LIX: Implicitly Infusing Spatial Geometric Prior Knowledge Into Visual Semantic Segmentation for Autonomous Driving

Sicen Guo, *Graduate Student Member, IEEE*, Ziwei Long, Zhiyuan Wu, Qijun Chen, *Senior Member, IEEE*, Ioannis Pitas, *Life Fellow, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Despite the impressive performance achieved by data-fusion networks with duplex encoders for visual semantic segmentation, they become ineffective when spatial geometric data are not available. Implicitly infusing the spatial geometric prior knowledge acquired by a data-fusion teacher network into a single-modal student network is a practical, albeit less explored research avenue. This article delves into this topic and resorts to knowledge distillation approaches to address this problem. We introduce the Learning to Infuse "X" (LIX) framework, with novel contributions in both logit distillation and feature distillation aspects. We present a mathematical proof that underscores the limitation of using a single, fixed weight in decoupled knowledge distillation and introduce a logit-wise dynamic weight controller as a solution to this issue. Furthermore, we develop an adaptively-recalibrated feature distillation algorithm, including two novel techniques: feature recalibration via kernel regression and feature consistency quantification via centered kernel alignment. Extensive experiments conducted with intermediate-fusion and late-fusion networks across various public datasets provide both quantitative and qualitative evaluations, demonstrating the superior performance of our LIX framework when compared to other state-of-the-art approaches. Source code is available at https://mias.group/LIX.**

***Index Terms*—Semantic segmentation, spatial geometric prior knowledge, data-fusion, knowledge distillation.**

I. Introduction

In THE domain of data-driven autonomous driving perception, a prevailing consensus among researchers asserts that "the increased availability of well-annotated training data is strongly correlated with improved learning performance." When examining visual semantic segmentation as an illustrative case, data-fusion networks, equipped with duplex encoders to acquire knowledge from both RGB images and spatial geometric information, consistently demonstrate superior performance compared to conventional single-modal networks trained solely on RGB images [1], [2]. This performance improvement is due to their ability to learn heterogeneous features from diverse data sources [3], [4]. RGB images primarily capture rich color and texture cues, whereas other visual data sources, commonly designated as "X", contain informative spatial geometric priors. The fusion of these features allows for a more comprehensive understanding of the driving environment [5], [6].

However, a significant limitation of data-fusion networks stems from their dependence on the availability of the "X" data, which can be problematic in scenarios devoid of range sensors [7]. Additionally, when the accuracy of the "X" data falls below expectations, possibly due to issues such as camera-LiDAR calibration errors [8], [9], the fusion of these heterogeneous features can potentially lead to a degradation in the overall performance of visual semantic segmentation [5]. As a result, the implicit infusion of spatial geometric prior knowledge from a teacher network (a data-fusion network trained with RGB-X data) to a student network (a single-modal network trained exclusively with RGB images) emerges as an interesting research direction worth pursuing. It is reasonable to consider that knowledge distillation (KD) techniques can be a viable solution to achieve this objective.

Existing KD techniques are generally classified into two main categories: logit distillation (LD) and feature distillation (FD). The former approaches [10], [11], [12] train the student network to replicate the logits of the teacher network by minimizing the divergence or distance between the probability

Received 14 March 2025; revised 20 August 2025; accepted 24 September 2025. Date of publication 23 October 2025; date of current version 10 November 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288; in part by the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, under Grant HMHAI-202406; in part by the Fundamental Research Funds for the Central Universities, NIO University Program (NIO UP); in part by Xiaomi Young Talents Program; and in part by European Commission-European Union through the Horizon Europe (Horizon Research and Innovation Actions) under Grant 101093003 (TEMA) HORIZON-CL4-2022-DATA-01-01. The associate editor coordinating the review of this article and approving it for publication was Prof. Vittoria Bruni. *(Corresponding author: Rui Fan.)*

Sicen Guo, Ziwei Long, Zhiyuan Wu, and Qijun Chen are with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: guosicen@tongji.edu.cn; zwlong@tongji.edu.cn; gwu@tongji.edu.cn; qjchen@tongji.edu.cn).

Ioannis Pitas is with the Department of Informatics, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece (e-mail: pitas@csd.auth.gr).

