Self-Supervised 3D Monocular Object Detection by Recycling Bounding Boxes

Sugirtha T[1], Sridevi M[1], Khailash Santhakumar[2], Hao Liu[3]
B Ravi Kiran[3], Thomas Gauthier[3] and Senthil Yogamani[4]
[1]NIT Tiruchirappalli, India [2]SASTRA University, India [3]Navya, France [4]Valeo, Ireland

Abstract

Modern object detection architectures are moving towards employing self-supervised learning (SSL) to improve performance detection with related pretext tasks. Pretext tasks for monocular 3D object detection have not yet been explored yet in literature. The paper studies the application of established self-supervised bounding box recycling by labeling random windows as the pretext task. The classifier head of the 3D detector is trained to classify random windows containing different proportions of the ground truth objects, thus handling the foreground-background imbalance. We evaluate the pretext task using the RTM3D detection model as baseline, with and without the application of data augmentation. We demonstrate improvements of between 2-3 % in mAP 3D and 0.9-1.5 % BEV scores using SSL over the baseline scores. We propose the inverse class frequency re-weighted (ICFW) mAP score that highlights improvements in detection for low frequency classes in a class imbalanced dataset with long tails. We demonstrate improvements in ICFW both mAP 3D and BEV scores to take into account the class imbalance in the KITTI validation dataset. We see 4-5 % increase in ICFW metric with the pretext task.

1. Introduction

3D object detection is a crucial perception task in modern autonomous driving applications, used upstream for scene understanding, object tracking and trajectory prediction and decision making. Initially, autonomous cars are equipped with LiDAR sensors and most 3D detectors rely on LiDAR data to perform 3D object detection. LiDAR provides precise distance measurement which makes it feasible to detect accurate 3D bounding boxes. But, they are expensive to be deployed in autonomous cars. Recent autonomous cars use single monocular camera and hence monocular 3D object detection (3D OD) became a research focus in computer vision community.

RTM3D [9] is a monocular 3D object detector based on the CenterNet architecture [19], we shall use this model as our baseline model to evaluate SSL methods for object detection. In this paper we evaluate the use of multi-object labeling pretext task proposed by authors in [7] as self-supervision to improve the 3D monocular object detection.

3D monocular detection requires an expensive annotation process. Self supervised learning methods provide auxiliary or pretext tasks that use cheaply available labels, to help the downstream primary task of monocular 3D-OD. In summary, the contributions of our paper are as follows: 1. We evaluate the performance of baseline RTM3D detector with self-supervised multi-object labeling pretext task under 2 settings (i) different number of random windows as a hyper-parameter (ii) Under the use of data augmentations along with self-supervision. 2. Propose the class frequency sensitive detection score (ICFW) that measures improvement in low frequency critical classes i.e. in pedestrian and cyclist classes.

Extensive analysis on KITTI dataset [4] demonstrates that our proposed data augmentations improve performance under various conditions: occlusions, contrasted/shadowed pixels, changing the diversity of viewpoints of objects seen in the dataset.

2. Related Work

2D OD on image plane is inadequate for reliable autonomous driving scenario because it does not provide an accurate estimation of 3D objects sizes and space localization. In other words, 2D OD methods have limited performance in following scenarios namely occlusion, object pose estimation and 3D position information. A 3D bounding box provides precise information about the size of the object and its position in 3D space.

3D Object Detection : 3D OD methods are usually part of the 4 following categories: (i) 2D proposal generation [6], (ii) Geometric constraints [20], (iii) Key-points detection [19] or (iv) Direct 3D proposal generation [10]. RTM3D[9] and its extension KM3D [10] using CenterNet [19] to regress a set of 9 projected keypoints corresponding to a 3D cuboid in image space (8 vertices of the cuboid and its center). They also perform direct regression for the object's distance, size and orientation. These values are then

