Article

AutoGranularity: Enhancing BEV Accuracy via Adaptive Processing from Scene Complexity

Yong Chan Lee [1] **and Eun Kyu Lee** [2]

[1, 2] Incheon National Univ; {yong26boy, eklee}@inu.ac.kr

Abstract

BEV MAP is effective in identifying the surrounding environment of autonomous vehicles. In particular, camera-based BEV MAP generation models are attracting attention because they can generate BEV MAPs while maintaining real-time performance at a lower cost than LIDAR-based models. However, since they rely on 2D images, there is an accuracy issue caused by object occlusion and the lack of depth information.

To solve this problem, this study proposes AutoGranularity. It is a method that increases accuracy by preventing feature contamination through separating the batch data input to the image backbone. The resulting degradation in real-time performance is mitigated through an adaptive model structure based on Scene Complexity. Here, Scene Complexity is measured by combining the number and variance of objects detected in camera images and Shannon entropy.

AutoGranularity, which adopts an adaptive structure in SinBEVT and applies the Per-sample forward pass method only when the surroundings of autonomous vehicles are complex, shows performance between SinBEVT and the Per-sample forward pass method in terms of IoU. In addition, the real-time performance degradation caused by the Per-sample forward pass method is mitigated by reducing the inference time by more than 0.1 seconds.

The source code of AutoGranularity is available at: https://github.com/defenxe/CoBEVT/tree/main/nuscenes

Keywords: Autonomous Driving, BEV map understanding, Adaptive Model

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

For smooth autonomous driving, it is essential to understand surrounding obstacles and the road environment. To achieve this goal, it is common to generate a BEV MAP using LIDAR and camera sensor data. However, LIDAR has many limitations because of its high cost and long computation time. To address this problem, BEV MAP generation methods using cameras have been actively studied. However, there is a limitation in improving accuracy due to object occlusion and the lack of depth information.

https://doi.org/10.3390/xxxxx

In this study, a new method is proposed to compensate for the limitations of camera-based BEV MAP generation models. The data separation method of the image backbone improves accuracy by leveraging the data processing characteristics of internal modules in the image backbone. In the case of EfficientNet-b4 used in this study, when batch data are passed through the image backbone without separation, statistical contamination and contextual contamination between samples occur during the normalization and SE block processes. The data separation method of the image backbone improves model accuracy by preventing such feature contamination.

However, separating batch data and passing them individually through the image backbone can cause long computation times due to per-sample processing. Therefore, in this study, an adaptive model method based on Scene Complexity is proposed to maintain the advantage of short computation time in camera-based BEV MAP generation models. Based on the number of objects detected in the camera and Shannon entropy, this method applies the data separation strategy of the image backbone only when the autonomous vehicle is driving in a complex environment with many objects, while enabling short computation times in simpler environments.

The contributions of this paper are as follows:

1. By leveraging the data separation method of the image backbone, we alleviate the accuracy bottleneck, which is the most critical limitation of camera-based BEV MAP generation models.

2. We propose the concept of Scene Complexity, which combines the number of objects detected in the camera and Shannon entropy, and apply the proposed model structure only in complex environments. Through this adaptive structure, short computation time can be maintained.

2. Related Works

2.1. Transformer based Semantic Segmentation

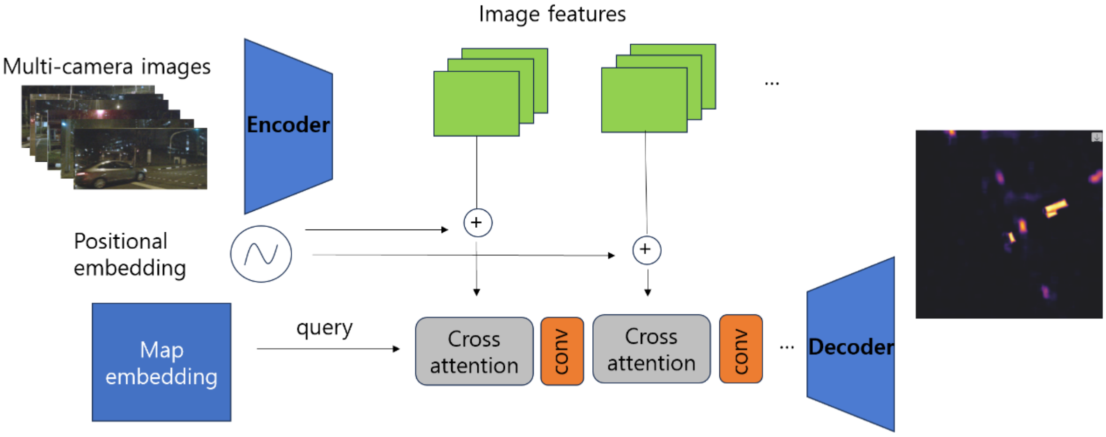


Figure 1. The framework of Cross View Transformer (adapted from [1])

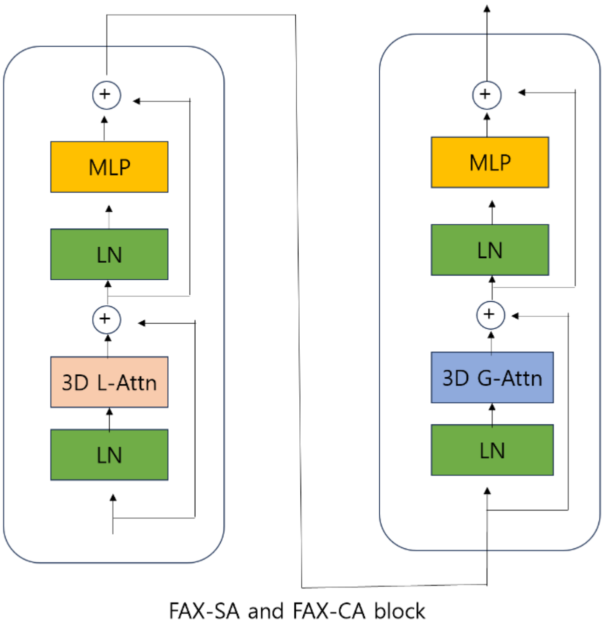


Figure 2. Fax Self Attention and Fax Cross Attention Structure (adapted from [2])

Several studies have been conducted on semantic segmentation using transformer models. Cross View Transformer [1] performs BEV semantic segmentation by applying positional embeddings to image features, as shown in Figure 1, and then passing them through cross-attention and convolution structures together with map embeddings. SinBEVT [2] of CoBEVT, which is used as the baseline in this study, improves performance by replacing the cross-attention mechanism in CVT with the fax cross-attention shown in Figure 2. Fax attention reflects both fine-grained details and global context through 3D local attention and 3D global attention.