Rui Fan is with the College of Electronic and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Key Laboratory of Intelligent Autonomous Systems, the State Key Laboratory of Autonomous Intelligent Unmanned Systems, and the Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Tongji University, Shanghai 201804, China, and also with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: rui.fan@ieee.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2025.3618378, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3618378

distributions generated by both networks. On the other hand, the latter approaches [13], [14], [15], [16], [17], [18] leverage the abundant information available in the activations, neurons, and features of the teacher network's intermediate layers to provide guidance and supervision for the training of the student network.

Nonetheless, directly applying these existing techniques to our specific problem remains challenging due to three considerations. First, while there are several existing KD methods for cross-modal distillation, such as [19], [20], [21] (distillation from RGB images and LiDAR point clouds to LiDAR point clouds only) and [22], [23] (distillation from multi-frame point clouds to single-frame point clouds), they are not specifically designed for the task addressed in this study. Direct application of such algorithms to our task is infeasible due to the differences in network architectures that arise from distinct input modalities. Second, differences in architecture between teacher and student networks result in discrepancies in heterogeneous feature characteristics, such as dimensions, magnitudes, and distributions. These discrepancies form a barrier to the effective infusion of spatial geometric prior knowledge [24]. Third, comprehensive feature consistency measurement should be emphasized and requires further exploration, as it directly impacts the overall performance of FD [13], [15], [16].

This article introduces **L**earning to **I**nfuse "**X**" (**LIX**) framework (see Fig. 1), designed to implicitly infuse spatial geometric prior knowledge acquired from a data-fusion teacher network into a single-modal student network. We make three major contributions to address the above-mentioned limitations. We begin by revisiting the LD theory based on decoupled knowledge distillation (DKD) [25] and reformulate its loss as a weighted combination of target class LD (TCLD) and non-target class LD (NCLD) losses. By deriving the gradient of the LD loss function with respect to the student logit, we expose the limitations of the DKD algorithm, which relies on a single, fixed weight. Consequently, we design a dynamic weight controller (DWC), capable of generating a weight for each logit, thereby improving the overall LD performance. As for FD, we first introduce an adaptive feature recalibration approach based on kernel regression, which aligns the features of teacher and student networks across various dimensions (spatial, channel, magnitude, and distribution). Finally, we resort to the centered kernel alignment (CKA) [26] algorithm based upon Hilbert-Schmidt independence criterion (HSIC) [27] to formulate our novel FD loss, which quantifies the feature consistency between teacher and student networks. These contributions collectively improve the effectiveness of implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving. In a nutshell, our contributions can be summarized as follows:

- We implicitly infuse spatial geometric prior knowledge into visual semantic segmentation by distilling an RGB-X data-fusion teacher network into a single-modal student network that operates solely on RGB images.
- We present the novel dynamically-weighted LD (DWLD) algorithm, which extends the DKD algorithm, by assigning an appropriate weight to each logit, resulting in better performance compared to the baseline algorithm.
- We introduce the novel adaptively-recalibrated FD (ARFD) algorithm that performs feature recalibration via kernel regression and feature consistency measurement leveraging HSIC-based CKA.
- We have conducted extensive experiments using representative RGB-X semantic segmentation networks on multiple public datasets to quantitatively and qualitatively validate the effectiveness of our introduced novel LD and FD techniques.

The remainder of this article is structured as follows: Sect. II reviews related works. Sect. III details our proposed LIX. Sect. IV compares LIX with other state-of-the-art (SoTA) methods and presents the ablation study results. Limitations and extendability of our proposed LIX framework are discussed in Sect. V. Finally, in Sect. VI, we summarize this article and provide recommendations for future work.

II. Related Work

A. RGB-X Semantic Segmentation

According to the data-fusion stage, SoTA RGB-X semantic segmentation networks can be grouped into three classes: early-fusion, intermediate-fusion, and late-fusion networks [28], [29]. Early-fusion approaches generally combine RGB and X images at the input level. Such a straightforward yet simplistic data-fusion strategy has limitations in capturing a deep understanding of the environment [28]. In contrast, intermediate-fusion approaches [6], [30], [31] typically extract heterogeneous features from RGB and X images using duplex encoders. These features are subsequently fused within the encoder to fully exploit their inherent characteristics. Similar to intermediate-fusion methods, late-fusion approaches [32], [33], [34] use two parallel encoders (one for RGB images and one for X data) to extract heterogeneous features. However, these methods primarily focus on data fusion within the decoder. In this article, we utilize two representative data-fusion models as our baseline networks, namely SNE-RoadSeg [6] (a computationally intensive, intermediate-fusion network) and MFNet [32] (a lightweight, late-fusion network), to comprehensively validate the effectiveness of our proposed LIX framework.

