2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

BridgeTA: Bridging the Representation Gap in Knowledge Distillation via Teacher Assistant for Bird's Eye View Map Segmentation

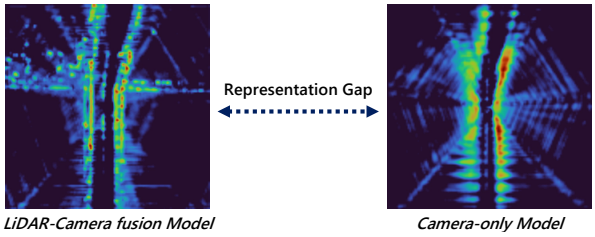
Beomjun Kim[1,2], Suhan Woo[1], Sejong Heo[2], and Euntai Kim[1,3,†]

***Abstract*— Bird's Eye View (BEV) map segmentation is one of the most important and challenging tasks in autonomous driving. Camera-only approaches have drawn attention as cost-effective alternatives to LiDAR, but they still fall behind LiDAR-Camera (LC) fusion-based methods. Knowledge Distillation (KD) has been explored to narrow this gap, but existing methods mainly enlarge the student model by mimicking the teacher's architecture, leading to higher inference cost. To address this issue, we introduce BridgeTA, a cost-effective distillation framework to bridge the representation gap between LC fusion and Camera-only models through a Teacher Assistant (TA) network while keeping the student's architecture and inference cost unchanged. A lightweight TA network combines the BEV representations of the teacher and student, creating a shared latent space that serves as an intermediate representation. To ground the framework theoretically, we derive a distillation loss using Young's inequality, which decomposes the direct teacher-student distillation path into teacher-TA and TA-student dual paths, stabilizing optimization and strengthening knowledge transfer. Extensive experiments on the challenging nuScenes dataset demonstrate the effectiveness of our method, achieving an improvement of 4.2% mIoU over the Camera-only baseline, up to 45% higher than the improvement of other state-of-the-art KD methods. The code will be available at https://github.com/kxxbeomjun/BridgeTA.**

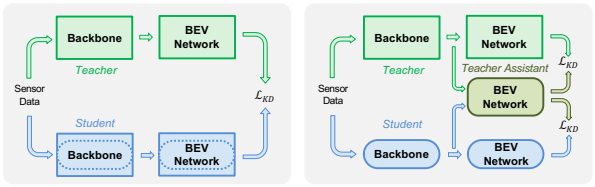
I. Introduction

Bird's Eye View (BEV) map segmentation is a fundamental task in perception [1], [2], [3], [4], playing a crucial role in autonomous driving [5], [6]. By analyzing road components from a top-down perspective, BEV map segmentation provides essential spatial understanding, which is critical for ensuring the safety of autonomous vehicles. To balance cost and efficiency, recent research has increasingly focused on Camera-only methods [1], [2], [7], [8], [9], achieving notable performance improvements. However, despite these advancements, Camera-only methods still lag behind LiDAR-Camera fusion-based approaches in terms of performance [2], [10], [11].

To bridge this performance gap, BEV-based Knowledge Distillation (KD) methods, which distill knowledge from a fusion-based teacher to a camera-only student, have been attracting significant attention [12], [13], [14]. In these methods, the representation gap due to the different input modalities between the teacher and student, as visualized in Figure 1a, serves as a major obstacle. Existing approaches attempt to reduce this gap by designing the student model to follow the structure of the teacher model [15], [16], as depicted schematically in Figure 1b. However, such methods increase the computational burden on the student, undermining the main advantage of KD, which is to achieve strong performance with a compact student. Additionally, many methods [14], [17] perform suboptimal distillation simply by forcing the student to replicate the representation of the teacher without adequately considering the inherent differences between them.



(a) Visualization of BEV feature representation gap



(b) Existing KD methods vs. **BridgeTA**

Fig. 1: Due to fundamental differences in sensor modalities, the LiDAR-Camera fusion model and Camera-only model capture distinct information for semantic map segmentation, leading to the representation gap shown in (a). To cost-effectively address this gap, we propose BridgeTA, as illustrated in (b).

In this paper, we propose **BridgeTA**, a novel distillation framework to **Bridge** the representation gap through an innovative **T**eacher **A**ssistant network. Our TA structure can be applied without any modification to the student model and is used only during training. Consequently, it introduces no additional inference cost. Uniquely, our TA is constructed by combining the representations of the teacher and student, requiring no individual data input or dedicated backbone for the TA itself. Unlike previous TA-based KD methods [18], [19], this design makes our training process highly cost-efficient. By leveraging TA, we decompose the direct teacher-student distillation path into teacher-TA and TA-student dual paths, thereby alleviating the severe representation mismatch between the teacher and the student. To theoretically support

† Corresponding author.
[1]B. Kim, S. Woo, and E. Kim are with Yonsei University, South Korea, {kbj,wsh112,etkim}@yonsei.ac.kr.
[2]B. Kim and S. Heo are with Hyundai Motor Company, South Korea, {kim.beomjun,sejong.heo}@hyundai.com.
[3]E. Kim is with Korea Institute of Science and Technology, South Korea.

our framework, we design our distillation loss based on *Young's inequality* [20]. Furthermore, to maximize the effect of distillation, we also perform multi-level distillation through three distinct schemes. Rather than merely mimicking representations, our approach improves distillation efficiency by explicitly transferring the relative geometric relationships essential for encoding BEV representations. This enables the student to effectively capture crucial spatial dependencies.