There is also an approach that adds additional modules outside the Transformer. As an example, BEVFormer v2 [3] uses perspective prediction to directly involve camera images in training. This addresses the issue of insufficient 3D learning information. In addition, temporal BEV is used to incorporate past BEV representations into the prediction of the current BEV.

In addition, methods that perform 3D semantic segmentation instead of BEV semantic segmentation have been proposed. VoxFormer [4] places pixels in camera images into a 3D space using depth estimation. Through deformable cross-attention, patterns from image data are applied to the corresponding pixels. After that, self-attention is used to infer missing voxel data. On the other hand, Fiery [5] generates a 3D map through geometric transformation and then projects it into 2D. MaxViT [6] is a model that alternately and repeatedly applies local (window) and global (grid) attention, achieving strong global representations while maintaining linear computational complexity.

2.2. V2V Perception

Methods that use V2V Perception to improve the performance of MAP generation models have been studied. V2V Perception is a method that improves accuracy by sharing information about blind spots among autonomous vehicles. SCOPE [7] learns a unified model by aggregating data from multiple autonomous vehicles. Past data are used to capture temporal context through a Pyramid LSTM. In addition, images are transformed into multiple scales and multiplied by confidence maps so that only highly reliable regions are used. Finally, data near a reference location are utilized to adjust localization errors between vehicles.

CoCa3D [8] converts image features into distance-wise probability distributions of object locations through depth estimation. It improves these probability tables by incorporating depth estimation information and 3D features shared from other agents, and then fuses them with its own 3D features. The resulting 3D map is converted into a BEV map to improve efficiency.

HM-ViT [9] is a model designed to perform heterogeneous collaborative perception in multi-agent scenarios. VALUE information is determined by edge type, while QUERY and KEY information are determined by node type. The KEY and VALUE vectors are generated by the sender, whereas the QUERY vector is generated from the receiver's features. These QUERY, KEY, and VALUE vectors are fed into a Transformer to enable heterogeneous V2V Perception. In addition, the window size is adjusted according to inter-token distance, allowing the model to focus on fine-grained details when tokens are close and to share global context when they are far apart.

CoAlign [10] is a model that achieves robustness to object localization errors without requiring labeled data. It corrects errors by adjusting object positions based on the center coordinates of overlapping regions between object bounding boxes. The reliability of each bounding box is computed using variance, and boxes with low reliability are excluded. Intermediate features with adjusted positions are then fused.

V2X-ViT [11] is a model for cooperative perception among heterogeneous agents. Different types of agents, such as infrastructure and autonomous vehicles, perform self-attention to reduce heterogeneity. FedBEVT [12] replaces only the cooperative perception component of CoBEVT [2] with Federated Learning. In addition, when agents have an insufficient number of cameras, masking is applied to camera-missing regions to improve performance.

Robosac [13] is a method for identifying attackers in V2V Perception. The ego robot treats data from robots that produce inconsistent results as attackers and excludes them. Unlike existing methods, this approach can handle previously unseen attacks. MADE [14] is a model proposed as an improvement over Robosac [13]. If the match loss statistics and collaborative reconstruction loss statistics fall below a predefined threshold, an attacker is assumed to be present.

Where2comm [15] is a model designed to improve the communication efficiency of V2V Perception. Only regions with high scores in the feature map are shared with other agents, while regions with low scores are requested from other agents. In addition, a confidence map indicating whether the data are sufficiently reliable is applied as a weighting factor to the feature map. Unlike Where2comm [15], How2comm [16] applies the confidence map as a weight to the feature map from the beginning and propagates it through all subsequent operations. Furthermore, since different regions of each agent's map are likely to contain important information, Exclusive Spatial Attention is used to distinguish complementary information. In addition, missed objects are recovered from other agents through Temporal Cross Attention, or maps altered by time delays are inferred using a Flow Generator.

3. Problem Definition

The LiDAR-based MAP generation method shows higher accuracy than camera-based methods. LiDAR constructs maps in three dimensions using laser reflection time, which enables accurate depth measurement. On the other hand, camera-based BEV MAP generation methods cannot accurately measure the distance to objects because they generate BEV maps from two-dimensional images. In addition, problems such as object occlusion, where objects behind are blocked by objects in front, also exist. To address these issues, it is necessary to improve the accuracy of the model. This study focuses on the image backbone structure of the previous model, SinBEVT.

