance in autonomous driving vehicles. Our proposed network is tested in various weather conditions and is able to maintain its accuracy despite the challenging visual conditions. (b) The proposed network adopts a powerful but lightweight transformer backbone for the feature extractor. Additionally, using the NAS-FPN, we adapt the output of the transformer such that its features can be used by the subsequent cascade object proposals network and detection heads. (c) The proposed network utilizes a novel shape and edge-aware region proposal network and end-to-end knowledge distillation

learning strategy that can provide additional contexts during training. The edge prediction is generated alongside the segmentation mask to ensure the algorithm can run efficiently. (d) A high-resolution mask that helps the network learn fine-grained features is incorporated during training, specifically in dense regions with many objects. This region-enhanced detection strategy can guide the network to focus on the most salient parts of the image.

METHODOLOGY

The proposed network architecture is illustrated in Figure 1. Our proposed approach can be categorized as two-stage object detection with a cascaded region of interest (RoI) heads. The transformer and FPN backbone acts as the features extractor and pass these features to the RPN and RoI heads. Then, the RPN generates the initial class-agnostic bounding boxes as proposals to the RoI heads. Afterward, the RoI heads determine which classes each proposal belongs to. Meanwhile, in the last stage of the RoI head, the instance segmentation head will generate the segmentation masks to refine the bounding box detection further. Lastly, we will explain each component of the proposed network in detail in the subsequent section.

A. Transformer Backbone

The proposed backbone and feature extractor is inspired by Mix-Vision Transformer (MiT) network [13]. There are three main improvements we made in the MiT architecture. First of all, we substitute the activation function to Mish [14]. Mish is more robust than the standard ReLU because it can retain small gradient for negative inputs. Secondly, we integrate neural architecture search feature pyramid network (NAS-FPN) [6] to adapt its output for the detection heads. Since MiT was originally intended for a semantic segmentation network, NAS-FPN is needed to adapt the outputs of the transformer blocks and pass them into the detection heads. In the NAS-FPN, the output height and width of each transformer block are interpolated to $\frac{H}{4} and \frac{W}{4}$, respectively. Afterward, each of those outputs is passed into 2D convolution layers. The output of the NAS-FPN is a tensor with the dimension of $\frac{H}{4} \times \frac{W}{4} \times 256$. Lastly, we utilize contextual-spatial patch embedding (CSPE) [15] instead of the overlap patch embedding (OPE) that is used in the original work [13]. The main advantage of CSPE is its ability to change the positional encoding based on the input feature. The embedding of each image patch x_{ij} can be formulated as follows ini Formula 1:

$$P_{ij} = \frac{(x_i \boldsymbol{W}^Q)(x_j \boldsymbol{W}^K)^T + (x_i \boldsymbol{W}^Q)\boldsymbol{r}_{ij}^T}{\sqrt{d_z}} \quad (1)$$

where W^Q and W^K are unique and learnable parameters. Meanwhile, d_z denotes the output sequence from self-attention.

Lastly, MiT transformer has five variants with different sizes on their embedded dimensions. In Section 3, we shall compare the results of our models with different variants. MiT-B0 has the lowest number of parameters and the fastest among other variants. Meanwhile, the most accurate is MiT-B5 which is also the slowest and has the highest number of parameters.

B. Cascaded Region Proposal Network (CRPN)

Prior works on RPN employ heuristic anchor generation algorithms to produce the initial object proposals in an arbitrary location. The downside of such algorithms is the reliance on predetermined shapes and aspect ratios that requires prior domain-specific knowledge. Moreover, the anchors need to be aligned with the features from the backbone network. Consequently, conventional anchor-based RPN methods introduce unnecessary inductive biases and poorly aligned proposals that can hamper network training. To address this problem, we utilize cascade region proposal network (CRPN) which generates dynamic proposals over an offset field using adaptive convolution [16].

CRPN is expected to generate a set of n bounding boxes prediction. The class-agnostic CRPN consists of two parts, an anchor generator and a proposal refinement network. In the anchor generation part, a grid that is defined by the kernel size of dilated convolution produces a set of proposals. Each proposal is loosely evaluated

according to the intersection over union (IoU) with the ground-truth. Then, k-number of proposals are selected to be passed on to the class-agnostic proposal refinement network. Afterward, the IoU loss ($\mathcal{L}_{\text{IoU}} = 1 - IoU$) is calculated for each proposal to optimize the anchor generator part. In our proposed method, we empirically set k=1,000.

