Journal of Multimedia Information System VOL. 9, NO. 3, September 2022 (pp. 183-190): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2022.9.3.183

A Survey for 3D Object Detection Algorithms from Images

Han-Lim Lee[1], Ye-ji Kim[1], Byung-Gyu Kim[1*]

Abstract

Image-based 3D object detection is one of the important and difficult problems in autonomous driving and robotics, and aims to find and represent the location, dimension and orientation of the object of interest. It generates three dimensional (3D) bounding boxes with only 2D images obtained from cameras, so there is no need for devices that provide accurate depth information such as LiDAR or Radar. Image-based methods can be divided into three main categories: monocular, stereo, and multi-view 3D object detection. In this paper, we investigate the recent state-of-the-art models of the above three categories. In the multi-view 3D object detection, which appeared together with the release of the new benchmark datasets, NuScenes and Waymo, we discuss the differences from the existing monocular and stereo methods. Also, we analyze their performance and discuss the advantages and disadvantages of them. Finally, we conclude the remaining challenges and a future direction in this field.

Key Words: 3D Object Detection, Autonomous Driving, Monocular Image Processing, Multi-View Image Processing, Stereo Image Processing.

I. INTRODUCTION

Three-dimensional (3D) object detection is considered important in computer vision applications that are deeply related to the real world, such as augmented reality, autonomous driving, and robotics. Most 3D object detection methods use RGB image as input with sensor devices that provide depth information such as LiDAR and Radar. Although LiDAR-based research [1-4] has developed a lot, interest in camera-based [5-11] has recently increased. This is because LiDAR is too expensive, and information on far away objects is not available. Also, the LiDAR point cloud is also very sparse, so very efficient transformation algorithm is required. Camera-based 3D object detection is much more advantageous in aspect of price and has rich information about far away objects. However, the disadvantage is that the accuracy is lower than the methods using sensor devices because the depth information must be estimated only based on the images.

Survey work for various approaches of 3D object detection has been published several times before [12-13]. However, existing survey works [12-13] have mainly explained the difference between the modality methods, such as camera based, point cloud, and sensor fusion. In this paper, we will make survey of the latest models for 3D perception and classification with only camera image. In particular, an analysis of the multi-view-based method that has recently getting attention is also included. There are three approaches for image-based object detection, which is depending on the number of cameras used.

The first is monocular-based 3D object detection [5-6], which is an approach of detection using only one image from a camera as an input. To predict 3D information of the objects of interest, the depth information must be obtained well. Since only one camera is used, there is a problem that it is difficult to estimate information on depth. To solve this problem, some methods have recently been introduced to predict 3D information by utilizing geometry prior. Nevertheless, monocular-based methods perform poorly compared to other methods due to lack of informative features.

The second approach is stereo-based 3D object detection [7-8], which utilizes disparity estimation for two images obtained by placing two cameras on the left and right sides. Since more accurate information on depth can be obtained by disparity estimation, comparable performance can be achieved even when compared to the LiDAR-based methods.

The last approach to introduce is multi-view 3D object detection [9-11]. The multi-view method utilizes multiple cameras in the ego car to make the field of view surrounding the car as input. In the existing monocular and stereo methods, 3D object detection and map segmentation were

Manuscript received September 9, 2022; Revised September 19, 2022; Accepted September 20, 2022. (ID No. JMIS-22M-09-032)
Corresponding Author (*): Byung-Gyu Kim, +82-2-2077-7293, bg.kim@sookmyung.ac.kr
[1]Department of IT Engineering, Sookmyung Women's University, Seoul, Korea, hl.lee@ivpl.sookmyung.ac.kr, yj.kim@ivpl.sookmyung.ac.kr, bg.kim@sookmyung.ac.kr

considered as separate tasks. However, a technique using multi-camera images has the advantage of being able to generate a BEV feature map centered on ego cars.

This work focuses on reviewing the state-of-the-art approaches for monocular, stereo, and multi-view 3D object detection mentioned above. We summarize the challenges and discuss the future research.