Extensive experiments on the nuScenes dataset [21] demonstrate that BridgeTA enhances BEV representation for BEV map segmentation while maintaining a cost-effective Camera-only setting during inference. Notably, BridgeTA decomposes the single distillation path into dual paths and achieves superior performance, significantly narrowing the gap between Camera-only and fusion-based models without incurring any additional cost or latency.

To summarize, our main contributions are as follows:

- We introduce a novel TA-based distillation framework that bridges the representation gap between LC fusion teachers and Camera-only students without changing the student architecture or inference cost.
- We theoretically ground our dual-path distillation loss with *Young's inequality*, providing a tight upper bound for stable and effective knowledge transfer.
- We design a multi-level distillation scheme and conduct extensive experiments on the nuScenes dataset [21], where BridgeTA achieves a 4.2% mIoU improvement over the Camera-only baseline, up to 45% higher than the gains of other state-of-the-art KD methods.

II. Related Work

1) BEV map segmentation: Recently, BEV map segmentation has attracted significant attention in autonomous driving perception. Fusion-based methods combine LiDAR's precise geometric and 3D spatial information with camera inputs to deliver superior performance and can be categorized according to the fusion stage, such as early fusion [22], deep fusion [2], [10], and late fusion [23], [24]. While these approaches enhance BEV representations, they introduce high computational overhead and LiDAR dependency costs. As a result, Camera-only methods [25], [26] have recently attracted increasing attention. To improve BEV representation and prediction performance, Camera-only approaches have adopted advanced techniques such as attention mechanisms [8], [27], [1] and diffusion models [28]. However, the increasing complexity of these methods often leads to higher inference latency, diminishing their efficiency advantages over fusion-based methods. To overcome this limitation, we propose a novel distillation framework that improves the performance of Camera-only methods without additional inference cost or complexity.

2) Knowledge Distillation (KD): Initially, KD was introduced as an effective approach for transferring knowledge from a high-capacity teacher to a lightweight student for model compression [29]. Then, it has been extended to vision tasks such as semantic segmentation [30] and object detection [12]. While early KD methods focused primarily on logit distillation [31], recent studies have explored feature distillation [32], [33] to provide richer guidance. However, KD methods often struggle to close representation and performance gaps between teacher and student when these gaps are large. Some approaches have addressed this by introducing a teacher assistant (TA) to mediate distillation, but existing TA-based methods [18], [19] typically require extra data inputs and a dedicated backbone for the TA, resulting in inefficient, resource-intensive training. Other works have tried to reduce the gap by aligning the student's structure with the teacher's [15], [16], but this usually incurs significant inference cost. In this work, we propose a cost-effective TA-based distillation framework that overcomes these limitations by efficiently bridging the teacher-student gap without increasing training or inference complexity.

III. Method

In this section, we detail the novelty and cost-effectiveness of the BridgeTA framework. We first present the architectures of the teacher, TA, and student models. Next, we introduce the distillation path decomposition method leveraging Young's inequality, and subsequently provide detailed descriptions of each level in the proposed multi-level distillation scheme. An overview of BridgeTA is shown in Figure 2.

A. Model architecture

1) Teacher and Student: BridgeTA is developed based on the BEVFusion [2] codebase, which is widely used for BEV map segmentation tasks. The teacher model utilizes both the LiDAR and Camera branches, while the student uses only the Camera branch. To enable effective knowledge transfer, we align the teacher's channel dimensions and decoder with the student's structure. Unlike previous methods [15], [16] that modify the student architecture, our approach fully preserves the original student model.

2) Teacher Assistant (TA): We propose a novel TA-based distillation framework to address both the representation and performance gaps in BEV map segmentation. BridgeTA introduces a TA network that requires no individual data input or dedicated backbone, fusing the teacher's LiDAR BEV feature with the student's Camera BEV feature, as in Figure 2. The TA shares BEV decoder and head architecture with both teacher and student for consistency. This structure decomposes the direct teacher-to-student path into teacher-to-TA and TA-to-student. As a result, the student receives finer-grained and less divergent guidance, rather than a single large distillation signal from the teacher. Therefore, the training process becomes more stable and effective, allowing better alignment.

B. Distillation Loss Formulation

Our framework, which decomposes the direct distillation path into dual-paths, is theoretically grounded by recalling the classical *Young's inequality* [20], which states that for any $a, b \in \mathbb{R}$ and $\lambda > 0$,