B. Knowledge Distillation

KD methods have demonstrated significant potential in balancing accuracy and efficiency across various applications. In object detection, instance-aware distillation (InsDist) [35] enhances feature learning by decoupling instance-related features and modeling inter-instance relationships. RadarDistill [36] improves Radar representations by transferring knowledge acquired from LiDAR point clouds through feature distillation. In image segmentation, MSTNet-KD [37] enables a compact student network to learn semantic information extracted from a complex teacher network via multi-level semantic alignment. Moreover, in object tracking, the distilled Siamese tracker (DST) [38] adopts a specialized distillation strategy combined with mutual learning among students, enabling more effective knowledge transfer from teacher to student. These innovations highlight KD's versatility across

modalities and tasks, underscoring its broad applicability in real-world scenarios.

The first LD method [12] guides the student network to mimic the distribution characteristics of soft targets generated by the teacher network. Unlike this original KD method [12], channel-wise KD (CWKD) [11] normalizes activations within each channel to generate a probability map. Subsequently, the Kullback-Leibler (KL) divergence [12] is utilized to minimize the discrepancy between the probability maps generated by teacher and student networks. Moreover, weighted soft label distillation (WSLD) [39] dynamically adjusts the sample-wise bias-variance trade-off during model training by re-weighting the soft label loss. However, in these previous works [11], [12], [39], [40], the TCLD and NCLD terms are treated as interdependent and coupled terms. The recent study DKD [25] suggests that the coupled formulation of KD limits the effectiveness and flexibility of knowledge transfer, and it is necessary to reformulate the classical KD loss into two independent and decoupled terms.

LD methods depend exclusively on output logits and do not incorporate supervision to ensure feature consistency between teacher and student networks. Such feature-level supervision has been demonstrated to be crucial for effective representation learning [41]. The first FD method, FitNet [42], directly matches feature activations between teacher and student networks. Attention transfer (AT) [13] resorts to both activation-based and gradient-based spatial attention maps to realize FD. Several subsequent studies, *e.g*, similarity-preserving (SP) KD [17] and variational information distillation (VID) [18], have sought to reduce the dimension of features to accommodate the increasing depth of networks. Factor transfer (FT) [14] introduces a paraphraser and a translator to extract and transfer structured knowledge, whereas relational knowledge distillation (RKD) [16] measures structural consistency between teacher and student networks using distance-wise and angle-wise losses. Neuron selectivity transfer (NST) [15] formulates KD as a distribution matching problem by aligning neuron selectivity patterns through maximum mean discrepancy minimization. More recent methods [43], [44], [45] enable the model to concentrate on the most informative parts of the data. For instance, masked distillation (MasKD) [43] employs masks to emphasize essential feature regions, while augmentation-free dynamic curriculum distillation (Af-DCD) [44] adjusts task difficulty based on the student's learning progress. Moreover, prime-aware adaptive distillation (PAD) [45] prioritizes informative samples based on uncertainty metrics, making it particularly beneficial for class-imbalanced datasets.

However, the aforementioned algorithms are not designed specifically to infuse spatial geometric prior knowledge for semantic segmentation. In this article, we make significant advancements in both LD and FD strategies. We introduce a logit-wise dynamic weight controller to address the limitations of using a single, fixed weight in DKD. Additionally, we develop an adaptively-recalibrated feature distillation algorithm that includes feature recalibration and feature consistency quantification. Extensive experiments demonstrate the superior performance of our proposed LIX framework over all other algorithms reviewed.

III. Methodology

A. Overall Workflow

As depicted in Fig. 1, our proposed LIX framework can implicitly infuse spatial geometric prior knowledge acquired from a data-fusion teacher network into a single-modal student network. DWLD first reformulates the LD loss as a weighted combination of TCLD and NCLD losses. Then, a dynamic weight controller generates a weight for each student's logit and formulates the overall LD loss. As for FD, an adaptive feature recalibration approach first aligns the features of teacher and student networks across various dimensions (spatial, magnitude, and distribution). Subsequently, the CKA algorithm based on HSIC quantifies the consistency between different features and formulates our FD loss. In summary, the overall loss can be formulated as a combination of the initial loss of ground truth and distillation losses.