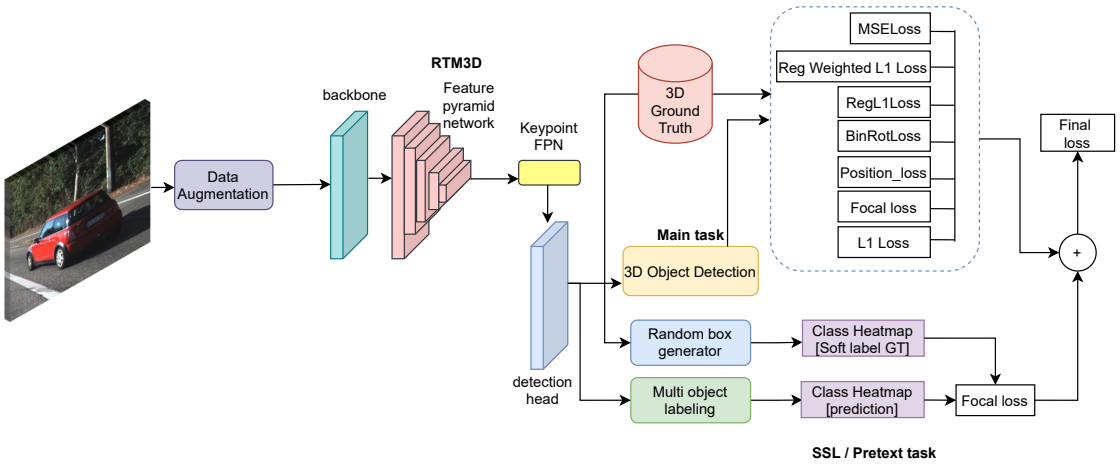


Figure 1: Self supervised learning setup with monocular 3D object detection as main task, and multi-object labeling as pretext task. Following authors in [7], we train the main and pretext task in a multi-task learning setup by summing the two losses with equal static weights.

used for offline initialization of an optimizer to estimate 3D bounding boxes under geometric constraints. CenterNet provides basic data augmentation such as affine transformations (shifting, scaling) and random horizontal flipping. Over these, KM3D [10] adds coordinate independent augmentation via random color jittering.

SMOKE [11] regresses 3D bounding box directly from image plane which eliminates 2D bounding box regression. It represents an object by a single keypoint and these keypoints are projected as 3D center of each object. Mono3D [2] is a region proposal based method that uses semantics, object contours and location priors to generate 3D anchors. It generates proposal by performing exhaustive search on 3D space and uses non-maximal suppression for filtering. SMOKE augments the training samples with horizontal flipping, scaling and shifting. GS3D [8] predicts the guidance of cuboid and performs feature extraction by projecting region of guidance. GS3D performs monocular 3D detection without augmenting the training data.

2.1. Self-supervised learning

Supervised learning requires human endeavour to create high quality annotations whereas self-supervised learning (SSL) creates labels by their own models without need for human labour. In the area of computer vision, various clues like optical flow [14], tracking [16], inpainting [15], sound [13] and colorization [18] are being utilized as a pretext task which help the primary task generalize better. Authors [12] train the network to solve jigsaw puzzles and fine tune it for object localization and detection where [5] differentiates real and artifact images and transfer it to object detection. Authors in [18] trained the model for colorization purpose and modulated it for object detection. Authors in [1] estimate 3D object properties such as location, dimension etc. via SSL and differentiable rendering, which eliminates the need for 3D annotations. [7] created three auxiliary tasks which reused bounding box labels for self supervision in order to improve 2D object detection performance. In this paper, we reiterate this recycling bounding box task [7], reusing the same method to achieve better 3d-monocular object detection.

3. SSL for Object detection

Authors in [7], discuss the use of three different pretext tasks to improve the performance of the main/downstream object detection task. Key pretext tasks include :

- **Multi-Object Labeling (MOL)** : To handle the imbalance in foreground and background w.r.t an object detection task, authors propose the usage of a random window (W) with partial or complete intersection with objects in the given sample (image I, bounding boxes $\{B\}$) pair to artificially increase the number of bounding boxes. The multi-object soft label corresponding to the random window assigns area ratios of each class's GT boxes B within the random window W. Only the classifier head of Faster R-CNN and R-FCN models were trained in a multi-task setting (detection as the main task, random window classification as an auxiliary task). We have demonstrated this on a KITTI image sample in figure 2.
- **Closeness & FG segmentation** : Authors in [7] also proposed a closeness label that measures the distances from the center of a GT box to those of other GT boxes.

Figure 2: Visualization for 3 different random windows selected from the input image domain, along with their soft label generated by using the proportion of different classes & background within each window.

While the segmentation task performs binary foreground/background segmentation, where foreground is created from union of regions of all bounding boxes.

In our study, we only evaluate the performance of the MOL pretext task.

3.1. Data augmentations

Box-Mixup : Motivated by the work on Mixup data augmentation [17] Box-MixUp is proposed to augment an image with object patches from other images, thus providing the same advantages of MixUp but localized over multiple regions in the image. The augmented sample can be expressed as:

$$\begin{aligned} \widetilde{x} &= (0.5x_A + 0.5x_B) \cdot M_B + x_A \cdot (1 - M_B) \\ \widetilde{y} &= y_A \cup y_B \end{aligned} \tag{1}$$

Box-Cut-Paste : Following cut-paste data augmentation [3] in Box Cut-Paste, we paste the pixels under the bounding box mask from one image onto the reference image. This can be expressed as :

$$\begin{aligned} \widetilde{x} &= x_A \cdot (1 - M_B) + x_B \cdot M_B \\ \widetilde{y} &= y_A \cup y_B \end{aligned} \tag{2}$$

The augmentations are demonstrated in figure 3.