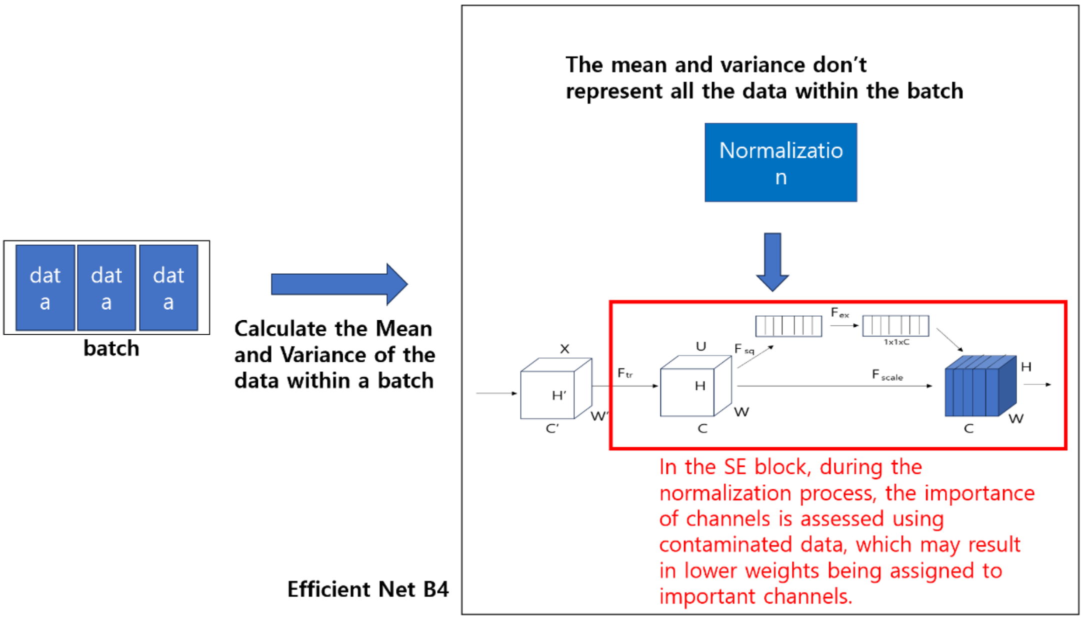

The mean and variance don't represent all the data within the batch
Normalizatio n
dat a
dat a
dat a
batch
Calculate the Mean and Variance of the data within a batch
X
U
F sq
F ex
1x1xC
H'
W'
C'
F tr
H
W
C
F scale
H
C
W
In the SE block, during the normalization process, the importance of channels is assessed using contaminated data, which may result in lower weights being assigned to important channels.
Efficient Net B4

Figure 3. Accuracy degradation caused by batch-wise processing (adapted from [17])

The EfficientNet-b4 model used as the image backbone consists of multiple modules. Among them, the batch normalization and SE block modules may lead to a decrease in accuracy when data are input in batch units, as shown in Figure 3. Batch normalization computes the mean and variance of data based on the entire batch. When features within the same batch differ significantly from each other, the resulting statistics may fail to represent all samples. Such statistical contamination in batch normalization can lead to a degradation in model accuracy. In addition, the SE block module also encounters issues with batch-wise data processing. The SE block module assigns weights by evaluating the importance of channels. When features contaminated by batch normalization are fed into the SE block, the weights of important channels may be underestimated. Therefore, separating batch data before inputting them into the image backbone is required to improve the accuracy of camera-based BEV MAP generation models.

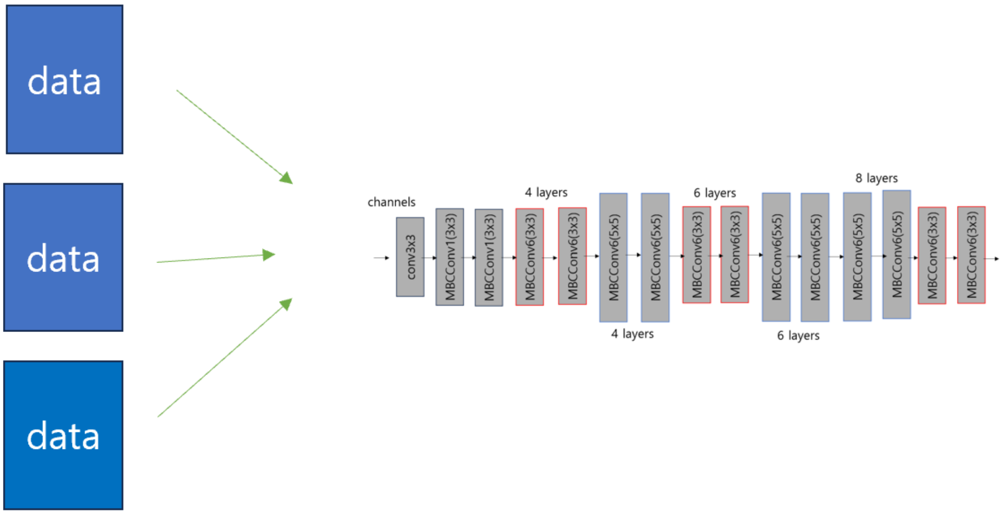


Figure 4. Separation of batch data input to the image backbone (adapted from [18])

However, as shown in Figure 4, separating batch data and passing them through the image backbone significantly increases computational cost. This undermines the real-time performance, which is a key advantage of camera-based methods. This study proposes a method that improves the accuracy of camera-based BEV MAP generation models while preserving real-time performance as much as possible.

4. Proposed Methodology

4.1. Overall Framework

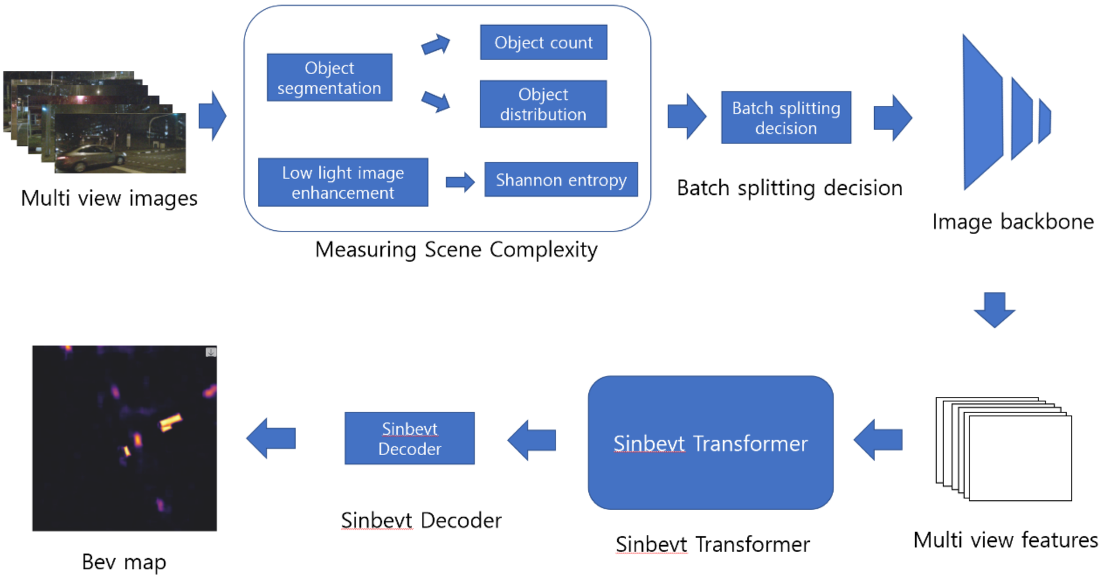


Figure 5. The overall framework of AutoGranularity

This study adopts a method of measuring scene complexity before inputting batch data into the image backbone. When the scene complexity, calculated by combining object count, Shannon entropy, and object distribution, exceeds a predefined threshold, the model structure is modified to separate the batch data before they are input into the image backbone. The overall structure is shown in Figure 5.