II. IMAGE-BASED 3D OBJECT DETECTION

2.1. Monocular-Based 3D Object Detection

Monocular 3D object detection is a task that estimates 3D information such as location, direction, and size of an object of interest using a single image as an input. Only by augmenting the 2D image feature or designing an efficient algorithm, the available feature can be refined. To solve the problem of severely lacking depth information compared to LiDAR-based methods [1-4, 12-13] or other camera-based methods [5-11], geometry prior is recently being used together.

Geometry Uncertainty Projection Network for Monocular 3D Object Detection (GUPNet) [5] proposed a GUP module that represent the inferences for depth as distributions using the geometry information. Since it uses depth as a continuous value rather than a discrete value, more accurate depth estimation is possible. As you can see in Fig 1, the GUP module estimates the depth as the distribution form. Another key design of GUPNet is hierarchical task learning (HTL) algorithm. 2D/3D height estimation is a very important issue in 3D object detection, as it can lead to incorrect depth estimation results. HTL strategy is to train the next task after the current task is well-trained. It proposed to reduce the instability in height estimation, which occurs frequently in the early of training. As shown in Table 1, GUPNet achieved comparable performance compared to CaDDN [11], a network using a camera with LiDAR sensor.

Learning Auxiliary Monocular Contexts Help Monocular 3D Object Detection (MonoCon) [2] proposed a method

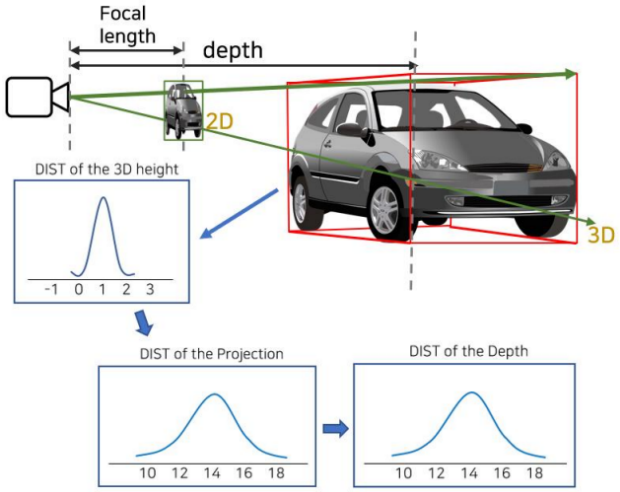


Fig. 1. An overview of GUP module [5].

using only image without any extra information such as lidar, CAD model or depth estimation module. The main idea of MonoCon is to utilize well-posed 2D contexts for auxiliary learning tasks to solve ill-posed problem. As shown in Fig. 2, it generates monocular contexts about geometric information. There are four types of auxiliary contexts: 1) The heatmaps of the projected 8 corner and center points of the 3D bounding boxes, 2) The offset vectors from the center point of 2D bounding box to the projected 8 corner points of 3D bounding box, 3) The size of 2D bounding box, 4)

Fig. 2. Examples of monocular context [6].

Table 1. Performance comparison on the car category in the Kitti official test set [16] – average precision of bird's eye view (AP_{BEV}).

Methods	Type	AP_{BEV} (IoU=0.7)			AP3d (IoU=0.7)		
		Easy	Moderate	Hard	Easy	Moderate	Hard
GUPNet [5]	Mono	0.303	0.212	0.182	0.201	0.142	0.118
MonoCon [6]	Mono	0.311	0.221	0.190	0.225	0.165	0.140
Stereo R-CNN [7]	Stereo	0.619	0.413	0.334	0.476	0.302	0.237
Disp R-CNN [8]	Stereo	0.738	0.523	0.436	0.585	0.379	0.319
CaDDN [11]	Mono & LiDAR	0.279	0.189	0.172	0.192	0.134	0.115
DSGN [12]	Stereo & LiDAR	0.829	0.651	0.566	0.735	0.522	0.451

The residual of a keypoint location. Even though it utilized simple contexts for an additional feature learning, MonoCon showed better performance than GUPNet [5].