$$|ab| \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2, \tag{1}$$

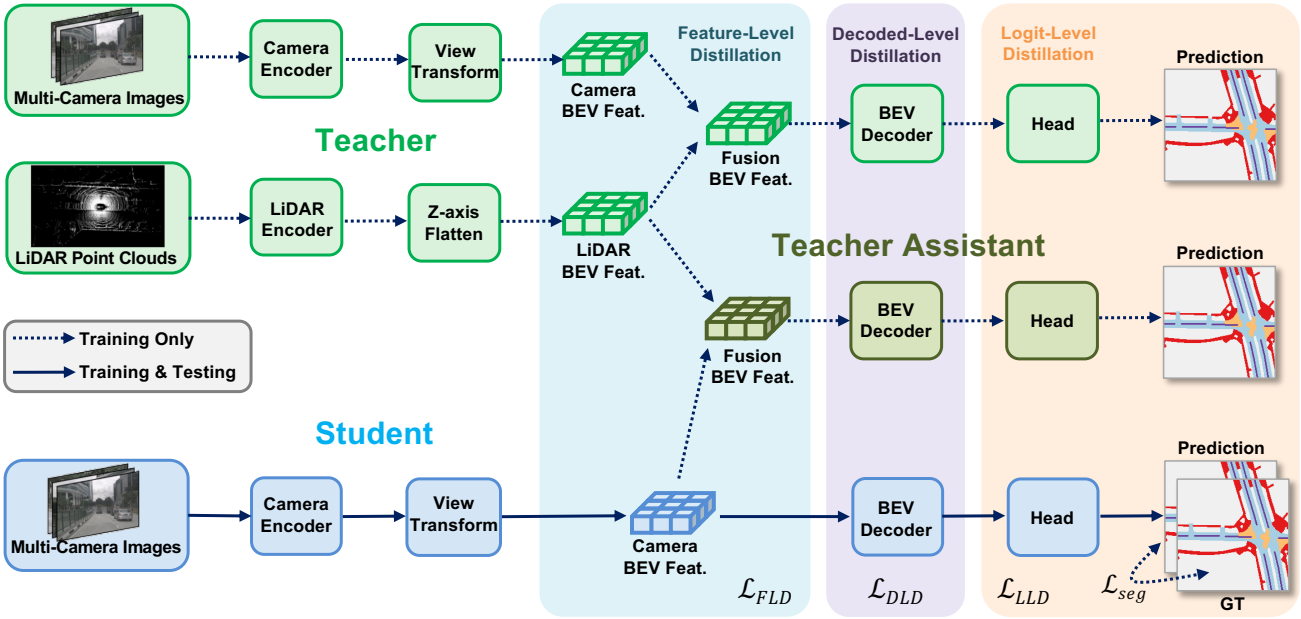


Fig. 2: **Overview of BridgeTA framework.** We propose a novel distillation framework to bridge the representation gap between LiDAR-Camera fusion teacher (green) and Camera-only student (blue) through a Teacher Assistant network (dark green). We design a multi-level distillation framework with three distinct distillation modules for the student model to learn rich BEV representation from the teacher model through an efficient KD process. The teacher model and TA network are used only during training, ensuring that inference relies solely on the Camera, without LiDAR.

this inequality extends to vectors or matrices using the L2 norm or Frobenius norm, respectively. Specifically, for any X, Y and $\lambda > 0$,

$$2\langle X, Y\rangle \leq \lambda \|X\|^2 + \frac{1}{\lambda}\|Y\|^2. \tag{2}$$

Expanding the squared norm of a sum gives:

$$\|X + Y\|^2 = \|X\|^2 + 2\langle X, Y\rangle + \|Y\|^2, \tag{3}$$

and by applying Eq. (2) to the cross-term in Eq. (3) and then setting $\lambda = \varepsilon$ leads to:

$$\|X + Y\|^2 \leq (1 + \varepsilon) \cdot \|X\|^2 + \left(1 + \frac{1}{\varepsilon}\right) \cdot \|Y\|^2. \tag{4}$$

Finally, substituting $X = R^S - R^{TA}$, $Y = R^{TA} - R^T$ into Eq. (4), we obtain the following inequality:

$$\begin{aligned}\|R^S - R^T\|^2 \leq (1 + \varepsilon) \cdot \|R^S - R^{TA}\|^2 \\ + \left(1 + \frac{1}{\varepsilon}\right) \cdot \|R^{TA} - R^T\|^2,\end{aligned} \tag{5}$$

where R^S, R^T, and R^{TA} denote the student, teacher, and TA representations at each distillation level, respectively.

Here, the left-hand side represents the direct teacher-to-student distillation, while the right-hand side is our proposed dual-path distillation loss. If the sum of the teacher-to-TA and TA-to-student losses converges, the upper bound ensures that the student and teacher representations also become close. In other words, effective optimization of our dual-path objective leads to successful alignment of the student with the teacher.

Furthermore, to achieve the tightest upper bound of the inequality, we define the following objective function, Eq. (6) shown below:

$$f(\varepsilon) = (1 + \varepsilon) \cdot a^2 + \left(1 + \frac{1}{\varepsilon}\right) \cdot b^2, \tag{6}$$

where $a = \|R^S - R^{TA}\|$, $b = \|R^{TA} - R^T\|$, and $\varepsilon > 0$. To find the optimal value, we take the derivative of $f(\varepsilon)$ with respect to ε and solve for zero, yielding

$$\frac{df}{d\varepsilon} = a^2 - \frac{b^2}{\varepsilon^2} = 0 \quad \Longrightarrow \quad \varepsilon^* = \frac{b}{a}. \tag{7}$$

Here, ε^* is an optimal value which achieves the tightest possible upper bound for our decomposition. Setting $\varepsilon = \varepsilon^*$ ensures a strong theoretical guarantee that the teacher-student representation gap will be effectively bridged through our TA-based distillation framework.