4.2. Scene Complexity Definition

Two elements are used in the definition of Scene Complexity: object count and Shannon entropy.

Object count refers to the total number of objects captured in images acquired from six cameras mounted on an autonomous vehicle. The types of objects are classified into a total of 12 categories, as shown in Table 1.

Object types	**Types**		
	dynamic	***static***	***divider***
Object	Car, truck, bus, trailer, constriction, pedestrian, motorcycle, bicycle	Lane, road segment	Road divider, lane divider

Table 1. Table of Object Types

Shannon entropy is introduced to measure the complexity of the surrounding environment of autonomous vehicles. It depends on the diversity of pixel values, and its value increases when there are many boundaries and when patterns are complex. Therefore, hazardous environments such as complex urban areas, intersections, and roads with numerous pedestrians can be measured using Shannon entropy.

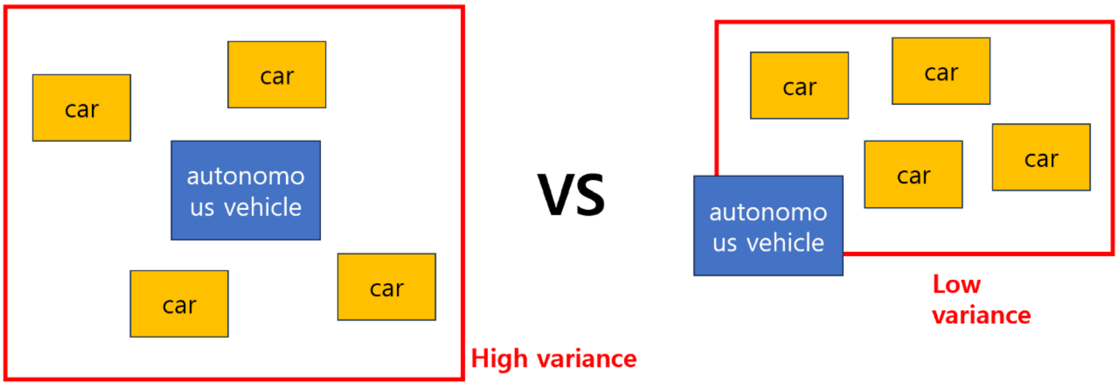


Figure 6. The figure of Object Distribution

Object distribution is used to measure risk according to the spatial distribution of objects around the autonomous vehicle, as shown in Figure 6. Even when the same number of objects are located near the autonomous vehicle, the complexity differs between situations where the vehicle is surrounded by objects and situations where objects are concentrated in only one direction. Object distribution evaluates the dispersion of surrounding objects by first obtaining the coordinates of objects detected by the cameras and then computing their variance. In this process, the autonomous vehicle itself is excluded from the variance calculation.

Scene Complexity is computed by combining object count, Shannon entropy, and object distribution. To ensure equal contributions from each component, scaling is applied before summation. Therefore, when the Scene Complexity defined in this manner exceeds a threshold, structural adaptation of the adaptive model is triggered.

The reason for adopting object count and Shannon entropy in the definition of Scene Complexity lies in their complementary characteristics. When considering only object count, roads with simple structures, such as highways, may be misclassified as complex environments solely due to a large number of detected objects. On the other hand, when Shannon entropy alone is used to define Scene Complexity, only small differences may appear between entropy values across different datasets due to the low-light image enhancement described later, potentially leading to errors. Therefore, this study defines Scene Complexity by combining these two criteria.

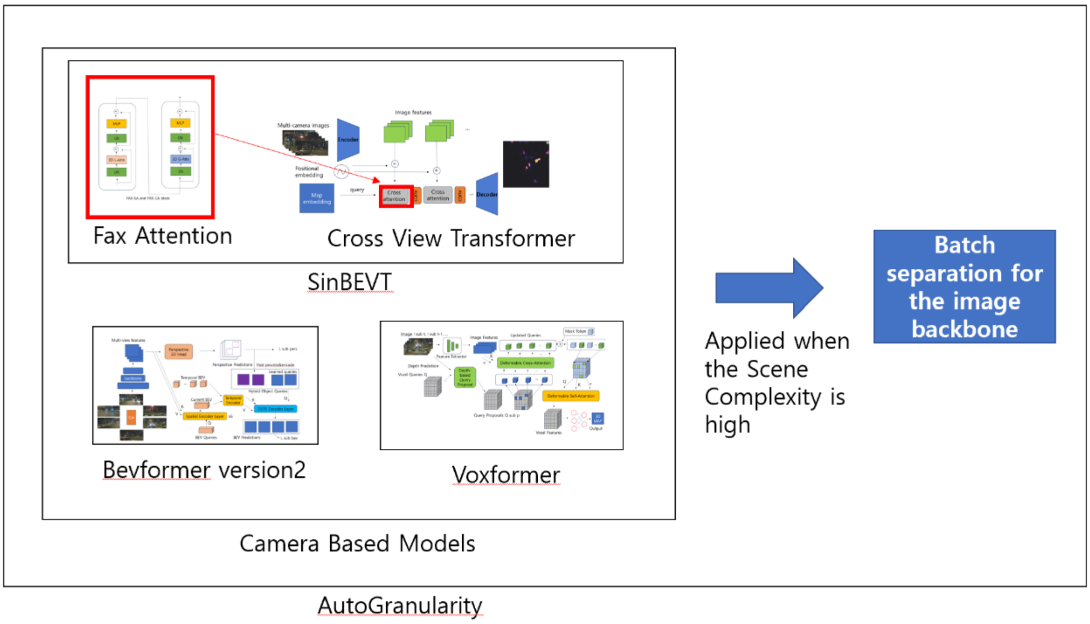
4.3. Adaptive Structural Design

If the adaptive model determines that the scene complexity of a camera image is high, the structure is deformed in a direction that increases accuracy. At this time, additional modules for object count and Shannon entropy are used to improve performance and prevent possible errors.