2.2. Stereo-Based 3D Object Detection

Compared to the monocular-based detection, stereo-based can obtain richer depth information by conducting the disparity estimation using left and right images. Therefore, this method can reduce the ill-posed problem that has not been solved in the monocular method. Despite the low-priced of setting it up. It shows comparable performance compared to LiDAR-based approaches without using expensive sensor devices.

The first model to introduce is Stereo R-CNN based 3D Object Detection for Autonomous Driving (Stereo R-CNN) [7]. It is a network that simultaneously detects and associates objects from stereo images. The algorithm is simple: First, the backbone network extracts the 2D feature from left and right images. Second, the extracted features are input into the stereo region proposal network (RPN) to concatenate them. And then, they align the proposed Region of interest (RoI) to each left-right feature map.

Finally, the aligned features are utilized to estimate the 3D bounding boxes by predicting the key-points of 3D boxes and conducting stereo regression. This approach outperformed other state-of-the-art image-based methods over 30% average precision of bird's eye view and 3D boxes.

The Second is Stereo 3D Object Detection via Shape Prior Guided Instance Disparity Estimation (Disp R-CNN) [8]. Li et al. proposed a more advanced network that Stereo R-CNN [7]. The existing stereo-based 3D object detection conduct the disparity estimation of the full-frame level. However, this approach often fails to generate the accurate disparity for low textured objects like vehicle. Also, since the area of the object of interest is much smaller than the background, there are many unnecessary computations. To solve these problems, Disp R-CNN proposes object detection based on the disparity estimation of the instance level. The process of instance-level estimation is as follows: 1) specifying the object region in the feature map after RoIAlign, 2) estimating disparity of the instance-level using full-frame disparity and coordinates of the left border of left-right bounding boxes.

The instance-level disparity estimation is illustrated Fig. 3. Meanwhile, most driving datasets [16-18] do not provide ground-truth for instance disparity, so there is a problem that the disparity cannot be learned. To solve the problem, they needed to create pseudo ground-truth generation without LiDAR points. They proposed a process that uses CAD model to generate dense pseudo ground-truth. As a result, it is possible to learn the object shape prior. Disp R-CNN

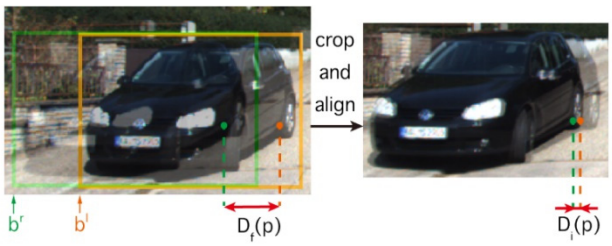


Fig. 3. Instance-level disparity estimation [8].

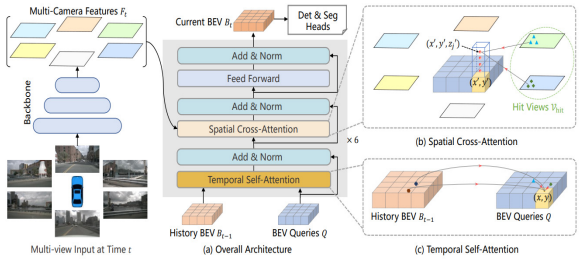


Fig. 4. Overall architecture of BEVFormer [9].

achieved not only state-of-the-art performance, but also faster inference time than other stereo-based models by reducing computation with novel disparity estimation.

As shown in the Table 1, Disp R-CNN performs worse than DSGN [14] which is a stereo 3D object detection method based on LiDAR. But it is better than Stereo R-CNN.

2.3. Multi-View 3D Object Detection

Many existing 3D Object detection used KITTI Datasets [13] as training datasets. The KITTI dataset is a collection of data with two RGB cameras, so only the Monocular and Stereo methods were possible. After that, multi-camera driving datasets such as NuScenes [17] and Waymo [18] datasets are appeared, allowing multi-view 3D object detection research.