4. Experiments

Baseline Model : RTM3D [9] is a real-time network that is based on the CenterNet architecture, which enables both fast training cycles and small inference time. Authors have already provided a set of baseline data augmentation which include flip, affine transformations, stereo dataset augmentation using the left/right images in the KITTI dataset.

In our study we present the following comparisons. The RTM3D baseline model trained with multi-object labeling as pretext task in two settings: 1. with pretext task alone 2. pretext task along with data augmentation.

SSL task: Multi-Object-Labelling: Self supervised learning setup for our evaluation is shown in Fig. 1. We evaluate the performance of the Multi-Object-Labelling pretext task proposed by authors in [7]. We evaluated the performance at different number of random windows as a hyper-parameter : two, four, eight, and six-teen random windows. We evaluated uniform different distribution of scales of random windows.

5. Results

We evaluate our models on the KITTI 3D detection benchmark which consists of 7,481 labeled training samples and 7518 unlabeled testing samples. Since the ground truth labels for the test set are not available, we evaluated our model by splitting the training set into 3711 training samples and 3768 validation samples. We experiment with ResNet-18 as the backbone. We implemented our deep neural network in Pytorch and trained using Adam optimizer with learning rate of 1.25*1e-4 for 200 epochs. We trained our network with a batch size of 16. Our model achieved best speed with 33 FPS on a NVIDIA GTX 2080Ti GPU.

Metrics : The KITTI benchmark evaluates the models by Average Precision (AP) of each class (Car, Pedestrian and Cyclist). We use the Mean Average Precision (mAP), the mean value of the Average Precision (AP) over all classes using equation (3) :

$$\text{mAP}_{3D} = \frac{1}{|C|} \sum_{c \in C} \text{AP}_c \tag{3}$$

where $C = \{\text{car, pedestrian, cyclist}\}$.

Inverse Class Frequency Weighted (ICFW) mAP : We introduce a new metric that is used to demonstrate gains in a class imbalanced KITTI dataset. As the proposed metric is weighted by inverse of the class frequency, the gains over minority classed are favoured. The relative frequency (denoted by f_c and in blue) of car, pedestrian and cyclist classes in the validation classes are shown in Table 2. This is evaluated by the following formula in equation (4):

$$w_c := \frac{f_c^{-1}}{\sum_{c \in C} f_c^{-1}} \in [0, 1] \text{ and } \sum_{c \in C} w_c = 1 \tag{4}$$

The values of w_c are shown in Table 2. Now the ICFW

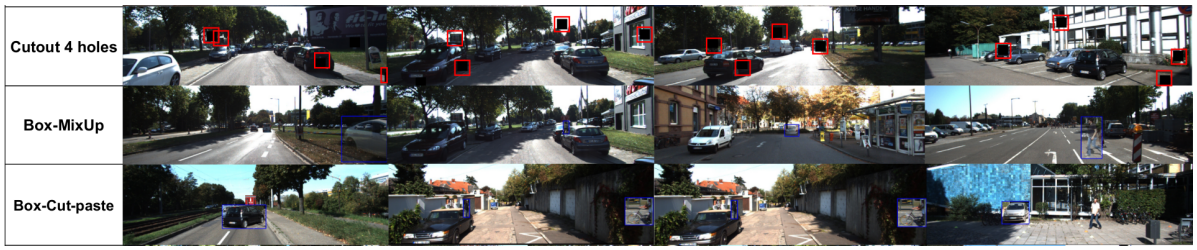


Figure 3: The three data augmentations : Box-mixup, Box-Cutpaste and Cutout that were used along with MOL pretext tasks.

Table 1: mAP and ICFW mAP scores for both 3D and BEV detection bounding boxes. Green refers to positive gains, while red refers to negative drops in performance over the baseline.