- **Object segmentation module:** Objects captured in camera images have different levels of risk depending on their type. To reflect this, object segmentation is performed to classify object types according to the criteria in Table 1, and the object count values are differentially adjusted based on the object type. Pedestrians, motorcycles, and bicycles are vulnerable to severe impact in the event of an accident because humans are directly exposed, while buses and trucks can be dangerous due to their large size and blind spots. Therefore, in this study, pedestrians, buses, motorcycles, bicycles, and trucks are classified as high-risk objects. When a high-risk object is detected, the object count is increased by 2. For other objects, the object count is increased by 1 upon detection.

- **Low-light image enhancement module:** Shannon entropy, which is used to measure the complexity of the environment surrounding an autonomous vehicle, has time-dependent limitations. At night, the diversity of pixel values decreases due to insufficient illumination, which increases the probability of errors. To prevent this, color correction is performed through low-light image enhancement. Gamma correction and histogram equalization are applied to ensure that Shannon entropy operates normally even at night.

4.4.



Difference from Related Work

Figure 7. Relationship between AutoGranularity and Related Work (adapted from [1], [2], [3], [4])

If the adaptive model determines that the scene complexity of a camera image is high, the structure is deformed in a direction that increases accuracy. At this time, additional modules for object count and Shannon entropy are used to improve performance and prevent possible errors. AutoGranularity is a method designed by transforming a camera-based model into an adaptive model. As shown in Figure 7, when the scene complexity is low, the BEV map is generated using the camera-based model. On the other hand, in environments where the scene complexity exceeds a threshold, AutoGranularity separates the batch data to be input into the camera-based model and passes each sample individually through the image backbone. Through this process, it addresses the accuracy degradation problem, which is considered a fundamental bottleneck of camera-based models.

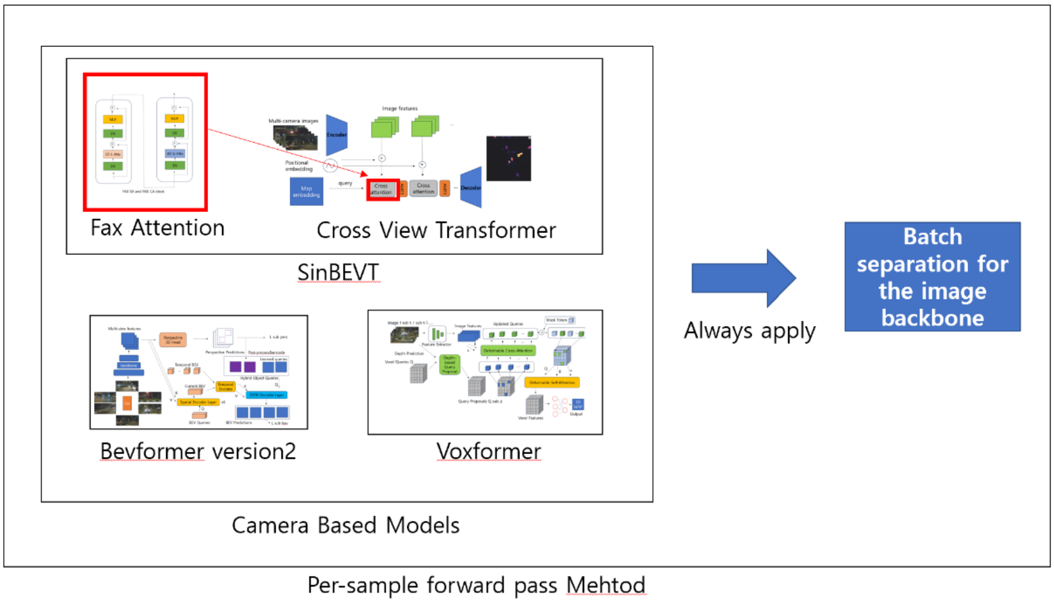


Figure 8. Relationship between Per-sample forward pass Method and Related Work (adapted from [1], [2], [3], [4])

The per-sample forward pass method is an approach in which the adaptive structure of AutoGranularity is not applied. Not only when the scene complexity is high, but also at all times, the batch data are separated and fed individually into the image backbone of camera-based models. Since the batch is always separated during input, statistical contamination and contextual contamination do not occur, resulting in higher accuracy compared to AutoGranularity. However, it should be noted that the inference time becomes longer, which may prevent real-time performance from being achieved.

5. Experiment

5.1. Datasets and Evaluation Metrics

The nuScenes dataset is a dataset collected in Boston and Singapore. A total of 1,000 sequences with a sampling rate of 2 Hz and a duration of 20 seconds per sequence are provided, and the dataset consists of 40,000 sampled frames in total. Data were collected using one LiDAR, five radars, and six cameras, and only camera data were used in this study. The evaluation was conducted on a BEV plane with an area of 100 m × 100 m, and the grid resolution was set to 0.5 m.

The main evaluation metric is Intersection over Union (IoU). In this study, an IoU threshold of 0.50 was used as the performance evaluation criterion. To compare performance differences between complex and simple environments, IoU with occlusion at 0.50 was also evaluated, which reflects accuracy in more complex environments with occlusions. In addition to IoU, the loss value and inference time were measured as evaluation metrics. Subsequently, performance differences among the model with batch separation, the baseline model, and the adaptive model were compared. The adaptive model was configured to change its structure when the Scene Complexity exceeded 2.90.

For visual performance comparison, the BEV maps generated by the models were examined. Overall, since the IoU values between the baseline model and the proposed model fluctuated, the evaluation was conducted at steps where the IoU differences were not extreme.