In this task, bird's-eye-view (BEV) representation, which is a map centered on ego-car, can be used to intuitively visualize the location, size, and orientation of the object of interest. In this paper, we introduce transformer-based approaches that recently shown excellent performance. In 2D object detection, the initial transformer-based model was DETR [19], which used object query to perform detection on the output of the decoder. The transformer-based model does not require NMS processing, and this is the same in 3D detection. Note that in multi-view detection, when feeding images into a model, do not divide the image into patches like 2D detection. They used multi-camera images as input like patches.

BEVFormer [9] generates bird's eye view (BEV) Features that combines temporary and spatial information to perform 3D object detection and map segmentation tasks. The main components of the BEVFormer are BEV query,

spatial cross attention, and temporal self-attention.

First, BEV query is a grid shaped query with the same size as BEV plane and consists of learnable parameters. Therefore, the space of the real world can be represented by BEV query. BEV query is first used in the temporal self-attention step to query temporal information from the previous BEV features. Then, BEV query is used to find spatial information through spatial cross attention between multi-view features. These output features are used as input to the feed-forward network, and as a result, the BEV feature is updated. The updated feature is used as the input of the next encoder layer.

After doing the same work on the 6 encoder layers, the BEV features BEV_t at the timestamp t is completed. With that, they input to the detection head and segmentation head to predict the 3D bounding boxes and map segmentation. Finally, the BEV_t is used as input for temporal self-attention at the next timestamp t-1. The temporal self-attention proposed by BEVFormer shows excellent performance of the velocity estimation than other existing camera-based method.

In Table 2, the mean average velocity error (mAVE) decreased by more than 5% compared to other camera-based models. Also, the recent camera-only methods tend to add temporal self-attention to improve the performance.

PETR [10] proposed a new position embedding for multi-view 3D object detection. They made a 3D coordinates generator module to represent 2D features like 3D features. It transforms the camera frustum space to the 3D world space. Meshgrid coordinates are shared by multi-view features, so the coordinates on 3D world space can be calculated by reversing the 3D projection expression with different camera parameters.

The pipeline of PETR is as follows. First, 2D feature is extracted from each view image using a backbone network such as ResNet [20] or VovNet-99 [21]. The 3D coordinate and 2D feature are then used together as inputs to the 3D position encoder to generate the 3D position-aware feature. The decoder uses these results and object query as input to predict the class and 3D bounding boxes of objects in each scene. As you can see in Fig. 5., 3D position embedding shows that the information related to position can be found well in surrounding images.

In general, 3D object detection is designed based on a Cartesian coordinate system using a perpendicular axis. PolarFormer [11] applied the polar coordinate system, noting that the real world seen in each camera from a ego-car perspective is a wedge shape. It is illustrated in Fig. 6. with the BEV map on cartesian coordinate and polar coordinate. When polar coordinate is applied, it can better represent nearby objects as in real world. Although it is difficult to implement grid shape as non-rectangular to apply polar coordinate to deep learning networks, this paper implements it in a novel way.

The pipeline of the Polar Former is as follows. First, the image of each view is fed into the backbone and FPN to extract the multi-scale 2D image feature. These feature maps are used as input to the cross-plane encoder, which converts the column of each feature into a polar ray in sequence-to-sequence format. In the polar alignment module, the generated polar relays are combined to create a polar BEV map. Then, the multi-scale BEV map is fed into the Polar BEV encoder to learn richer information across all feature scales and generate more refined BEV features. Finally, polar head utilizes multi-scale polar BEV features to predict the 3D bounding box on the Polar coordinate system and classifies the category of the object.

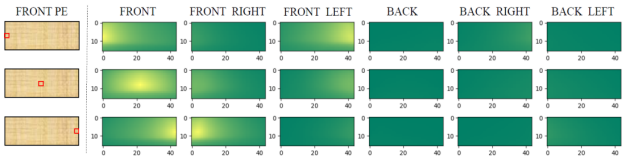


Fig. 5. The visualization of 3D position embedding similarity.