IoU=0.5	**mAP**$_{2D}$	**mAP**$_{BEV}$	**mAP**$_{3D}$	**ICFW mAP**$_{2D}$	**ICFW mAP**$_{BEV}$	**ICFW mAP**$_{3D}$
Baseline (B)	41.44	21.17	19.12	33	15.1	14.65
Self-Supervised Learning (SSL)						
B + 8W	0.85	0.53	0.46	0.83	0.7	0.54
B + 16W	0.59	-0.75	-0.59	0.57	-1.88	-1.73
B + 32W	1.4	0.29	0.12	1.75	0.12	-0.17
Data Augmentation (DA)						
B + Cutout4	-0.91	0.11	-0.71	-2.79	0.15	-0.54
B + BoxMixup	0.39	0.29	0.21	0.53	0.12	0.04
B + Cutpaste	1.63	1.10	0.34	3.22	1.91	0.49
SSL + DA						
B + 16W + Cutout	1.54	1.27	0.43	2.17	**2.81**	1.02
B + 16 W + box mixup	1.2	1.67	1.66	1.42	2.57	**2.59**
B + 16 W + boxmixup cutout	**3.51**	**1.84**	1.01	**5.57**	2.53	1.02
B +16 W + cutpaste cutout	2.87	1.38	**2.26**	5	1.13	1.19
B +16 W + cutpaste	0.98	0.67	0.72	1.61	0.65	0.73

mAP is evaluated with equation (5) as:

$$\text{ICFW mAP}_{3D} = \sum_{c \in C} w_c \text{AP}_c \tag{5}$$

Table 1 shows the results of DA and SSL evaluated on KITTI dataset. It tabulates the mAP2D, mAP3D & mAP-BEV over all classes for IoU=0.5.

SSL vs SSL+DA : We observe an improved performance in both mAP$_{3D}$ and mAP$_{BEV}$ scores when using the MOL-SSL pretext task. In our study we evaluate the effect of 3 data augmentation schemes on the pretext task and thus the performance of the main task. They are detailed and demonstrated in Figure 3. The combination of SSL-MOL along with cutout and box-mixup augmentations provide the largest of gains. This is attributed to the gains that cutout provide for truncated objects, and box-mixup for varied foreground/background variations.

DA vs SSL+DA : The proposed data augmentations alone without pretext tasks do also provide gains over the baseline while remaining lower in performance than combination SSL+DA. We hypothesize that DA provide augmented samples that help to improve the pretext task by imposing that the n/w should learn features to classify random windows in augmented samples as well as the primary 3D detection task. Cutout alone performed a bit worse over baseline, while in combination with the MOL pretext task, performance is greater than Cutout/SSL tasks alone.

6. Conclusion

In this study we evaluated the performance of the multi-object labeling pre-text task training in conjunction with the main monocular 3d-object detection task. As expected due to correlation between the classification task and the 3d localization task, we see an improvement of 1-2 points in the mAP scores for 3D and BEV metrics, besides the evident gains in mAP 2D scores. We also evaluated the performance of using data augmentation schemes along with the pretext task, in our case we evaluated the performance of box-mixup (a bounding box version of instance mixup), cutout and box-cut-paste, and their combinations. We observed that the addition of data augmentation strategies improved the diversity of samples received by the MOL pretext task, and thus directly contributed in improving the performance of the main 3d detection task.

A. Acknowledgements

Authors would like to thank Navya to have provided the opportunity to perform applied research in key engineering problems.

B. Class frequencies in KITTI-3D

Table 2: Class freq. & inverse weights on the validation set.

Class	**Car**	**Pedestrian**	**Cyclist**
Frequency f_c	0.82	0.12	0.05
Inverted w_c	0.04	0.27	0.69

C. Example results

Fig. 4 shows the detection results of our proposed Box-Mixup data augmentation on KITTI in different scenarios. Ex. Occluded objects, missed detections and misclassification by baseline. It shows baseline, Box-Mixup predictions on left panel and their corresponding BEV representations on right panel.

References

[1] D. Beker, Hiroharu Kato, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Monocular differentiable rendering for self-supervised 3d object detection. In *ECCV*, 2020. 2

[2] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[3] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1310–1319, 2017. 3

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[5] S. Jenni and P. Favaro. Self-supervised feature learning by learning to spot artifacts. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2733–2742, 2018. 2

[6] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[7] Wonhee Lee, Joonil Na, and Gunhee Kim. Multi-task self-supervised object detection via recycling of bounding box annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3

[8] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[9] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 644–660, Cham, 2020. Springer International Publishing. 1, 3

[10] Pei-Xuan Li. Monocular 3d detection with geometric constraints embedding and semi-supervised training. *ArXiv*, abs/2009.00764, 2020. 1, 2

[11] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020. 2

[12] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[13] Andrew Owens, Jiajun Wu, Josh H. McDermott, W. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2

[14] Deepak Pathak, Ross B. Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6024–6033, 2017. 2