C. Multi-Level Knowledge Distillation

To maximize distillation effectiveness in bridging both the representation and performance gaps, we propose a multi-level distillation framework comprising three schemes: Feature-Level Distillation (FLD), Decoded-Level Distillation (DLD), and Logit-Level Distillation (LLD). The primary advantage of multi-level distillation is that it captures complementary information across network stages, enabling comprehensive knowledge transfer. These three schemes share a common concept of a dual distillation path between the teacher-TA and TA-student, as illustrated in Figure 3 and Figure 4. We employ mean square error (MSE) between

corresponding representations to unify the loss formulation across all levels. Specifically, the teacher-to-TA and TA-to-student objectives are defined as follows:

$$\mathcal{L}^{X}_{t2ta} = \frac{1}{H_X \times W_X} \sum_{i}^{H_X} \sum_{j}^{W_X} \left\| (R^{TA}_X)_{i,j} - (R^{T}_X)_{i,j} \right\|^2,$$
$$\mathcal{L}^{X}_{ta2s} = \frac{1}{H_X \times W_X} \sum_{i}^{H_X} \sum_{j}^{W_X} \left\| (R^{S}_X)_{i,j} - (R^{TA}_X)_{i,j} \right\|^2, \tag{8}$$

where R^T_X, R^{TA}_X, and R^S_X denote the teacher, TA, and student representations at each distillation level X. In addition, H_X, W_X represent the corresponding spatial dimensions. The precise forms of these variables depend on the chosen distillation level (e.g., feature, decoded, or logit), and are specified in the respective subsections below.

1) Feature-Level Distillation (FLD): At the feature level, the representation gap between the teacher and student arises from their different input modalities: the teacher captures both geometric information from LiDAR and semantic texture from Cameras, while the student relies only on Camera inputs. To bridge this gap, we propose distillation at the BEV feature level using a TA network that fuses the teacher's LiDAR BEV features with the student's Camera BEV features, thereby forming an intermediate representation that integrates both geometric richness and semantic information. Through this process, the student not only learns from the teacher's LiDAR-enhanced representation, but also implicitly adapts its own Camera features to better align with the LiDAR modality for more effective fusion.

For the FLD stage, we instantiate the general loss formulation in Eq. (8) by specifying the variables as follows:

$$(R^T_X,\, R^{TA}_X,\, R^S_X) = (F^T_{fus},\, F^{TA}_{fus},\, F^S_{cam}),$$

where F^T_{fus} and F^{TA}_{fus} are the fused BEV features of the teacher and TA, and F^S_{cam} is the Camera BEV feature of the student. The spatial size is $H_f \times W_f$. We define these features as $F^T_{fus} \in \mathbb{B}^{C^T_f \times H_f \times W_f}$, $F^{TA}_{fus} \in \mathbb{B}^{C^{TA}_f \times H_f \times W_f}$, and $F^S_{cam} \in \mathbb{B}^{C^S_c \times H_f \times W_f}$. To facilitate distillation, we set the channel and spatial dimensions to be identical, i.e., $C^T_f = C^{TA}_f = C^S_c$ and $H_f \times W_f$. The resulting feature-level distillation loss is then formulated as

$$\mathcal{L}_{\text{FLD}} = (1+\varepsilon^*) \cdot \mathcal{L}^F_{ta2s} + \left(1 + \frac{1}{\varepsilon^*}\right) \cdot \mathcal{L}^F_{t2ta}, \tag{9}$$

where ε^* is optimally chosen as discussed previously.

2) Decoded-Level Distillation (DLD): The decoded level serves as an intermediate stage between the feature level and the prediction head, playing a vital role in refining high-level scene understanding and spatial relationships. Distilling at this stage enables the transfer of more structured and semantically rich information, thereby helping the student model to interpret complex scenes more accurately.

For the DLD stage, we instantiate the general dual-path loss formulation in Eq. (8) by specifying as:

$$(R^T_X,\, R^{TA}_X,\, R^S_X) = (F^T_{dec},\, F^{TA}_{dec},\, F^S_{dec}),$$

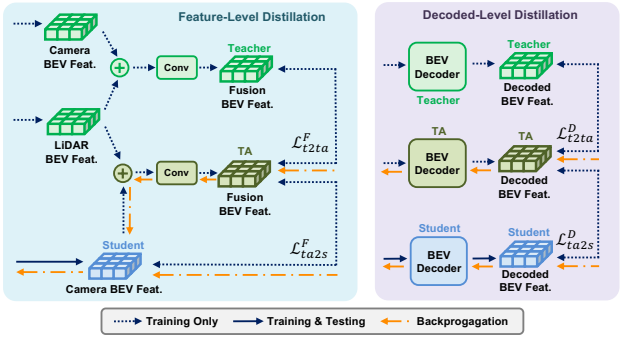


Fig. 3: Illustration of **Feature-Level Distillation (FLD)** and **Decoded-Level Distillation (DLD)**. Both FLD and DLD utilize dual-path distillation with the TA network to bridge the BEV representation gap between the teacher and student models.

where F^T_{dec} and F^{TA}_{dec} denote the decoded features of the teacher and TA networks, and F^S_{dec} is the decoded feature of the student. The spatial size at this stage is $H_d \times W_d$. All decoded features are defined as $F^T_{dec}, F^{TA}_{dec}, F^S_{dec} \in \mathbb{B}^{C_d \times H_d \times W_d}$, where C_d, H_d, and W_d denote the shared channel and spatial dimensions. The resulting decoded-level distillation loss is then formulated as

$$\mathcal{L}_{\text{DLD}} = (1+\varepsilon^*) \cdot \mathcal{L}^D_{ta2s} + \left(1 + \frac{1}{\varepsilon^*}\right) \cdot \mathcal{L}^D_{t2ta}, \tag{10}$$

where ε^* is optimally chosen as described previously.

