Regularizing Neural Networks for BEV Semantic Segmentation via Inter-class Hierarchy and Spatially-aware Weight Adjustment

Jeongbin Hong[1,2], Dooseop Choi[1,2], Muhammad Atta ur Rahman[1,2], Kyounghwan An[2], Kyoung-Wook Min[2]
[1]*dept. Artificial Intelligence, University of Science and Technology*
[2]*dept. Autonomous Driving Intelligence Research, Electronics and Telecommunications Research Institute*
Daejeon, South Korea
{hjb3880@etri.re.kr, d1024.choi@etri.re.kr, rahman@etri.re.kr, mobileguru@etri.re.kr, kwmin92@etri.re.kr}

***Abstract*—In this work, we propose an effective multi-label semantic segmentation framework for Bird's Eye View (BEV) perception. While existing BEV frameworks typically employ a separate model for binary segmentation of each semantic class—achieving state-of-the-art performance per class—this design is impractical for real-world autonomous driving applications. A more feasible solution for deployment demands a single model capable of performing multi-label or multi-class segmentation across multiple object categories. To this end, we introduce two key strategies that enhance segmentation quality without modifying the architecture of existing BEV models. First, we incorporate multi-class prediction to capture inter-class hierarchies, allowing the model to learn dependencies between semantic categories such as roads, vehicles, and pedestrians. This improves semantic reasoning and boundary delineation, especially in spatially overlapping regions. Second, we propose a spatially-aware weight adjustment (SAWA) that emphasizes rare object zones on the BEV map. This addresses the inherent class imbalance and spatial sparsity of BEV segmentation tasks.**

***Index Terms*—Bird's Eye View, Semantic Segmentation, Autonomous Driving, Class Imbalance**

I. Introduction

A Bird's Eye View (BEV) map represents the surrounding road environment from a top-down perspective, centered on the ego-vehicle. Due to its comprehensive 360-degree global context, the BEV map is a critical representation for autonomous driving and is widely utilized in downstream tasks such as 3D object detection and segmentation [1]–[3]. Research on BEV map generation has largely followed two main paradigms. The first is the LSS-based approach, which originates from the pioneering Lift-Splat-Shoot (LSS) framework [4]. The second is based on transformers or query-driven architectures, often referred to as DETR-based methods [5].

LSS-based approaches rely on explicit depth estimation from images and employ a two-stage "lift and splat" pipeline. Specifically, 2D image features are first lifted into a 3D frustum volume, and then splatted onto a 2D BEV plane [4]. These methods are typically fast and computationally efficient. However, they suffer from limited performance due to their dependence on the accuracy of explicit depth estimation [2], [3], [6].

In contrast, DETR-based methods leverage transformers to model 2D image features more effectively, thereby achieving significantly improved accuracy without requiring explicit 3D scene construction or depth estimation [3], [7]. Despite these advantages, such approaches remain computationally expensive and resource-intensive [2], [7], [8].

While BEV perception models have continued to advance, most existing methods focus solely on binary segmentation for individual classes. This design choice allows optimal segmentation performance per class but leads to inefficiencies, as each model is restricted to recognizing a single class. Although this issue could be mitigated by adapting the model head for multi-class segmentation, such modifications often result in a severe degradation of per-class performance.

To address this limitation, we propose a multi-label semantic segmentation framework tailored for BEV perception. Specifically, our approach integrates multi-class prediction to model the inter-class hierarchy among drivable area, vehicle, and pedestrian classes, enabling the network to learn their semantic dependencies and spatial relationships. In addition, we introduce a spatially-aware weight adjustment (SAWA) that modulates the loss contribution of each region based on the presence of rare objects, thereby mitigating the effects of class imbalance and improving segmentation accuracy in critical areas. The contributions of this study are summarized as follows:

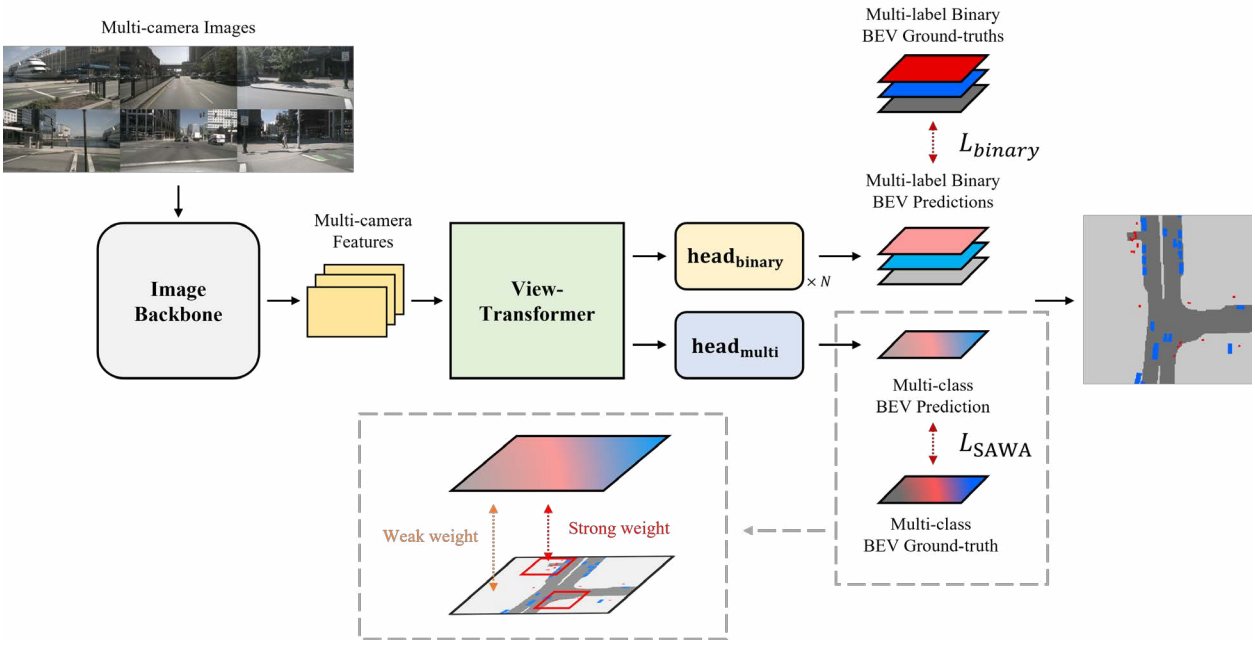

Fig. 1. The overall architecture of our approach for multi-label BEV segmentation. The multi-class segmentation head $\text{head}_{\text{multi}}$ is added to a standard view transformer-based BEV model without altering the architecture. Subsequently, the SAWA is applied to the multi-class prediction branch through the loss $\mathcal{L}_{\text{SAWA}}$. This mechanism operates by assigning higher weights to patches that contain rare objects, where the BEV space is pre-divided into $k \times k$ patches based on the ground-truth BEV map.

- We propose an inter-class hierarchical modeling approach that leverages the semantic dependencies among classes to improve segmentation performance. This is implemented via a multi-class prediction branch that captures the structural hierarchy across classes in the BEV space.
- We introduce a spatially-aware weight adjustment strategy that adjusts the training focus to spatial regions containing rare objects on the BEV map. This mitigates the effects of class imbalance and enhances accuracy in semantically important areas.
- We achieve noticeable performance improvements in the multi-label segmentation task with only the addition of a simple loss function without modifying the structure of the existing model.

II. RELATED WORK

A. LSS-based BEV Transformation

Lift-Splat-Shoot (LSS) [4] lifts each input image into a 3D frustum based on discretized depth distributions and then projects (splats) it onto the BEV plane. FIERY [9] extends this formulation by incorporating temporal information to perform A probabilistic future prediction of the BEV. BEVDet [10] proposes a modular end-to-end framework that efficiently detects 3D objects using a BEV encoder and a CenterPoint-based head. BEVNeXt [7] presents a modernized dense BEV framework that integrates depth estimation via a conditional random field (CRF) module and performs long-term temporal aggregation through the Res2Fusion module.