5.2. Comparative Evaluation

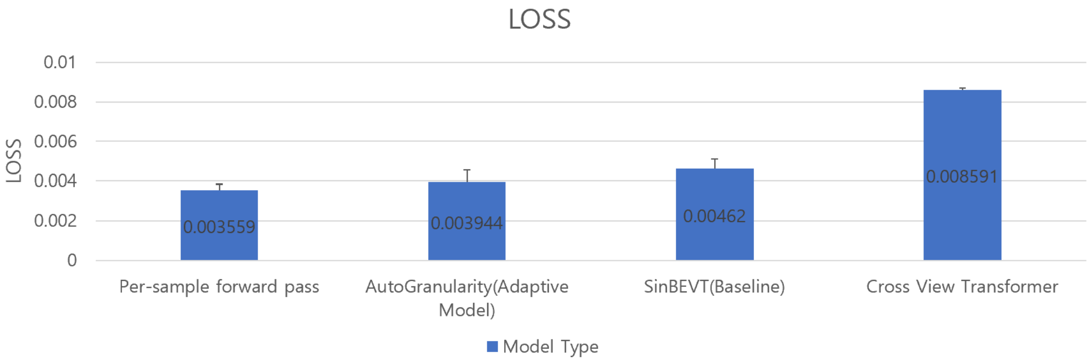


Figure 9. Comparative analysis of loss for AutoGranularity and comparative models

Figure 9 above shows the LOSS graphs of the Per-sample forward pass method, AutoGranularity, SinBEVT, and the Cross View Transformer. The average value of Scene Complexity is 2.93, and in this experiment, a value of 2.90, which is slightly lower than the average, was adopted and applied as the Scene Complexity threshold of AutoGranularity in order to overcome the low accuracy that is considered a bottleneck of camera-based models. The difference between the LOSS of AutoGranularity and that of the Per-sample forward pass method is 0.000385, and the difference between the LOSS of AutoGranularity and that of SinBEVT is 0.000676. It can be observed that the LOSS value of AutoGranularity is closer to that of the Per-sample forward pass method than to that of SinBEVT and is significantly low. However, in the IoU experiment below, the IoU value of AutoGranularity was obtained at a level between those of SinBEVT and the Per-sample forward pass method. Considering this result, it is more plausible that the low LOSS value is due to higher accuracy in the background surrounding objects rather than higher accuracy in object detection itself. From this, when AutoGranularity is applied to SinBEVT, the object detection accuracy is measured at an intermediate level between the two comparison models, while the accuracy of the surrounding background is relatively high. As a result, the probability of errors such as falsely detecting objects in locations where no objects actually exist is reduced.

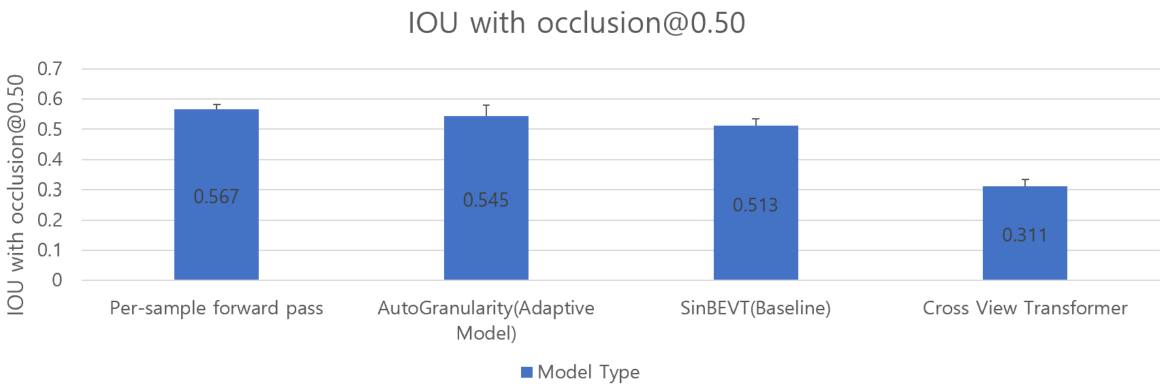


Figure 10. Comparative analysis of IOU with occlusion@0.50 for AutoGranularity and comparative models

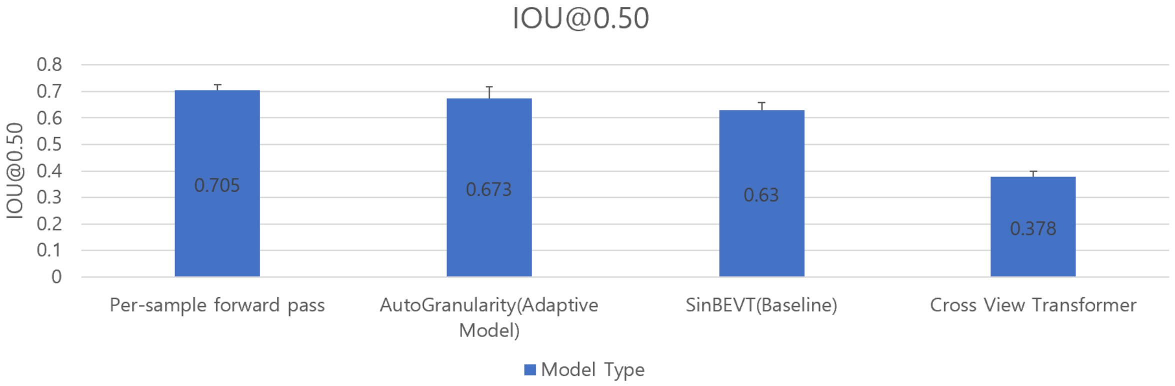


Figure 11. Comparative analysis of IOU@0.50 for AutoGranularity and comparative models

Figure 10 and Figure 11 above show the IoU graphs in an environment with occlusion and in a baseline environment, respectively. Camera-based models are fundamentally vulnerable to occlusion caused by other objects. Therefore, to evaluate the performance of AutoGranularity in occluded environments, the IoU performance is compared across the occluded environment and the baseline environment together with the comparison models. The Per-sample forward pass method exhibited an IoU decrease of 0.138 under occlusion. AutoGranularity showed an IoU decrease of 0.128 in the occluded environment. SinBEVT decreased by 0.117, and the Cross View Transformer decreased by 0.067. From these results, it is confirmed that the performance degradation caused by occlusion becomes more severe as the overall IoU performance increases. Nevertheless, AutoGranularity still demonstrates intermediate performance between the Per-sample forward pass method and SinBEVT. Therefore, AutoGranularity fulfills its objective by improving the insufficient IoU performance of SinBEVT in occluded environments while mitigating the real-time performance degradation of the Per-sample forward pass method.