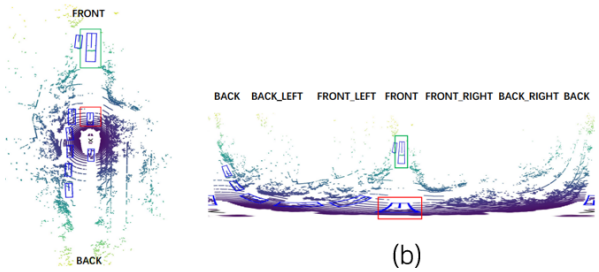


Fig. 6. The object detection results (a) based on Cartesian coordinate system and (b) based polar coordinate system.

Table 2. Performance comparison on nuScenes test set.

Methods	Modality	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFormer [6]	Multi-view	0.569	0.481	0.582	0.256	0.375	0.378	0.125
PETR [7]	Multi-view	0.504	0.441	0.593	0.249	0.383	0.808	0.132
Polar Former [8]	Multi-view	0.572	0.493	0.556	0.256	0.364	0.440	0.127
SSN [4]	LiDAR	0.569	0.463	-	-	-	-	-
CenterPoint [3]	LiDAR	0.655	0.580	-	-	-	-	-

In Multi-view 3D object detection, features are extracted using a combination of backbone and FPN [22] to find objects of various sizes. In addition, they usually use ResNet-101 [20] and VovNet2-99 [21] and Swin- Transformer [23] as a backbone network. So far, it is difficult to detect real-time 3D object due to the large latency while extracting image features from the Backbone network. Also, it is still difficult to detect objects such as pedestrians and cyclists smaller than cars. As small object detection well in autonomous driving is important to ensure safety, research on this aspect is also required.

III. DATASETS

In this section, we will analyze two frequently used datasets in 3D object detection. KITTI dataset [14] released in 2012, and most monocular and stereo studies still widely use it. KITTI dataset uses only two RGB cameras, it can be used for the monocular method and the stereo method. Since multi-view images are not provided on the KITTI dataset, other new datasets were needed for research related to multi-view. Meanwhile, they provide point cloud information surrounding the ego-car by installing a laser scanner on the car. Although it is an old dataset, various tasks can be studied with KITTI data using various sensor modalities. In addition to object detection, they have opened several benchmarks such as flow, depth, odometry, and line detection. NuScenes dataset [17] was inspired by the KITTI dataset. To collect this dataset, they installed 6 cameras looking in various directions on the ego-car, one LiDAR, and five radars, thus providing more meta information than KITTI [14]. KITTI provides only 22 scenes, while it provides 1K scenes. In addition, KITTI has 15 k annotated frames, labeling only for 8 classes, while it is about 2.7 times more than that, and the number of classes is 23. Therefore, the nuScenes dataset is a much larger dataset in many aspects.

3.1. KITTI Dataset

3.1.1. Datasets

KITTI dataset [17] consists of 7481 training images and 7518 test images. The test dataset does not have ground-truth, so the dataset for validation is part of the training dataset. The answer label consists of three classes: Car, Cylist, and Pedestrian.

3.1.2. Evaluation Metrics

KITTI benchmark evaluates performance with average precision (AP). AP_{3D} represents the AP of 3D bounding boxes, and AP_{BEV} represents the AP of bird's eye view. Difficulties are defined in three levels: easy, moderate, and hard. The criteria for occlusion and truncation are different depending on each level.

3.2. NuScenes Dataset

3.2.1. Datasets

The nuScenes [17] dataset contains 1,000 driving scenes collected by Boston and Singapore. Each scene is about 20 seconds long, and there are 3D bounding box annotations at 2 Hz for 23 object classes. It is a driving scene taken with 6 cameras, so there are about 1.4 M camera images.