B. DETR-based BEV Transformation

DETR [5] introduces a transformer-based object detection method using object queries for 2D images, which has inspired the development of DETR-based BEV transformation approaches [1], [2], [11]–[13]. DETR3D [11] proposes a multi-view 3D object detection framework by projecting 3D queries onto the 2D image plane. Subsequent models [1], [2], [12], [13] further enhance performance by incorporating 3D positional embeddings. CVT [2] and PointBEV [3] address the high computational and memory demands that limit the deployment of DETR-based BEV models, offering more efficient alternatives for BEV perception.

Although these prior works have shown promising performance, they are fundamentally limited to single-label binary segmentation tasks for individual classes. When extended to multi-label segmentation, performance often degrades significantly. To overcome this limitation and build more practical BEV perception models for real-world scenarios, we propose a new multi-label segmentation methodology tailored for BEV frameworks.

C. Loss Functions for the Class Imbalance Problem

Focal loss [14] was originally introduced to address extreme class imbalance in dense object detection tasks

by down-weighting well-classified examples and focusing training on hard negatives. It has since been widely adopted in various segmentation tasks, especially where rare classes are underrepresented. Generalized focal loss [15] extends focal loss to multi-class segmentation settings by incorporating both classification and localization uncertainties. Asymmetric loss [16] penalizes false negatives more heavily than false positives to address the costs of asymmetric misclassification, particularly in the medical and semantic segmentation domains. Dice loss [17] directly optimizes the overlap between the predicted and ground-truth masks. This makes it especially suitable for scenarios involving small or sparse foreground classes in segmentation tasks.

Loss functions designed to address class imbalance, such as focal loss, have shown clear performance improvements for sparse object categories. However, these methods are limited to pixel-level comparisons only. In the context of BEV perception for autonomous driving, it is crucial not only to classify each BEV pixel correctly but also to capture the spatial context and relationships among surrounding objects [18]. Therefore, a loss function that accounts for the spatial structure of the BEV space and inter-object interactions is necessary.

III. PROPOSED METHODOLOGY

In this paper, we introduce the spatially-aware weight adjustment strategy through the multi-class prediction to address the challenge of class imbalance in the multi-label BEV segmentation task. The overall architecture is illustrated in Fig. 1. Without modifying the existing View Transformer (VT) responsible for projecting multi-view images to the BEV space, we simply append a single segmentation head for multi-class prediction at the final stage of the model. To further improve segmentation performance for rare classes such as pedestrians, while minimizing performance degradation for more frequent classes, we apply the proposed weighting scheme to the loss map. The new auxiliary loss is based on the focal loss. The general focal loss for each pixel for multiple classification is defined as follows:

$$\mathcal{L}_{\text{focal}}(x, y) = -\alpha_y (1 - p_y(x))^{\gamma} \log(p_y(x)) \tag{1}$$

p_y denotes the predicted probability for the ground-truth class y, typically obtained using the softmax function. A balancing factor α_y that adjusts the importance of class, used to mitigate class imbalance. The focusing parameter γ reduces the loss contribution from well-classified examples and emphasizes hard examples.

We extend this formulation to a spatially-aware weight adjustment objective that facilitates the understanding of spatial context. Fig. 2 shows the process. For this, we divide the BEV ground-truth map of size $n \times n = N$ into $k \times k$ non-overlapping patches. A spatially-aware weight $w_{k(i)}$ is applied to each pixel i that belongs to a patch $k(i)$ containing at least one pixel of a rare class:

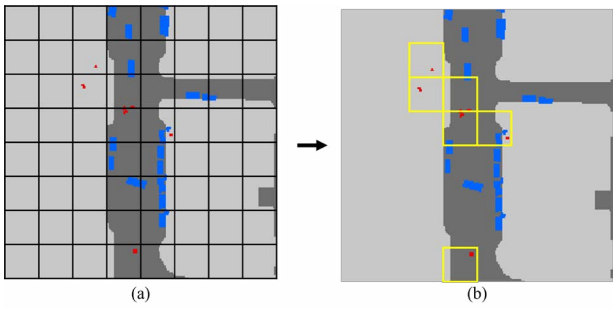

Fig. 2. Illustration of the SAWA process. The BEV map (a) is first divided into $k \times k$ patches. Subsequently, the higher weight is assigned to the corresponding patches in the loss map that spatially align with BEV regions containing rare classes, such as the red points representing pedestrians, as illustrated in (b). This strategy balances class importance while preserving generalization.

$$w_{k(i)} = \begin{cases} \lambda, & \text{if a rare class exists} \\ 1.0, & \text{otherwise} \end{cases} \tag{2}$$

Therefore, the average loss for the entire map is defined as follows:

$$\mathcal{L}_{\text{SAWA}} = \frac{1}{|N|} \sum_{i \in N} w_{k(i)} \cdot \mathcal{L}_{\text{focal}}(x_i, y_i) \tag{3}$$

$\mathcal{L}_{\text{SAWA}}$ denotes the spatially-aware weight adjustment loss. We integrate this with existing C class-wise binary focal loss objectives $\mathcal{L}_{\text{binary}}$ to obtain the final total loss $\mathcal{L}_{\text{Total}}$:

$$\mathcal{L}_{\text{Total}} = \sum_{j=1}^{C} \mathcal{L}_{\text{binary}}(j) + \mathcal{L}_{\text{SAWA}} \tag{4}$$

Unlike pixel-wise weighting methods that may cause overfitting to specific pixels or classes, our formulation, which integrates $\mathcal{L}_{\text{binary}}(j)$ and $\mathcal{L}_{\text{SAWA}}$, considers both global and local context in the BEV space. This balances the importance of rare object regions while preventing excessive overfitting, thus enhancing generalization across all classes. We adopt 2.0 for the weight factor λ.

IV. EXPERIMENT

A. Dataset

All experiments are conducted on the nuScenes [19] dataset, a large-scale benchmark for autonomous driving. This dataset comprises 1,000 driving scenes collected from Boston and Singapore, covering diverse weather and lighting conditions, as well as a wide range of driving scenarios. The 1,000 scenes are divided into 700 for training, 150 for validation, and 150 for test set. Each

scene spans approximately 20 seconds and is composed of image frames recorded at 2Hz intervals (i.e., every 0.5 seconds), resulting in a total of approximately 40,000 annotated samples. The images are captured from 6 synchronized, monocular cameras mounted around the ego-vehicle to provide a full 360-degree surround view. In addition, the dataset includes 3D point cloud data collected from one LiDAR and 5 radars, enabling multimodal perception tasks.

B. Implementation Details

We perform semantic segmentation for three classes, drivable-area, vehicle, and pedestrian. Following previous work [2], [4], we define the BEV map as a grid of size 200×200, where each pixel represents a 0.5×0.5m area. All input images are resized and top-cropped to a resolution of 224×480. No additional data augmentation is applied to the images. For baseline comparisons, we adopt CVT and BEVFormer [1] as representative models. To ensure consistency, we use the same optimizer and scheduler—AdamW [20] and OneCycleLR [21]—for both. For CVT, we follow the original configuration, using a learning rate of 4×10^{-3} and a weight decay of 1×10^{-7}, training for 30 epochs. BEVFormer is trained similarly to the original setup, with a learning rate of 2×10^{-4} and a weight decay of 1×10^{2} for 24 epochs. The loss weights for the three classes (drivable-area, vehicle, and pedestrian) are set to 1, 8, and 32, respectively. These weights are applied to both $\mathcal{L}_{\text{binary}}$ and $\mathcal{L}_{\text{SAWA}}$. Each experiment is conducted three times using different random seeds, and the mean performance across runs is reported as the final result. All experiments were performed on the nuScenes validation dataset.