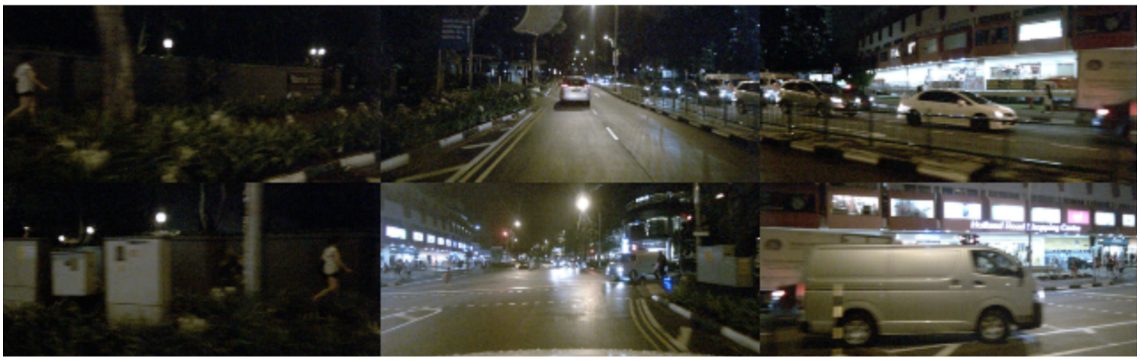


Figure 12. Camera images used for BEV map generation

The comparison of the generated BEV MAPs among SinBEVT, AutoGranularity, and the Per-sample forward pass method is conducted at a STEP where the difference in IoU performance is neither extremely large nor extremely small. The image used to generate the BEV MAP is the camera image shown in Figure 12.

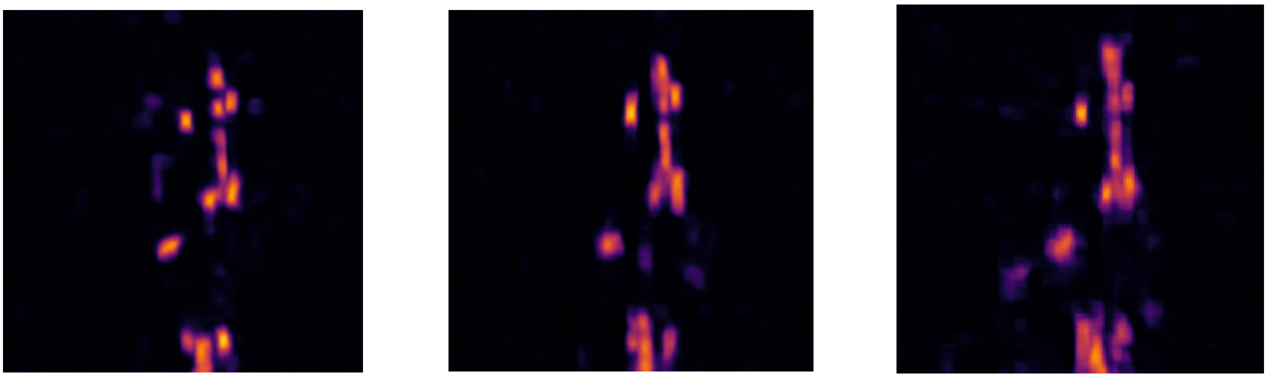


Figure 13. (Left) Per-sample forward pass Method, (Center) AutoGranularity, (Right) SinBEVT(Baseline Model)

Figure 13 shows the BEV MAP generated from Figure 12. Since the IoU values are highest in the order of the Per-sample forward pass method, AutoGranularity, and SinBEVT, models with relatively lower IoU values exhibit increased errors in the position, orientation, and size of the predicted boxes in the BEV MAP. In addition, object boundaries are estimated inconsistently, resulting in reduced overlap between the ground truth boxes and the predicted boxes.

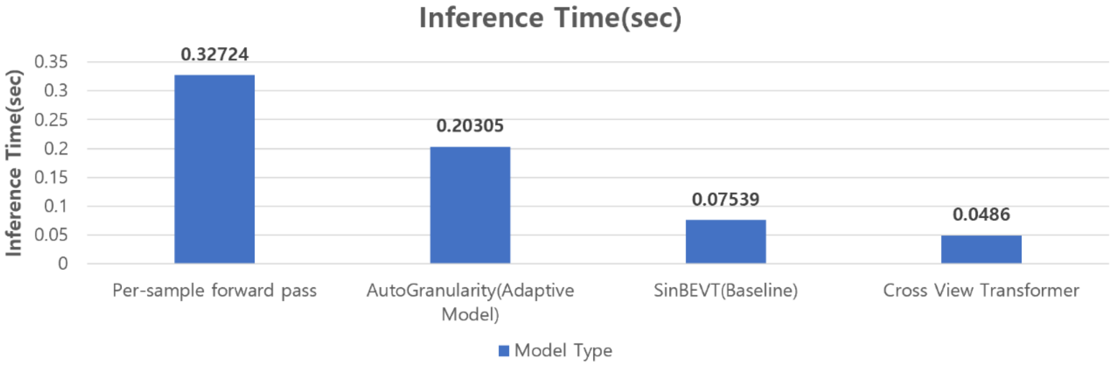


Figure 14. Comparative analysis of Inference Time for AutoGranularity and comparative models

Figure 14 presents the Inference Time of the Per-sample forward pass method, AutoGranularity, SinBEVT, and the Cross View Transformer. The reported Inference Time refers to the average inference time for a single batch. For the Per-sample forward pass method, AutoGranularity, and SinBEVT, one batch consists of eight samples, whereas for the Cross View Transformer, one batch consists of four samples. SinBEVT demonstrates very fast processing speed, consistent with the characteristics of camera-based models. AutoGranularity shows an inference time that is approximately 2.5 to 3 times slower than that of SinBEVT. However, considering that the Per-sample forward pass method is more than four times slower than SinBEVT, it can be confirmed that the degradation in real-time performance is alleviated to a certain extent. Although the inference time of AutoGranularity may be insufficient to stably operate in high-speed scenarios at the level of SinBEVT, it still demonstrates performance that is sufficient to maintain real-time operation.