Meanwhile, in the second part, the proposal refinement network measures the objectness score of each proposal. We substitute the loss function from the original CRPN to focal loss (2) because we generate more proposals in the prior stage. Moreover, unlike the $\mathcal{L}_{\text{IoU}}$ in the first stage of the CRPN, the loss function utilized to refine the proposal is generalized IoU loss ($\mathcal{L}_{\text{GIoU}}$) as it offers better stability and correlation between ground-truth and prediction [17]. Finally, the overall loss calculated in the object proposal can be formulated as (3) where λ_{IoU}, λ_{GIoU} = 10 and λ_{obj} = 1, we can see in Figure 2.

$$\mathcal{L}_{obj} = \frac{1}{\text{H}/4 \cdot \text{W}/4} \sum_{\text{u}=1}^{\text{H}/4} \sum_{\text{v}=1}^{\text{W}/4} \text{FL}\left(\text{O}(\text{u},\text{v}), \widehat{\text{O}}(\text{u},\text{v})\right) \tag{2}$$

$$\mathcal{L}_{\text{CRPN}} = \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}} \tag{3}$$

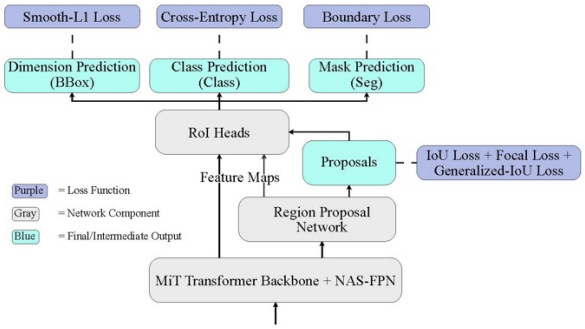

Figure 2. The loss functions utilized during training and their corresponding outputs.

C. Region of Interests (RoI) Heads

The RoI heads determine which classes each proposal corresponds with. The proposed method utilized multi-stage cascaded RoI heads architecture. Additionally, an edge-aware instance segmentation head is integrated into the last RoI head to refine the bounding box prediction. The network expects two outputs, bounding boxes predicted by H0, H1, and H2; and segmentation maps predicted by the segmentation head.

C.1 Bounding Box Classification and Localization

Bounding boxes from H0, H1, and H2 are evaluated sequentially with an increasingly higher IoU detection threshold for each stage. The detection thresholds for H0, H1, and H2 are 0.5, 0.7, and 0.9, respectively. With the increasing detection threshold, the RoI head in the later stage is expected to be more accurate. To evaluate the output of each stage, the network utilizes binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$) on n-number of class-agnostic bounding boxes for the classification part, as written in (4).

$$\mathcal{L}_{\text{BCE}} = \frac{1}{\text{n}} \sum_{\text{i}}^{\text{n}} \widehat{\text{B}_\text{i}} \cdot \log(\text{B}_\text{i}) + \left(1 - \widehat{\text{H}_\text{i}}\right) \cdot \log(1 - \text{B}_\text{i}), \tag{4}$$

where B and $\hat{B}$ are the class prediction and ground-truth pair of each bounding box. Meanwhile, the bounding boxes' dimension is evaluated with smooth-L1 loss (\mathcal{L}_{dim}). Thus, the overall bounding box classification and localization loss function in the RoI heads can be formulated as (\ref{eq:roi_loss}), where λ_{BCE} and λ_{dim} are the weights for $\mathcal{L}_{BCE}$ and $\mathcal{L}_{dim}$, respectively.

$$\mathcal{L}_{\text{RoI}} = \lambda_{\text{BCE}}\mathcal{L}_{BCE} + \lambda_{\text{dim}}\mathcal{L}_{dim} \tag{5}$$

C.2 Boundary-Aware Instance Segmentation Mask

In extreme lighting and weather conditions, most visual features such as texture and color are unreliable. A standard cross-entropy loss only measures the loss on each pixel without considering the object's shape. Thus, it is not sufficient to train a network in extreme lighting and weather conditions. Hence, a boundary-aware object mask is proposed to guide the network to estimate the object of interest based on its shape and boundary to obtain higher-quality segmentation masks.