TABLE I
IOU(%) COMPARISON OF MULTI-LABEL BEV SEGMENTATION ON THE NUSCENES VALIDATION DATASET.

Method	Drivable	Vehicle	Pedestrian	mIoU
CVT	76.8	31.7	10.8	39.8
+ $\mathcal{L}_{\text{SAWA}}$	**77.0**	**32.3**	**12.0**	**40.4**
BEVFormer	**78.6**	33.8	11.0	41.1
+ $\mathcal{L}_{\text{SAWA}}$	**78.6**	**34.5**	**11.8**	**41.6**

TABLE II
IOU(%) COMPARISON OF MULTI-LABEL BEV SEGMENTATION FOR THE DYNAMIC OBJECTS WITH VISIBILITY FILTERING ON THE NUSCENES VALIDATION DATASET.

Method	Vehicle	Pedestrian	mIoU
CVT	33.4	11.5	22.5
+ $\mathcal{L}_{\text{SAWA}}$	**34.2**	**12.7**	**23.5**
BEVFormer	36.0	11.4	23.7
+ $\mathcal{L}_{\text{SAWA}}$	**36.8**	**12.2**	**24.5**

C. Experimental Results

We compare our methodology with baseline models. As shown in TABLE I, introducing our proposed loss function, $\mathcal{L}_{\text{SAWA}}$ into both CVT [2] and BEVFormer [1] models improves multi-label semantic segmentation performance. Our approach continues to be effective even when visibility filtering is applied during training and evaluation, focusing on vehicles and pedestrians with less than 40% visibility (see TABLE II). All experimental results are the mean values of three experiments. Our proposed method appears to improve the BEV segmentation performance for all classes.

Notably, although the $\mathcal{L}_{\text{SAWA}}$ is applied based on the presence of pedestrians, the performance of other classes also improved or maintained. This is attributed to the patch-wise emphasis, which highlights not only the pedestrian regions but also neighboring objects that frequently co-occur or interact with them. Unlike pixel- or class-specific weighting schemes, this approach encourages the model to better capture the local and spatial context on the BEV map.

TABLE III
IOU(%) COMPARISON OF CVT MODELS BY VARYING NUMBER OF PATCHES.

Method	Number of Patches	Drivable	Vehicle	Pedestrian
CVT	-	76.8	31.7	10.8
+ $\mathcal{L}_{\text{SAWA}}$	5×5	76.7	32.0	11.9
+ $\mathcal{L}_{\text{SAWA}}$	8×8	76.8	31.9	11.5
+ $\mathcal{L}_{\text{SAWA}}$	**5×5 or 8×8**	**77.0**	**32.3**	**12.0**

TABLE IV
IOU(%) COMPARISON OF CVT ACCORDING TO MULTI-CLASS PREDICTION (MULTI-C) AND SAWA. THE BEST PERFORMANCE IS HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST IS INDICATED WITH AN UNDERLINE.

Method	Multi-C	SAWA	Drivable	Vehicle	Pedestrian
CVT			76.8	31.7	10.8
CVT	✓		**77.0**	<u>32.2</u>	<u>11.5</u>
CVT	✓	✓	**77.0**	**32.3**	**12.0**

D. Ablation Study

We conducted an ablation study to explore the optimal number of patches. Table III shows that randomly mixing 5×5 and 8×8 patches with a 50% probability yields the most effective performance across different classes for a BEV map of size 200×200. This mixing method can be interpreted to provide richer and more varied emphasis compared to the monotonous weights provided by a single strategy.

In TABLE IV, Multi-C means the case of adding only the multi-class segmentation head for the auxiliary loss

TABLE V
IOU(%) COMPARISON BETWEEN TWO TYPES OF AUXILIARY LOSSES. WE PERFORM A COMPARISON OF THE CONDITIONS UNDER WHICH PEDESTRIANS WERE EXCLUDED FROM EACH AUXILIARY LOSS.