3) Logit-Level Distillation (LLD): At the logit level, distillation directly influences model performance by aligning the predictions of teacher and student. To minimize representation and performance gaps, we extend the dual-path KD approach to the logit level using the TA network, enabling the student to better capture the teacher's logit distributions.

For the LLD stage, we instantiate Eq. (8) with

$$(R^T_X,\, R^{TA}_X,\, R^S_X) = (L^{T_T},\, L^{TA_{TA}},\, L^{S_S}),$$

where L^{T_T} and $L^{TA_{TA}}$ are the logit outputs of the teacher and TA heads (given their respective features), and L^{S_S} is the student's logit output. The spatial size at this stage is $H \times W$. The base loss of LLD module is formulated as:

$$\mathcal{L}^L_{Base} = (1+\varepsilon^*) \cdot \mathcal{L}^B_{ta2s} + \left(1 + \frac{1}{\varepsilon^*}\right) \cdot \mathcal{L}^B_{t2ta}, \tag{11}$$

where ε^* is chosen as in the FLD and DLD losses.

We further enhance logit-level distillation with a structure inspired by CrossKD [34]. In this setup, the decoded BEV representation from the student is passed not only to the student's Head but also to the teacher's and TA's Heads, generating S_S, S_T, and S_{TA} predictions, as shown in Figure 4. This design enriches the distillation paths and enables more comprehensive knowledge transfer, as the well-trained heads of the teacher and TA provide the student with diverse and informative guidance.

By applying each head to both its own feature and the student's feature, we obtain four logit outputs: L^{T_T}, $L^{TA_{TA}}$,

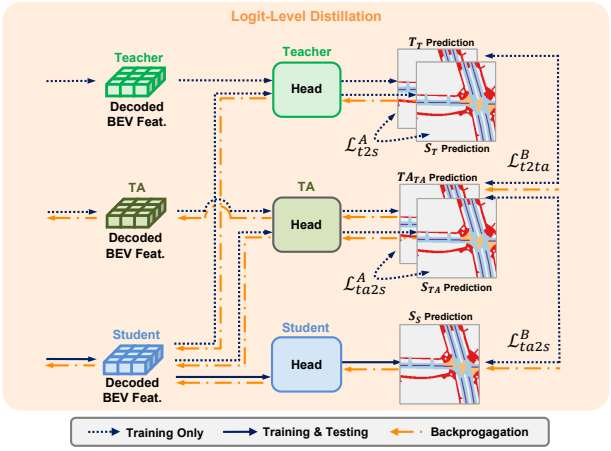


Fig. 4: Illustration of **Logit-Level Distillation (LLD)**. LLD utilizes multi-path distillation to effectively bridge both logit representation and performance gaps between the teacher and student models, with four distinct loss terms that provide richer guidance and accelerate training convergence.

L^{S_T}, and $L^{S_{TA}}$. This configuration allows us to employ KL divergence as an auxiliary distillation loss between outputs of the same head under different inputs, as shown below:

$$\begin{aligned} \mathcal{L}^L_{Aux} &= \mathcal{L}^A_{t2s} + \mathcal{L}^A_{ta2s}, \\ \mathcal{L}^L_{Aux} &= KL(L^{T_T} \parallel L^{S_T}) + KL(L^{TA_{TA}} \parallel L^{S_{TA}}). \end{aligned} \tag{12}$$

This additional supervision accelerates convergence and strengthens the student to learn robust and reliable logit representations, ultimately narrowing the performance gap. The overall logit-level distillation loss combines the dual-path regression and the auxiliary KL terms, leveraging both fine-grained regression and distribution-level alignment:

$$\mathcal{L}_{LLD} = \mathcal{L}^L_{Base} + \mathcal{L}^L_{Aux}. \tag{13}$$

Here, L^{T_T}, $L^{TA_{TA}}$, L^{S_S}, L^{S_T}, and $L^{S_{TA}}$ all denote logit outputs in $\mathbb{R}^{N_c \times H \times W}$, where N_c is the number of classes and $H \times W$ is the BEV prediction map size.

D. Training and Inference

1) Training: We formulate BEV map segmentation as a pixel-wise classification problem and optimize the student model with a segmentation loss. The total loss used to train BridgeTA is:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_1 \cdot \mathcal{L}_{FLD} + \lambda_2 \cdot \mathcal{L}_{DLD} + \lambda_3 \cdot \mathcal{L}_{LLD}, \tag{14}$$

where $\mathcal{L}_{\text{seg}}$ is the segmentation loss of the student model, which is applied to maximize the Intersection over Union (IoU) value. The hyperparameters λ_1, λ_2, and λ_3 are introduced to balance the contributions of the distillation losses, $\mathcal{L}_{\text{FLD}}, \mathcal{L}_{\text{DLD}}$, and $\mathcal{L}_{\text{LLD}}$, respectively.