[15] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 2

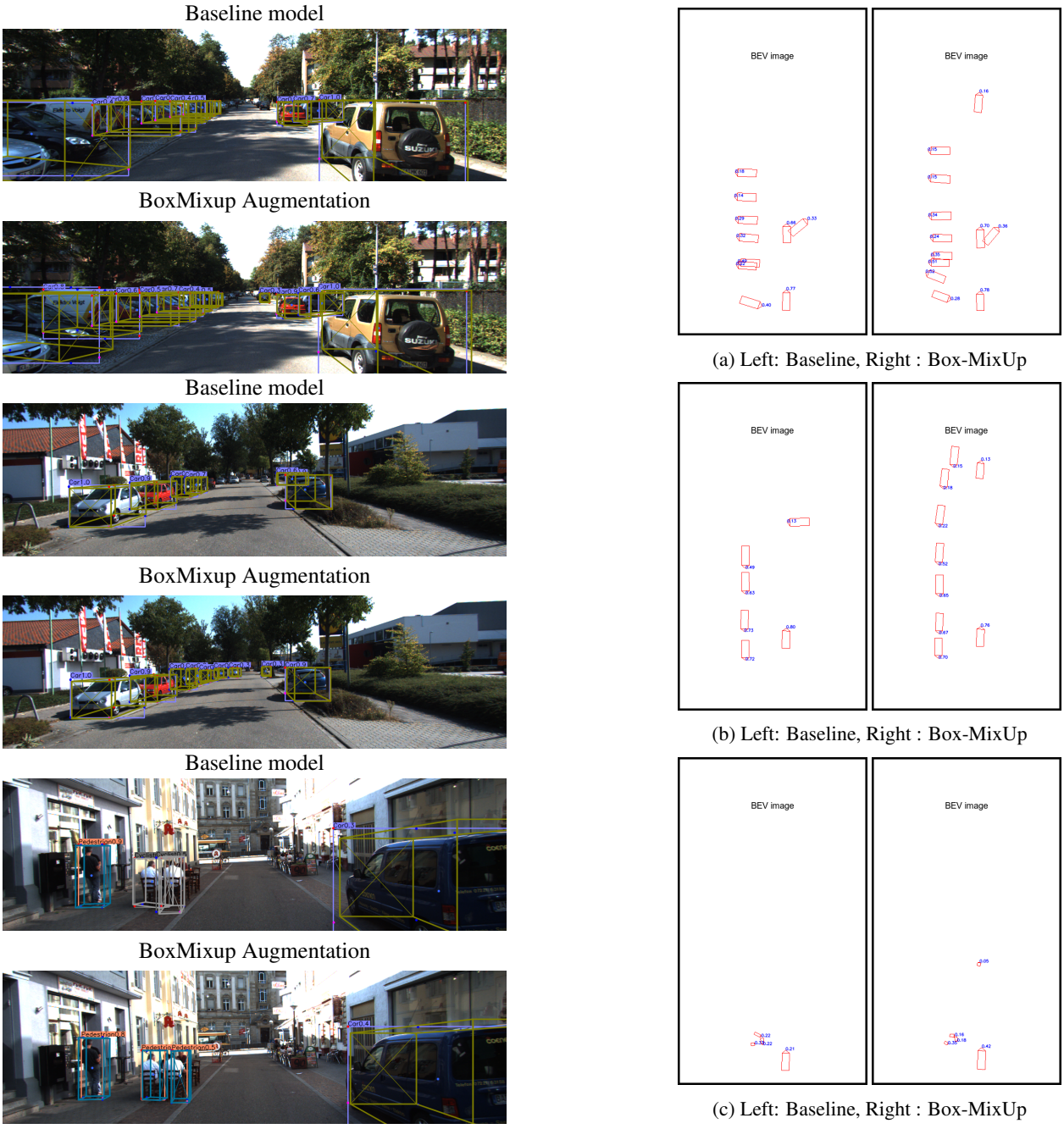
[16] Krishna Kumar Singh, Fanyi Xiao, and Yong Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3548–3556, 2016. 2

[17] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 3

[18] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016. 2

[19] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *ArXiv*, abs/1904.07850, 2019. 1

[20] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1



(a) Left: Baseline, Right : Box-MixUp

(b) Left: Baseline, Right : Box-MixUp

(c) Left: Baseline, Right : Box-MixUp

Figure 4: Illustration of Box-Mixup data augmentation in various scenarios. Each time contains the (baseline, Box-Mixup) prediction pair on the left panel, while the BEV representations (baseline, data augmented) pair on the right panel.