First of all, a shape-aware object mask considers the full shape of the object of interest despite visual obstruction caused by fog and insufficient lighting. By utilizing Dice Similarity Coefficient (DSC), the network takes into account the similarity of the overall shape

between each predicted instance segmentation mask (S) and the ground-truth ($\hat{S}$) based on their union normalized by the sum of their areas, as formulated in (6), like we see in Figure 3.

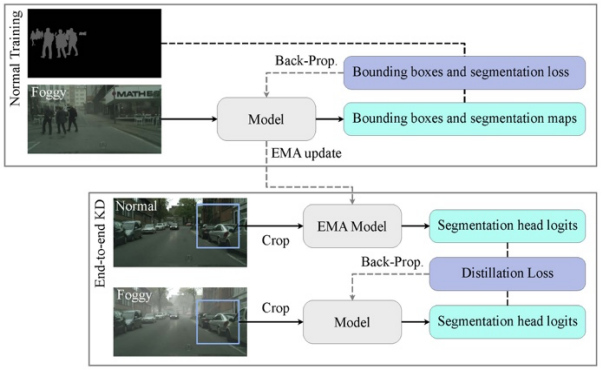

Figure 3. End-to-end knowledge distillation is performed when training the model in Foggy Cityscapes dataset. The bounding boxes and segmentation loss is described in (10), whereas the distillation loss is described in (11).

$$\mathrm{DSC}(S,\hat{S}) = \frac{2\sum(S \times \hat{S})}{\sum(S^2 + \hat{S}^2)} \tag{6}$$

The overall loss for all instances in the segmentation head can be calculated as the mean of DSC and pixel-by-pixel cross-entropy (CE).

$$\mathcal{L}_{\text{shape}} = \frac{1}{N}\sum^{N} \mathrm{CE}(S_N, \hat{S}_N) + \left(1 - \mathrm{DSC}(S_N, \hat{S}_N)\right) \tag{7}$$

Aside from the object's shape, foggy weather can also obstruct the visual edge or contour of an object, causing a degradation in the bounding box IoU. Therefore, a loss function based on Hausdorff distance (d_H) is added to improve the edge estimation between the prediction and ground-truth. Given a set of points along the contour of the predicted mask (A) and ground-truth (B), d_H can be formulated as (8). The edge loss itself can be simply calculated as $\mathcal{L}_{\text{edge}} = 1 - d_H$.

$$\mathrm{d_H} = \max\left\{\sup_{a\in A}\inf_{b\in B} \mathrm{d}(a,b), \sup_{b\in B}\inf_{a\in A} \mathrm{d}(a,b)\right\} \tag{8}$$

Finally, the overall loss function for the boundary-aware instance segmentation mask (9) can be formulated as the weighted sum of $\mathcal{L}_{\text{shape}}$ and $\mathcal{L}_{\text{edge}}$ with their respective weights, λ_{shape} and λ_{edge}.

$$\mathcal{L}_{\text{bound}} = \lambda_{\text{shape}}\mathcal{L}_{shape} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} \tag{9}$$

D. Total Loss Function

The loss functions are calculated from various outputs from each stage of the detection, as illustrated in Figure 2. The total loss is calculated as the sum of $\mathcal{L}_{\text{RPN}}$, $\mathcal{L}_{\text{RoI}}$, and $\mathcal{L}_{\text{bound}}$ as previously mentioned in (3), (5), and (9), respectively. Therefore, the total loss can be written as (10).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RoI}} + \mathcal{L}_{\text{bound}} \tag{10}$$

E. Training Strategies

Transformer-based models are known for being data-hungry. Despite its ability to scale better than CNN, transformers require more data to train and are slower to converge. Therefore, most transformers must be trained on huge data such as ImageNet before fine-tuned on the intended dataset [18]. For our backbone transformer, we utilize MiT pre-trained model using the weight from ImageNet [19]. Additionally, since our method is expected to be deployed in an outdoor environment, we have to consider some weather conditions such as rain and fog. As a result, we propose an end-to-end knowledge distillation strategy to allow our model to perform well in a dataset with extreme weather, such as foggy Cityscapes [12].