B. Dynamically-Weighted Logit Distillation

Most existing LD methods [11], [12] primarily resorted to more effective regularization and optimization methods, rather than proposing novel distillation strategies. In this subsection, we delve deeper into the LD theory and reframe its loss as the weighted combination of TCLD and NCLD losses. We derive the gradient of the LD loss with respect to the student logit and expose limitations of the traditional DKD algorithm [25], which relies on a single, fixed weight. Drawing inspiration from these findings, we introduce a novel LD method, which is capable of dynamically generating a weight for each logit, thereby improving the overall performance of spatial geometric prior knowledge infusion.

*1) **Reformulating Logit Distillation**:* Let $\hat{\boldsymbol{p}} = \left[\hat{p}_1, \ldots, \hat{p}_k, \ldots, \hat{p}_C\right]^{\top} \in \mathbb{R}^C$ be a column vector, storing probabilities $\hat{p}_i = P(y = i|\boldsymbol{q}) = \exp(z_i)/\sum_{j=1}^{C} \exp(z_j)$ $(\forall i \in [1, C] \cap \mathbb{Z})$ of pixel $\boldsymbol{q}$ belonging to C classes, where y denotes the predicted label of $\boldsymbol{q}$ and z_i denotes its logit with respect to the i-th class. Let $\boldsymbol{b} = \left[\hat{p}_k, 1 - \hat{p}_k\right]^{\top} \in \mathbb{R}^2$ be a binary probability vector of the target class k. Let $\hat{\boldsymbol{p}}_{\setminus k} = \hat{\boldsymbol{p}}/(1 - \hat{p}_k) = \left[\hat{p}_{1,\setminus k}, \ldots, \hat{p}_{k-1,\setminus k}, \hat{p}_{k+1,\setminus k}, \ldots, \hat{p}_{C,\setminus k},\right]^{\top} \in \mathbb{R}^{(C-1)}$ be a column vector, storing independently modeled probabilities among non-target classes (*i.e*, without considering the k-th class). The conventional LD uses the KL divergence [12] between output probabilities $\hat{\boldsymbol{p}}^{\mathcal{T}}$ and $\hat{\boldsymbol{p}}^{\mathcal{S}}$ of the teacher and student networks, respectively. Its loss function is expressed as follows[1]:

$$\text{LD} = \text{KL}\left(\hat{\boldsymbol{p}}^{\mathcal{T}} \| \hat{\boldsymbol{p}}^{\mathcal{S}}\right) = \hat{p}_k^{\mathcal{T}} \log \frac{\hat{p}_k^{\mathcal{T}}}{\hat{p}_k^{\mathcal{S}}} + \sum_{i=1, i\neq k}^{C} \hat{p}_i^{\mathcal{T}} \log \frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}, \quad (1)$$

which can be reformulated as follows [25]:

$$\text{LD} = \underbrace{\hat{p}_k^{\mathcal{T}} \log \frac{\hat{p}_k^{\mathcal{T}}}{\hat{p}_k^{\mathcal{S}}} + \left(1 - \hat{p}_k^{\mathcal{T}}\right) \log \frac{1 - \hat{p}_k^{\mathcal{T}}}{1 - \hat{p}_k^{\mathcal{S}}}}_{\text{KL}(\boldsymbol{b}^{\mathcal{T}} \| \boldsymbol{b}^{\mathcal{S}})}$$

[1] $\mathcal{T}$ and $\mathcal{S}$ denote teacher and student networks, respectively.