Method	Classes used for auxiliary loss			Drivable	Vehicle	Pedestrian
	Drivable	Vehicle	Pedestrian			
CVT				76.8	31.7	**10.8**
CVT+Binary	✓	✓		**76.9**	31.6	9.3
CVT+Multi	✓	✓		**76.9**	**32.1**	**10.8**

without SAWA. Interestingly, even this alone leads to a performance improvement. This observation suggests that the multi-class objective implicitly encourages class-specific exclusivity in latent space, which can help reduce classification ambiguity in multi-label task.

To analyze this in more detail, we conduct an ablation study in which the multi-class segmentation head is replaced with the binary segmentation head for each class. Table V shows the results. The +Binary denotes that the auxiliary loss, which is the same as the existing binary segmentation loss, is added. Therefore, The CVT+Binary means that two binary segmentation are performed on the drivable-area and the vehicle. On the other hand, only one existing loss is used for the pedestrian. In the experiment where auxiliary losses were applied only to drivable-area and vehicles, the +Binary does not meaningfully improve the performance for both classes, but significantly decreases the performance for pedestrian. This shows that simply performing the same binary segmentation does not improve the predictive performance for that class, but rather leads to performance degradation of other classes that have not been emphasized.

In contrast, the +Multi significantly improves the performance on the vehicle class, while preserving the performance on the pedestrian even without any auxiliary loss applied to it. These results suggest that the multi-class objective captures hierarchical relationships between semantic classes—such as the fact that roads are always beneath vehicles and can be occluded by them—which conventional binary models fail to account for. This structured formulation helps to overcome the limitations of binary-based approaches while properly regularizing with the binary objectives applied in parallel.

V. CONCLUSION

In this paper, we presented a new auxiliary loss design for multi-label BEV semantic segmentation. Specifically, we introduced the spatially-aware weight adjustment via multi-class prediction. This novel auxiliary objective effectively emphasizes the local regions containing rare classes, without hindering global contextual focus. Even this does not require a change in the existing BEV perception structure.

In addition, we demonstrated through experiments that it is more effective to perform multi-class segmentation together on multi-label binary segmentation task. This is because multi-class prediction captures the structural hierarchy between classes such as drivable-area, vehicles, and pedestrian, regularizing to construct a more realistic BEV space.

Through comprehensive experiments on the nuScenes dataset, we demonstrated that our method consistently improves segmentation accuracy across all object categories. These results offer a simple yet robust extension to existing BEV segmentation frameworks.

ACKNOWLEDGMENT

This work was supported by IITP grant funded by the Korea government (MSIT) (RS-2023-00236245, Development of Perception/Planning AI SW for Seamless Autonomous Driving in Adverse Weather/Unstructured Environment)

REFERENCES

[1] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 760–13 769.

[3] L. Chambon, E. Zablocki, M. Chen, F. Bartoccioni, P. Pérez, and M. Cord, "Pointbev: A sparse approach for bev predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 195–15 204.

[4] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[6] S.-W. Lu, Y.-H. Tsai, and Y.-T. Chen, "Toward real-world bev perception: Depth uncertainty estimation via gaussian splatting," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 17 124–17 133.

[7] Z. Li, S. Lan, J. M. Alvarez, and Z. Wu, "Bevnext: Reviving dense bev frameworks for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 113–20 123.

[8] X. Liu, C. Zheng, M. Qian, N. Xue, C. Chen, Z. Zhang, C. Li, and T. Wu, "Multi-view attentive contextualization for multi-view 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 688–16 698.

[9] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[10] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[11] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[12] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 531–548.

[13] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3262–3272.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[15] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[16] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 82–91.

[17] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[18] T. An, J. Kang, D. Choi, and K.-W. Min, "Crfnet: Context refinement network used for semantic segmentation," *ETRI Journal*, vol. 45, no. 5, pp. 822–835, 2023.

[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[21] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.