The proposed end-to-end knowledge distillation is illustrated in Figure 3. The main model is evaluated and updated normally using Foggy Cityscapes images with loss function written in (10). After each training step, the teacher model is updated with the exponentially moving average (EMA) algorithm to stabilize the pseudo-labels. Afterward, we load regular and Foggy Cityscapes from the same scene. The images are cropped to reduce memory consumption and ensure the model learns the segmentation map effectively from the small high-resolution patches rather than the long-range context provided by the whole image. Because the normal and synthesized fog images are supposed to be the same image, both teacher/EMA and student/main models are expected to have the same output logits. Therefore, we retrieve the output logits of the mask head of the EMA model (y_t) and use them as pseudo-labels for the main

model's logits (y_s) with cross-entropy loss (CE) as the loss function [20]. Thus, the distillation loss can be formulated as (11) where T denotes the temperature parameter and is dynamically set to the maximum softmax probability (τ), and we can see in Figure 4.

$$\mathcal{L}_{\text{distillation}} = \text{CE}\big(\sigma(y_t; T = \tau), \sigma(y_s; T = \tau)\big) \quad (11)$$

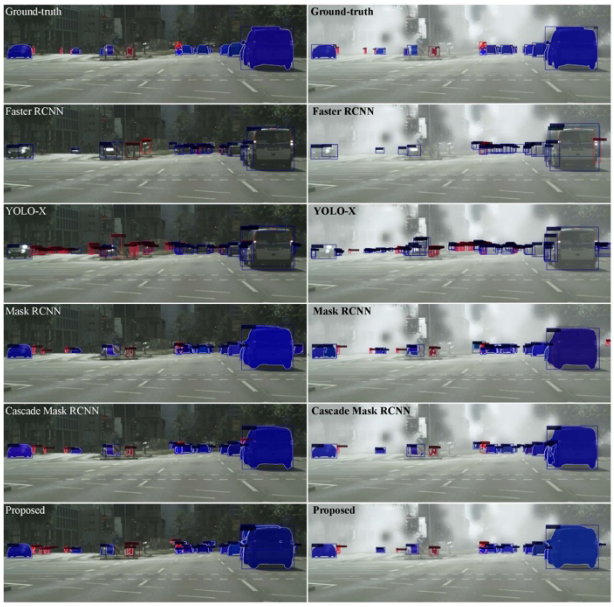

Figure 4. Visual comparison of prior and proposed methods. Images on the left column are from the regular Cityscapes, whereas images on the right column are from the Foggy Cityscapes with the beta value set at 0.01.

RESULT AND DISCUSSION

The proposed method is trained on a Ubuntu 20.04 desktop with AMD Ryzen 9 3950X CPU, 128 GB RAM, and NVidia RTX 3090-24GB VRAM GPU. To obtain fair comparison and benchmarking, the prior and proposed methods are written using Pytorch [21] framework with MMDetection [22] deep learning toolbox. The proposed method is trained for 12 epochs with a polynomial scheduler and AdamW optimizer. Based on the results of several experiments, the best initial learning rate to train our model is 6×10^{-5}.

The proposed and prior methods are evaluated in the standard Cityscapes [23] and Foggy Cityscapes datasets. The cityscapes dataset is focused on autonomous driving development. The data is taken from cameras and Lidar mounted on a car to simulate a driving situation. Meanwhile, Foggy Cityscapes artificially generates foggy scenes from Cityscapes' images. The evaluation metrics used are AP across scales (mAP, mAP^S, mAP^M, mAP^L) which are built-in inside MMDetection toolbox and have been used in most modern object detection datasets [24]. In addition, mean intersection over union (mIoU) is also calculated for methods that are able to produce a segmentation map.

Table 1. Bounding box and instance segmentation evaluation on Cityscapes [23] dataset with eight classes. Best values are written in bold.

Method	Backbone	AP	mIoU
F-RCNN [25]	ResNext-101	13.2	-
YOLO-X [26]	CSPDarknet	26.3	-
M-RCNN [27]	GVT-S	30.7	27.4
Cas. RCNN [28]	Swin-S	34.9	30.2
CBNet (Ours)	MiT-B5	**41.4**	**35.9**

Table 2. Bounding box and instance segmentation evaluation on Foggy Cityscapes dataset with eight classes. In the right-most column, we listed the mIoU drop with respect to the results on the regular Cityscapes dataset, which are listed in Table 1.