Moreover, the introduction of level-specific weights can be theoretically justified by Young's inequality. For each distillation level ℓ, the discrepancy between student and teacher representations can be bounded by the Teacher-TA and TA-Student terms. Extending this to all levels with weights λ_ℓ, we obtain:

$$\begin{aligned} \sum_\ell \lambda_\ell \|R^S_\ell - R^T_\ell\|^2 \le \sum_\ell \lambda_\ell \Big[&(1+\varepsilon_\ell)\|R^S_\ell - R^{TA}_\ell\|^2 \\ &+ \left(1 + \frac{1}{\varepsilon_\ell}\right) \|R^{TA}_\ell - R^T_\ell\|^2 \Big], \end{aligned} \tag{15}$$

where $\varepsilon_\ell > 0$ is a positive scalar. This formulation provides a principled explanation for introducing λ_ℓ, aligning the total loss with a theoretically grounded upper bound.

2) Inference: In our approach, we fully optimize the Camera-only student model through a novel distillation framework, enabling it to capture rich BEV representations comparable to those of the LC fusion-based teacher model. Furthermore, by leveraging a cost-effective distillation framework, we use only the Camera branch of the student model during the inference stage, excluding the teacher and TA networks at this stage, thereby avoiding any increase in computational cost or inference latency.

IV. Experiments

A. Implementation Details

1) Dataset: To assess the effectiveness of our method for BEV map segmentation, we conduct experiments on the nuScenes [21] dataset, a widely used benchmark. The dataset contains approximately 1.4M Camera images, 390K LiDAR sweeps, and HD maps covering 40K keyframes, collected in Boston and Singapore using a 32-channel LiDAR and six RGB Cameras. Annotations are provided for keyframes. nuScenes encompasses diverse challenging conditions, such as rain, night, and complex intersections, enabling a comprehensive evaluation of BridgeTA's robustness and effectiveness. We follow the experimental protocol of LSS [7] and implement BridgeTA using MMDetection3D [39].

2) Train setting: All experiments were conducted on 2 NVIDIA RTX A6000 GPUs with the teacher model frozen. The student model was trained for 20 epochs using a batch size of 6, a learning rate of 1e-4, and a Cosine Annealing schedule [40].

B. State-of-the-Art Comparison Results

1) KD methods: We compared BridgeTA with state-of-the-art (SOTA) KD methods in BEV networks, as shown in Table I. To ensure fairness, we re-implemented these methods on the BEVFusion [2] codebase to match our experimental settings. BridgeTA achieves higher accuracy across all classes, with up to 45% over other KD methods (4.2 vs. 2.9 mIoU). Moreover, as shown in Table III, since BridgeTA introduces no extra computational cost or latency over the baseline, it achieves up to 1.2× faster inference (16.2 vs. 13.3 FPS), 12% lower memory (7.6 vs. 8.6 GB), 14% fewer FLOPs (465.6 vs. 541.0 G), and 12% fewer parameters (31.8 vs. 36.2 M) than other KD methods during inference. These results underscore the efficiency and effectiveness of our KD framework.

Method	Modality	IoU↑(%)						
		Drivable	Ped. Cross	Walkway	Stopline	Carpark	Divider	Mean
SimDistill [15]	LC → C	82.4	57.3	61.3	51.1	54.4	48.3	59.2
MapDistill [16]	LC → C	82.6	57.4	61.7	51.6	54.5	48.8	59.5
BridgeTA (Ours)	LC → C	**83.3**	**58.6**	**62.9**	**53.6**	**56.6**	**50.1**	**60.8**

TABLE I: Comparison against state-of-the-art KD methods. We compare BEV map segmentation performance on the nuScenes validation set. BridgeTA surpasses all other KD methods, achieving top performance across all classes. LC → C denotes distillation from a LiDAR-Camera fusion teacher model to a Camera-only student model.

Method	IoU↑(%)						
	Drivable	Ped. Cross	Walkway	Stopline	Carpark	Divider	Mean
LSS [7]	75.4	38.8	46.3	30.3	39.1	36.5	44.4
CVT [8]	74.3	36.8	39.9	25.8	35.0	29.4	40.2
M^2BEV [35]	77.2	-	-	-	-	40.5	-
BEVFusion-C [2]	81.7	54.8	58.4	47.4	50.7	46.4	56.6
MapPrior [9]	81.7	54.6	58.3	46.7	53.3	45.1	56.7
X-Align [36]	82.4	55.6	59.3	49.6	53.8	47.4	58.0
MetaBEV [37]	83.3	56.7	61.4	50.8	55.5	48.0	59.3
DDP [38]	**83.6**	58.3	61.6	52.4	51.4	49.2	59.4
RGC [25]	81.7	57.1	60.5	51.7	53.8	**53.5**	59.7
BridgeTA (Ours)	83.3	**58.6**	**62.9**	**53.6**	**56.6**	50.1	**60.8**

TABLE II: Comparison against state-of-the-art Camera-only BEV map segmentation methods. We compare performance on the nuScenes validation set. BridgeTA achieves the highest mIoU overall and outperforms all other methods in 4 out of 6 classes.