3.2.2. Evaluation Metrics

The nuScenes benchmark use 7 defined metrics. The first metric is the average precision (AP) metric considering the 2D center distance on the ground plane. And there are five true positive (TP) metrics that measure translation, scale, orientation, velocity, and attribute errors. The meanings of metric in Table 2. are as follows: Average translation error (ATE) is a metric that calculates the Euclidean center distance with a mean average translation error. Average scale Error (ASE) is a metric, which aligns the center and orientation and then calculates the 1-intersection of union (IOU) between the 3D bounding. Average orientation error (AOE) calculates the yaw angle difference between the predicted bounding box and the ground-truth bounding box. Average velocity error (AVE) calculates the difference in absolute velocity. average attribute error (AAE) is a mean average attribute error, which means the error rate of object classification. These TP metrics are calculated separately for each class, and mATE, mASE, mAOE, mAVE, and mAAE, which are calculated on average, are used in Table 2. They also use a nuScenes detection score (NDS) by computing a weighted sum of AP and TP metrics.

IV. PERFORMANCE COMPARISON

We will compare the performance of the monocular and stereo detection models in the KITTI official benchmark [16]. As shown in Table 1, the monocular 3D object detection [5-6, 11] achieved similar performance with or without LiDAR. Stereo 3D object detection using only camera [7-8] achieved slightly lower performance than the LiDAR method, DSGN [9]. However, they outperformed CaDDN [5] using a single camera and LiDAR, which shows that the using stereo cameras can learning richer semantics.

In Table 2, the performance of multi-view methods using nuScenes [17] dataset was compared. Among the models that did not use Lidar [6-8], PETR [7] achieved the lowest performance because it did not utilize temporal information

for learning. As shown in Table 2, the temporal self-attention which has been first proposed by BEVFormer [6] reduced the error of velocity. Compared to SSN [4] and CenterPoint [3], which are models that utilize Point cloud together, these models achieved comparable performance.

V. CONCLUSION AND DICUSSION

In this paper, we have reviewed the monocular, stereo and multi-view 3D object detection methods. The camera-based methods still have many problems, such as incorrect 3D inference results or poor detection of small objects.

Monocular detection has a problem that geometric prior that can be obtained from one image is very insufficient. To solve these problems, modern monocular-based papers use the strategy to find contexts that can be learned in an image. Stereo-based 3D object detection often utilized disparity estimation to estimate the 3D information such as location, dimension, and orientation. Utilizing the disparity information to train the model enables more accurate detection, which means that parameters related to the camera are also important for 3D object detection. Finally, multi-view detection using the surrounding images has the advantage of being able to utilize more information such as camera parameters than monocular or stereo.

In the field of autonomous driving and robotics, most objects of interest usually move quickly. But there is a problem that all three approaches use the backbone with the large scale. This means that it is still difficult to apply to real-time detection. Therefore, efficient feature extraction will be important in future studies. In addition, there is an important problem that detection of small objects is still difficult. Enabling precise detection of small objects such as pedestrians or obstacles will also be an important issue for future research. If these problems are completely solved, image-based 3D object detection will be successful in the future, even without LiDAR system.

ACKNOWLEDGMENT

This research was supported by DNA+Drone Technology Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. NRF-2020M3C1C2A01080819).

REFERENCES

[1] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, and B. Gong, et al.,"Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601-9610.

[2] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, and W. Li, et al., "Cylindrical and asymmetrical 3D convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9939-9948.

[3] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11784-11793.

[4] X. Zhu, Y. Ma, T. Wang, Y. Xu, J. Shi, and D. Lin, "Ssn: Shape signature networks for multi-class object detection from point clouds," in *European Conference on Computer Vision*, Springer, 2020, pp. 581-597.

[5] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, and Q. Chu, "Geometry uncertainty projection network for monocular 3D object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2020, pp. 3111-3121.

[6] X. Liu, N. Xue, and T. Wu, "Learning auxiliary monocular contexts helps monocular 3D object detection," arXiv preprint arXiv:2112.04628, 2021, unpublished.

[7] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3D object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644-7652.