Method	AP	mIoU	AP Drop
F-RCNN [25]	12.9	-	**0.3**
YOLO-X [26]	22.1	-	4.2
M-RCNN [27]	26.4	22.2	4.3
Cas. RCNN [28]	31.8	27.8	3.1
CBNet (Ours)	39.1	34.5	2.8
CBNet-DD (Ours)	**39.5**	**34.5**	1.9

A. Comparison with Prior Methods

To test the versatility and effectiveness of our cascaded boundary-aware network (CBNet), the proposed method is evaluated along with popular prior methods in object detection tasks using the newest backbone networks, as listed in Table 1 and 2. Aside from minor adjustments for training on the Cityscapes dataset, prior methods are run with their default settings.

A.1. Evaluation on regular Cityscapes.

Table 1 summarizes the detection results in Cityscapes dataset validation set. The AP scores obtained by our proposed method are significantly higher than prior methods, with a 6.5 improvement in the overall AP compared with the second best, Cascade Mask R-CNN (Cas. RCNN). The drawback of our network is the low inference speed. On the other hand, the one-stage object detection method, YOLO-X, obtains higher fps to make up for the low AP scores. Meanwhile, the cascade design

in the RoI heads of Cascade Mask RCNN and CBNet proves to be capable of removing many false positives, as shown in Figure 4. However, CBNet improves upon this further by also using the cascade design in the object proposal network. Consequently, the AP^S that represents small bounding boxes of distant objects in CBNet is significantly higher than prior methods, and we can see in Figure 5 and Figure 6.

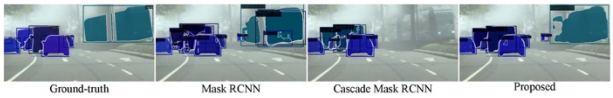

Figure 5. Visual comparison of prior and proposed methods' segmentation maps. The proposed method demonstrates its ability to accurately predict segmentation map on a challenging condition

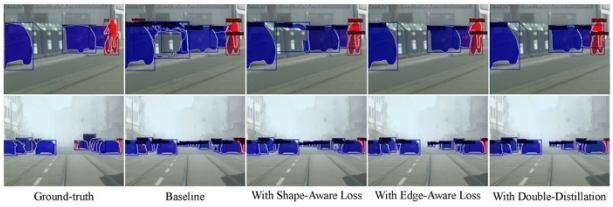

Figure 6. Visualization of the proposed shape and edge-aware losses along with knowledge distillation training. Each subsequent column starting from the "Baseline" adds to the preceding components, as listed in Table 3.

A.2. Evaluation on Foggy Cityscapes.

The main objective of the experiment in Foggy Cityscapes is to ensure the model is robust enough to be deployed during extreme weather. Therefore, the model must retain its detection capabilities in Foggy Cityscapes with the lowest AP drop possible. In Figure 5, the performance of instance segmentation methods in extreme weather is compared. CBNet is able to detect objects with high precision and recall rate due to the combination of the Cascade RPN and Cascade RoI heads. Meanwhile, Mask RCNN has low precision due to the low-quality proposals, whereas Cascade Mask RCNN has high precision but a low recall rate because most of the proposals generated by the RPN are not good enough for the Cascade RoI heads.

The quantitative assessments in Foggy Cityscapes are summarized in Table 2. Originally, the AP drop in baseline CBNet is 2.8%. However, after implementing shape-aware and edge-aware loss along with the end-to-end knowledge distillation (CBNet-KD), the proposed method is able to reduce the AP drop to 1.9%. Both CBNet and CBNet-DD are only inferior to Faster-RCNN in terms of AP drop. However, both models are vastly superior in terms of the overall AP scores. From the visual assessment provided in Figure 4, CBNet result is almost similar to Cascade Mask RCNN, safe for few misclassifications for small and distant objects.

B. Ablation Study

An ablation study is performed to understand the impact of the novel loss function and knowledge distillation training strategy. The ablation study is performed on the Foggy Cityscapes dataset so that knowledge distillation can be performed by using the model trained on the regular Cityscapes as the teacher. In the baseline CBNet, the segmentation head's losses are substituted by the standard cross-entropy loss.

B.1 Shape-aware and edge-aware loss.