Method	Latency (ms)	FLOPs (G)	Params (M)
Baseline	61.6	456.6	31.8
SimDistill [15]	69.2(+7.6)	493.5(+36.9)	32.4(+0.6)
MapDistill [16]	75.2(+13.6)	541.0(+84.4)	36.2(+4.4)
BridgeTA (Ours)	**61.6(+0.0)**	**456.6(+0.0)**	**31.8(+0.0)**

TABLE III: Computational efficiency comparison against state-of-the-art KD methods. Unlike other KD methods, BridgeTA incurs no additional computation cost or latency.

Method	$\mathcal{L}_{FLD}$	$\mathcal{L}_{DLD}$	$\mathcal{L}_{LLD}$		mIoU↑(%)
			$\mathcal{L}^L_{Base}$	$\mathcal{L}^L_{Aux}$	
Ours w/o TA	✓	✓	✓	-	57.8
	✓	✓	✓	✓	58.2
Ours	✓	-	-	-	58.5
	✓	✓	-	-	59.4
	✓	✓	✓	-	60.5
	✓	✓	✓	✓	**60.8**

TABLE IV: Ablation study of the TA network, multi-level distillation, and logit-level losses in BridgeTA.

2) Camera-only BEV map segmentation method: We also compare ours with state-of-the-art Camera-only BEV map segmentation methods. As shown in Table II, BridgeTA achieves the highest mIoU among them. Specifically, it improves performance by 4.2% mIoU over the baseline, BEVFusion-C, demonstrating effective knowledge transfer.

C. Ablation Studies

1) Effect of TA network and ε setting: After theoretically justifying the effectiveness of the TA network with Young's inequality, we empirically validate this in ablation experiments. In Table IV, Ours w/o TA, which excludes the TA network and applies only direct teacher-student distillation, achieves lower mIoU than our full approach. These results confirm that the TA network with dual-path distillation leads to more effective knowledge transfer.

We visualize the L_2 distance between teacher and student representations across all levels throughout training in Figure 5. The results show that TA network enables faster, more stable convergence with consistently lower L_2 distances compared to Ours w/o TA.

Furthermore, as shown in Figure 5, applying the theoretically derived optimal weight ε^* results in the smallest gap. This experimentally confirms that our mathematically determined value of ε^* is indeed optimal for bridging the teacher-student representation gap.

2) Multi-Level Distillation: Table IV further shows that applying multi-level distillation terms, FLD, DLD, and LLD, to BridgeTA progressively boosts performance. Notably, incorporating all three levels simultaneously yields the best performance, which further highlights the complementary nature of each distillation term.

3) Logit-Level Distillation (LLD): We conducted ablations on the LLD loss components. As shown in the rightmost columns of Table IV, combining both $\mathcal{L}^L_{Base}$ and $\mathcal{L}^L_{Aux}$

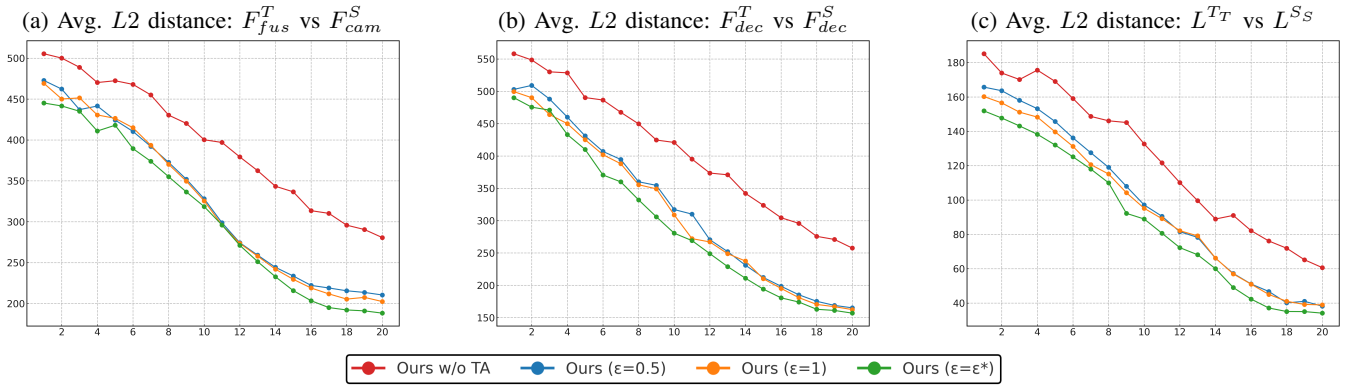


Fig. 5: **Visualization of the representation gap throughout training.** (a) FLD, (b) DLD, and (c) LLD plot the L_2 distance between teacher and student representations across epochs for different distillation settings. "Ours w/o TA" denotes multi-level distillation without a TA network. The X-axis represents epochs, and the Y-axis indicates L_2 distance.

Fig. 6: **Qualitative results on nuScenes.** By effectively distilling geometric and structural information from the teacher, BridgeTA predicts essential road elements like Lane Dividers and Walkways more accurately than the Camera-only baseline even under challenging night-time conditions, demonstrating the effectiveness of the TA network.

leads to the highest mIoU, indicating that both regression and distribution-based logit losses are essential.