6. Conclusion

6.1. Summary

In order to improve the accuracy of the camera-based BEV MAP generation model, this study proposes a method of separating the batch data input to the image backbone. In addition, to maintain real-time performance, the degradation of real-time capability is alleviated by adopting an adaptive approach that applies this method only in situations with high Scene Complexity. The IoU performance of AutoGranularity lies between that of SinBEVT and the Per-sample forward pass method in both non-occluded and occluded environments. However, the LOSS value of AutoGranularity is 0.003944, which is closer to that of the Per-sample forward pass method (0.003559) than to that of SinBEVT (0.00462). Considering this, AutoGranularity demonstrates higher accuracy in the background surrounding objects rather than in the objects themselves. In addition, by applying an adaptive structure, AutoGranularity reduces the inference time of the Per-sample forward pass method by more than 0.1 seconds, thereby alleviating the degradation in real-time performance.

6.2. Limitations

Since the inference time of the adaptive model represents only the average inference time, there remains a question as to whether AutoGranularity can truly secure real-time performance. Due to the adaptive structure, a shorter inference time is observed in simple environments, whereas a longer inference time is observed in complex environments. Therefore, in complex environments where the structure is transformed by the batch separation method, the adaptive model requires approximately 0.3 seconds to process a single batch. This indicates that there is uncertainty as to whether the adaptive model can stably maintain real-time performance in complex environments such as intersections.

Shannon entropy is a criterion for measuring the color diversity of pixels. Therefore, at night, the diversity naturally decreases, which interferes with the assessment of the complexity of the surrounding environment. Although gamma correction and histogram equalization are applied to address this issue, they result in a reduced dynamic range of the Shannon entropy values. As a consequence, since the Shannon entropy values do not vary significantly, its relative contribution to Scene Complexity is reduced compared to Object Count and Object Distribution.

6.3. Future Work

In order to address the narrow variation range of Shannon entropy, future research can be directed toward modifying the definition of Scene Complexity. By treating Shannon entropy as a separate criterion, it is possible to apply an additional filtering step to datasets that have already been categorized based on Object Count and Object Distribution. In addition, a simple form of feature extractor can be placed before the image backbone to compute feature entropy, which can then be used as a substitute for Shannon entropy.

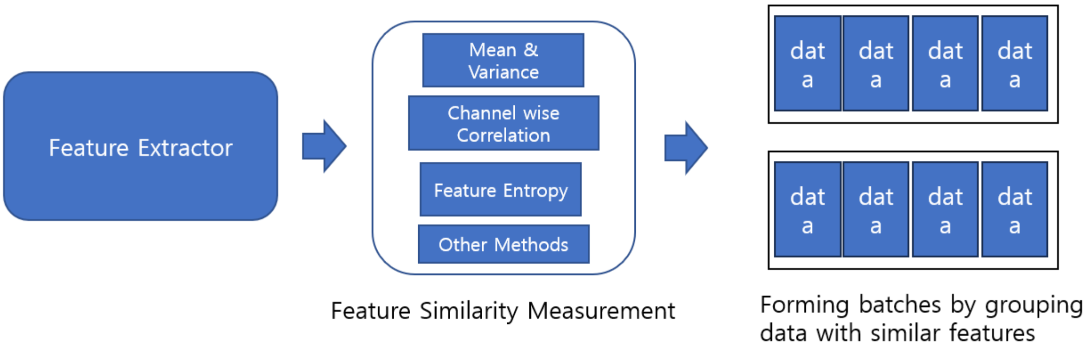


Figure 15. A batching strategy that groups data with similar feature representations

To resolve the real-time performance issue in complex environments, the batch separation method needs to be improved. The decrease in accuracy observed in the normalization modules and SE BLOCK of EfficientNet-B4 arises from the heterogeneous characteristics of data within a batch. Therefore, as illustrated in Figure 15, it may be effective to extract feature maps from camera images using a feature extractor and then compute inter-image feature similarity using statistics such as mean, variance, and channel-wise correlation. Based on the feature similarity determined in this manner, grouping data with similar characteristics into slightly smaller batch sizes and passing them through the image backbone may alleviate the degradation of real-time performance in complex environments.

References

1. B. Zhou and P. Kr¨ ahenb¨ uhl. Cross-view transformers for real-time map-view semantic segmentation. arXiv preprint arXiv:2205.02833, 2022.
2. Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, Jiaqi Ma. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. arXiv preprint arXiv:2207.02202, 2022.
3. Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, Jifeng Dai. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
4. Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, Anima Anandkumar. VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
5. Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, Alex Kendall. FIERY: Future Instance Prediction in Bird's-Eye View From Surround Monocular Cameras. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
6. Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik & Yinxiao Li. MaxViT: Multi-axis Vision Transformer. arXiv:2204.01697v4 [cs.CV] 9 Sep 2022.
7. Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, Liang Song. Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
8. Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, Yanfeng Wang. Collaboration Helps Camera Overtake LiDAR in 3D Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
9. Hao Xiang, Runsheng Xu, Jiaqi Ma. HM-ViT: Hetero-Modal Vehicle-to-Vehicle Cooperative Perception with Vision Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
10. Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, Yanfeng Wang. Robust Collaborative 3D Object Detection in Presence of Pose Errors. arXiv:2211.07214, 2022.
11. Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang & Jiaqi Ma. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. arXiv:2203.10638v3 [cs.CV] 8 Aug 2022.
12. Rui Song, Runsheng Xu, Andreas Festag, Jiaqi Ma, Alois Knoll. FedBEVT: Federated Learning Bird's Eye View Perception Transformer in Road Traffic Systems. IEEE TRANSACTIONS ONINTELLIGENT VEHICLES, VOL. 9, NO. 1, JANUARY 2024.
13. Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, Chen Feng. Among Us: Adversarially Robust Collaborative Perception by Consensus. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
14. Yangheng Zhao, Zhen Xiang, Sheng Yin, Xianghe Pang, Yanfeng Wang, Siheng Chen. MADE: Malicious Agent Detection for Robust Multi-Agent Collaborative Perception. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) October 14-18, 2024.
15. Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, Siheng Chen. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. Advances in Neural Information Processing Systems 35 (NeurIPS 2022).
16. Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, Lihua Zhang. How2comm: Communication-Efficient and Collaboration-Pragmatic Multi-Agent Perception. Advances in Neural Information Processing Systems 36 (NeurIPS 2023).
17. Jie Hu, Li Shen, Gang Sun. Squeeze-and-Excitation Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
18. Alexandr Pak, Atabay Ziyaden. Comparative analysis of deep learning methods of detection of diabetic retinopathy. Cogent Engineering, August 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.