The shape-aware loss successfully improves the bounding boxes AP by a small margin of 0.2 percent. However, shape-aware loss can remove most of the obvious false positives from the predictions, as shown in the first row of Figure 6. Combining these two loss functions improves the AP and mIoU by 0.6 and 0.9 percent, respectively. Moreover, shape-aware and edge-aware losses can improve the detection of distant objects, as illustrated in the second row of Figure 6, and we can see in Table 3.

Table 3. Ablation study of the proposed method and its components on Foggy Cityscapes dataset.

Ablation Settings			AP	mIoU
Shape-Aware Loss	Edge-Aware Loss	Knowledge Distillation		
-	-	-	38.6	33.6
✓	-	-	38.8 (↑0.2)	33.5 (↓0.1)
✓	✓	-	39.2 (↑0.6)	34.5 (↑0.9)
✓	✓	✓	39.5 (↑0.9)	34.5 (↑0.9)

B.2. End-to-end knowledge distillation training.

End-to-end knowledge distillation training simultaneously trains the model on regular and Foggy Cityscapes. Knowledge distillation training provides more stable training for the model trained on Foggy Cityscapes by providing additional regularization terms

based on the difference in their output logits. Thus, the model is less affected by the hyper-parameters, such as the initial learning rate and the random seed, and can minimize the AP drop between the two datasets. Unlike the noticeable improvement by shape-aware and edge-aware losses, the improvement by knowledge distillation training is more subtle but impactful nonetheless. In summary, using all three components, CBNet can improve both AP and mIoU scores by 0.9% and minimize the AP drop from extreme weather conditions to 1.9%.

CONCLUSION

In this work, a shape and edge-aware object detection framework is proposed. It leverages the powerful yet lightweight transformer network as a feature extractor and adapts the extracted features with NAS-FPN to improve the network's performance on multi-scale detection. The proposed network, CBNet, adopts the two-stage object detection approach with cascade RPNs as the object proposal and cascade RoI heads as the classification network. Most modern object detection frameworks can perform well in normal weather but perform significantly worse in extreme weather. This drawback could not be tolerated in vital real-world applications, such as autonomous driving. Thus, a benchmark and comparison with popular object detection frameworks have been performed to evaluate the effectiveness and robustness of the proposed network in normal and extreme weather conditions. Based on the overall AP score for bounding box prediction, CBNet obtains a significant 6.5% margin from the prior best-performing method, Cascade Mask RCNN, in the regular Cityscapes dataset. CBNet also performs better than Cascade Mask RCNN in instance segmentation tasks by a 5.7% margin. Meanwhile, in the Foggy Cityscapes dataset, CBNet-DD obtains 7.7% higher AP and 6.7% higher mIoU compared with Cascade Mask RCNN. Compared with prior methods, CBNet can minimize the AP and mIoU drop in extreme weather using the boundary-aware loss and the end-to-end knowledge distillation method. In the future, it would be beneficial to implement our proposed network on a more complex and challenging computer vision task, such as panoptic segmentation.

REFERENSI

[1] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, “Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 129–137.

[2] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, “What makes for effective detection proposals?,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 814–830, 2015.

[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” *arXiv preprint arXiv:1312.6229*, 2013.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[6] G. Ghiasi, T.-Y. Lin, and Q. V. Le, “Nas-fpn: Learning scalable feature pyramid architecture for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 7036–7045.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] W. Liu *et al.*, “Ssd: Single shot multibox detector,” in *European conference on computer vision*, Springer, 2016, pp. 21–37.

[9] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. Torr, “BING: Binarized normed gradients for objectness estimation at 300fps,” *Computational Visual Media*, vol. 5, no. 1, pp. 3–20, 2019.

[10] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[11] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, Springer, 2014, pp. 391–405.

[12] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[14] D. Misra, "Mish: A self regularized non-monotonic neural activation function," *arXiv preprint arXiv:1908.08681*, vol. 4, no. 2, pp. 10–48550, 2019.

[15] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10033–10041.

[16] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[17] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[18] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[19] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv.org. Accessed: Mar. 08, 2024. [Online]. Available: https://arxiv.org/abs/1503.02531v1

[21] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[22] K. Chen *et al.*, "MMDetection: Open MMLab Detection Toolbox and Benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[23] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[24] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[26] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[28] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relation- ships that could be construed as a potential conflict of interest.

Article History:
Received: 09 September 2024 | Accepted: 15 Oktober 2024| Published: 30 November 2024