D. Qualitative Results

We present the visualization of BEV map segmentation results in Figure 6 to highlight the effectiveness of BridgeTA and the role of the TA network in night-time scenarios. Since the baseline model relies solely on Camera inputs, it struggles to accurately predict essential road elements such as Lane Dividers and Walkways under low-light conditions. In contrast, BridgeTA leverages the fused LiDAR-Camera information from the teacher model to effectively distill both geometric and structural information. This enables the student to generate richer and more accurate feature representations even in night-time environments. This allows BridgeTA to deliver robust prediction results, significantly better than the baseline and Ours w/o TA. The results demonstrate the necessity and advantages of the TA network in handling challenging conditions, and BridgeTA consistently shows robust performance not only in night-time scenarios but also under other challenging driving scenarios.

V. Conclusions

In this paper, we proposed BridgeTA, a distillation framework for BEV map segmentation that leverages a lightweight TA network to mitigate the representation gap between an LC fusion teacher and Camera-only student. Unlike prior work, BridgeTA preserves the student architecture with no additional inference cost. Our framework is theoretically grounded by Young's inequality, which enables a principled decomposition of the distillation path into teacher-TA and TA-student with a tight alignment bound.

Extensive experiments demonstrate that BridgeTA effectively overcomes the representation gap and maintains robust performance even under adverse conditions. We believe our findings provide a promising direction for scalable and cost-effective BEV perception in autonomous driving.

Acknowledgements

This work was supported by the KIST Institutional Program (No. 2E33801-25-015), by a grant from the Korea Evaluation Institute of Industrial Technology (KEIT), funded by the Korean government (MOTIE) (No. 20023455), and by the Autonomous Driving Center, Hyundai Motor Company.

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] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[3] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5935–5943.

[4] C. Pan, Y. He, J. Peng, Q. Zhang, W. Sui, and Z. Zhang, "Baeformer: Bi-directional and early interaction transformers for bird's eye view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9590–9599.

[5] S. Ren, S. Chen, and W. Zhang, "Collaborative perception for autonomous driving: Current status and future trend," in *Proceedings of 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control*. Springer, 2022, pp. 682–692.

[6] J. Zhao, J. Shi, and L. Zhuo, "Bev perception for autonomous driving: State of the art and future perspectives," *Expert Systems with Applications*, vol. 258, p. 125103, 2024.

[7] 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.

[8] 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.

[9] X. Zhu, V. Zyrianov, Z. Liu, and S. Wang, "Mapprior: Bird's-eye view map layout estimation with generative models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8228–8239.

[10] M. Kim, G. Kim, K. H. Jin, and S. Choi, "Broadbev: Collaborative lidar-camera fusion for broad-sighted bird's eye view map construction," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 125–11 132.

[11] D.-T. Le, H. Shi, J. Cai, and H. Rezatofighi, "Diffuser: Diffusion model for robust multi-sensor fusion in 3d object detection and bev segmentation," *arXiv preprint arXiv:2404.04629*, 2024.

[12] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, "Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8637–8646.

[13] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Bevdistill: Cross-modal bev distillation for multi-view 3d object detection," *arXiv preprint arXiv:2211.09386*, 2022.

[14] S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, "Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5116–5125.

[15] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang, and D. Tao, "Simdistill: Simulated multi-modal distillation for bev 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7460–7468.

[16] X. Hao, R. Li, H. Zhang, D. Li, R. Yin, S. Jung, S.-I. Park, B. Yoo, H. Zhao, and J. Zhang, "Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 166–183.

[17] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[18] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[19] H.-I. Liu, C. Wu, J.-H. Cheng, W. Chai, S.-Y. Wang, G. Liu, H. Latapie, J.-C. Wu, J.-N. Hwang, H.-H. Shuai *et al.*, "Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 266–22 275.

[20] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*. Cambridge University Press, 1952.

[21] 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.

[22] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.

[23] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[24] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 172–181.

[25] Q. Chen and X. Qi, "Residual graph convolutional network for bird's-eye-view semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3324–3331.

[26] S. Woo, M. Park, Y. Lee, S. Lee, and E. Kim, "Location-aware transformer network for bird's eye view semantic segmentation," *IEEE Transactions on Intelligent Vehicles*, 2024.

[27] S. Chen, T. Cheng, X. Wang, W. Meng, Q. Zhang, and W. Liu, "Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer," *arXiv preprint arXiv:2206.04584*, 2022.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[29] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[30] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 319–12 328.

[31] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[32] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1921–1930.

[33] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *European conference on computer vision*. Springer, 2020, pp. 41–55.

[34] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, and Q. Hou, "Crosskd: Cross-head knowledge distillation for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 520–16 530.

[35] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "M ̂2 bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation," *arXiv preprint arXiv:2204.05088*, 2022.

[36] S. Borse, M. Klingner, V. R. Kumar, H. Cai, A. Almuzairee, S. Yogamani, and F. Porikli, "X-align: Cross-modal cross-view alignment for bird's-eye-view segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3287–3297.

[37] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for bev detection and map segmentation," *arXiv preprint arXiv:2304.09801*, 2023.

[38] Y. Ji, Z. Chen, E. Xie, L. Hong, X. Liu, Z. Liu, T. Lu, Z. Li, and P. Luo, "Ddp: Diffusion model for dense visual prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 741–21 752.

[39] M. Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection," https://github.com/open-mmlab/mmdetection3d, 2020.

[40] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.