[8] J. Sun, L. Chen, Y. Xie, S. Zhang, Q. Jiang, X. Zhou, and H. Bao, "Disp r-cnn: Stereo 3D object detection via shape prior guided instance disparity estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10548-10557.

[9] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, and T. Lu, et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," arXiv preprint arXiv:2203.17270, 2022, unpublished.

[10] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3D object detection," arXiv preprint arXiv:2203.05625, 2022, unpublished.

[11] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, and W. Hu et al., "PolarFormer: Multi-camera 3D object detection with polar transformers," arXiv preprint arXiv:2206.15398, 2022, unpublished.

[12] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782-3795, 2019.

[13] Z. Li, Y. Du, M. Zhu, S. Zhou, and L. Zhang, "A survey of 3D object detection algorithms for intelligent vehicles development," *Artificial Life and Robotics*, pp. 1-8, 2021.

[14] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555-8564.

[15] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12536-12545.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE. 2012, pp. 3354-3361.

[17] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, and Q. Xu, et al., "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11621-11631.

[18] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, and P. Tsui, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446-2454.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko,"End-to-end object detection with transformers," in *Proceedings of European Conference on Computer Vision,* Springer, 2020, pp. 213-229.

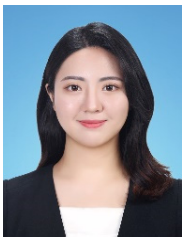
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[21] Y. Lee, J. W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and gpu-computation efficient backbone network for real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

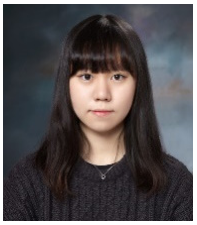
[22] T. Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117-2125.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, and Z. Zhang, 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.

AUTHORS



Han-Lim Lee received her B.S. in the Department of Statistics and Computer Science from Sookmyung Women's University, Korea, in 2022. In 2022, she joined the Department of IT Engineering for pursuing his M.S. degree at Sookmyung Women's University. Her research interests include deep learning algorithms for 3D object detection.



Ye-Ji Kim received her B.S. in the Department of Chemistry and IT Engineering from Sookmyung Women's University, Korea, in 2022. In 2022, she joined the Department of IT Engineering for pursuing his M.S. degree at Sookmyung Women's University.
Her research interests include deep learning algorithms for depth information estimation.

Byung-Gyu Kim has received his BS degree from Pusan National University, Korea, in 1996 and an MS degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a PhD degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST). In March 2004, he joined in the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea where he was a senior researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award in 2007.

From February 2009 to February 2016, he was associate professor in the Division of Computer Science and Engineering at Sun-Moon University, Korea. In March 2016, he joined the Department of Information Technology (IT) Engineering at Sookmyung Women's University, Korea where he is currently an associate professor.

In 2007, he served as an editorial board member of the International Journal of Soft Computing, Recent Patents on Signal Processing, Research Journal of Information Technology, Journal of Convergence Information Technology, and Journal of Engineering and Applied Sciences. Also, he is serving as an associate editor of Circuits, Systems and Signal Processing (Springer), The Journal of Supercomputing (Springer), The Journal of Real-Time Image Processing (Springer), Heliyon Journal (Elsevier), and International Journal of Image Processing and Visual Communication (IJIPVC). From 2018, he is serving as the Editor-in-Chief (EiC) of the Journal of Multimedia Information System. He also served as Organizing Committee of CSIP 2011 and Program Committee Members of many international conferences. He has received the Special Merit Award for Outstanding Paper from the IEEE Con-

sumer Electronics Society, at IEEE ICCE 2012, Certification Appreciation Award from the SPIE Optical Engineering in 2013, and the Best Academic Award from the CIS in 2014. He has been honored as an IEEE Senior member in 2015.

He has published over 260 international journal and conference papers, patents in his field. His research interests include software-based image and video object segmentation for the content-based image coding, video coding techniques, 3D video signal processing, wireless multimedia sensor network, embedded multimedia communication, and intelligent information system for image signal processing. He is a senior member of IEEE and a professional member of ACM, and IEICE.