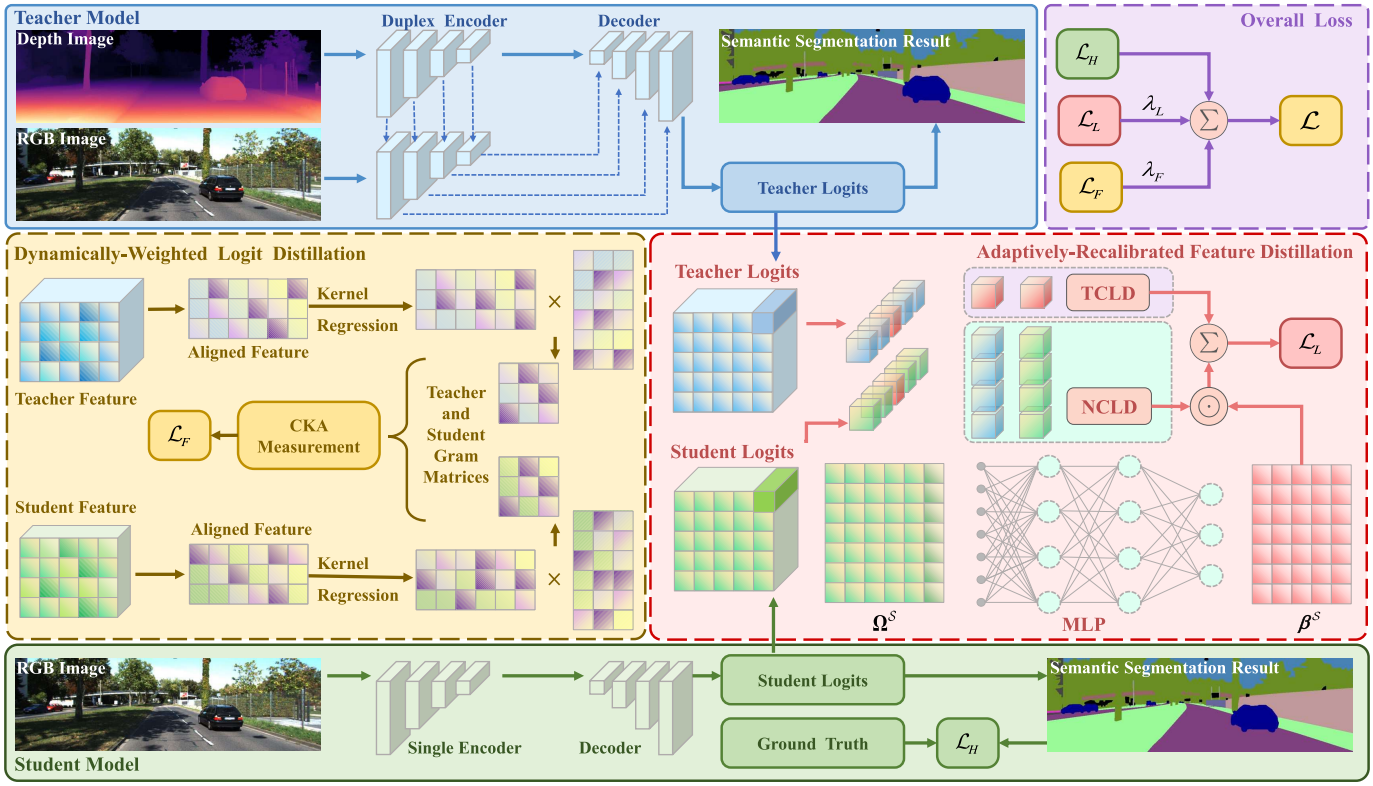

Fig. 1. An illustration of our proposed **LIX framework**, which consists of two key components: (a) **dynamically-weighted logit distillation** and (b) **adaptively-recalibrated feature distillation**.

$$+\left(1-\hat{p}_k^{\mathcal{T}}\right)\underbrace{\sum_{i=1,i\neq k}^{C}\hat{p}_{i,\backslash k}^{\mathcal{T}}\log\frac{\hat{p}_{i,\backslash k}^{\mathcal{T}}}{\hat{p}_{i,\backslash k}^{\mathcal{S}}}}_{\mathrm{KL}(\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{T}}\|\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{S}})}. \tag{2}$$

(2) can be rewritten as follows:

$$\mathrm{LD}=\underbrace{\mathrm{KL}\left(\boldsymbol{b}^{\mathcal{T}}\|\boldsymbol{b}^{\mathcal{S}}\right)}_{\mathrm{TCLD}}+\left(1-\hat{p}_k^{\mathcal{T}}\right)\underbrace{\mathrm{KL}\left(\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{T}}\|\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{S}}\right)}_{\mathrm{NCLD}}, \tag{3}$$

where the TCLD term represents the similarity between $\boldsymbol{b}^{\mathcal{T}}$ and $\boldsymbol{b}^{\mathcal{S}}$ (binary probability vectors of teacher and student networks), while the NCLD term denotes the similarity between $\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{T}}$ and $\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{S}}$. Obviously, both TCLD and the weight of NCLD are related to $\hat{p}_k^{\mathcal{T}}$, making them coupled. As $(1-\hat{p}_k^{\mathcal{T}})$ is often much smaller than the weight of the TCLD term (consistently 1), effects of the NCLD term are often suppressed. However, in [25], the authors claimed that the primary contribution of LD comes from the NCLD term, and reformulated (3) as follows:

$$\mathrm{DKD}=\alpha\mathrm{TCLD}+\beta\mathrm{NCLD}. \tag{4}$$

Two independent hyper-parameters α and β are used to balance the TCLD and NCLD terms [25]. Extensive experiments conducted across a variety of datasets demonstrate that setting α to 1, as consistent with (3), and assigning β as a positive integer between 1 and 10 yields improved distillation performance compared to conventional LD (detailed in Sect. IV-C). Nevertheless, upon revisiting (3), it becomes evident that $(1-\hat{p}_k^{\mathcal{T}})$ varies independently at each pixel. Thus, we propose a strategy to dynamically control the weight of the NCLD term across logits.

A logit value $z_k^{\mathcal{S}}$ approaching positive infinity indicates high confidence in the classification of $\boldsymbol{q}$. This implies that ambiguity arises in the student network when $z_k^{\mathcal{S}}$ is not sufficiently high. Differentiating DKD with respect to $z_k^{\mathcal{S}}$ yields the following expression:

$$\begin{aligned}\frac{\partial\mathrm{DKD}}{\partial z_k^{\mathcal{S}}}&=\alpha\frac{\partial\mathrm{TCLD}}{\partial z_k^{\mathcal{S}}}+\beta\frac{\partial\mathrm{NCLD}}{\partial z_k^{\mathcal{S}}}\\&=\alpha\frac{\mathrm{KL}\left(\boldsymbol{b}^{\mathcal{T}}\|\boldsymbol{b}^{\mathcal{S}}\right)}{\partial z_k^{\mathcal{S}}}+\beta\frac{\mathrm{KL}\left(\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{T}}\|\hat{\boldsymbol{p}}_{\backslash k}^{\mathcal{S}}\right)}{\partial z_k^{\mathcal{S}}}\\&=\alpha\partial\left(\hat{p}_k^{\mathcal{T}}\log\frac{\hat{p}_k^{\mathcal{T}}}{\hat{p}_k^{\mathcal{S}}}+\left(1-\hat{p}_k^{\mathcal{T}}\right)\log\frac{1-\hat{p}_k^{\mathcal{T}}}{1-\hat{p}_k^{\mathcal{S}}}\right)/\partial z_k^{\mathcal{S}}\\&=\alpha\left(-\frac{\hat{p}_k^{\mathcal{T}}}{\hat{p}_k^{\mathcal{S}}}\frac{\partial\hat{p}_k^{\mathcal{S}}}{\partial z_k^{\mathcal{S}}}+\frac{\left(1-\hat{p}_k^{\mathcal{T}}\right)}{\left(1-\hat{p}_k^{\mathcal{S}}\right)}\frac{\partial\hat{p}_k^{\mathcal{S}}}{\partial z_k^{\mathcal{S}}}\right)+0,\end{aligned} \tag{5}$$

which reveals that the gradient of the DKD loss with respect to $z_k^{\mathcal{S}}$ is only related to TCLD, and NCLD contributes zero gradients to the logit optimization. However, as discussed earlier, it is essential to pay more attention to NCLD, especially when the student network is less confident. Therefore, we are motivated to introduce a dynamic weight controller built upon the confidence of student's logits to dynamically balance TCLD and NCLD.

*2) **Dynamic Weight Controller**:* A thorough search of the relevant literature reveals no discussions on the adaptive control of the NCLD weight. We were inspired by a recent study [46] that discussed setting adaptive temperature for KD in graph neural networks. However, this approach uniformly affects all logits, failing to account for the varying levels of confidence across different classes and instances. As depicted

in Fig. 1, we assign each logit with an adaptive NCLD weight based on both the probability vector $\hat{\boldsymbol{p}}$ and the confidence $c = -\hat{\boldsymbol{p}}^{\top} \log \hat{\boldsymbol{p}}$ at the given pixel $\boldsymbol{q}$. Expanding to the entire logit tensor $\boldsymbol{Z}^{\mathcal{S}} \in \mathbb{R}^{C\times H\times W}$ generated by the student network, we first compute the confidence vector $\boldsymbol{c}^{\mathcal{S}} \in \mathbb{R}^{HW}$ for all pixels using the following expression:

$$\boldsymbol{c}^{\mathcal{S}} = -\left(\hat{\boldsymbol{P}}^{\mathcal{S}} \odot \log \hat{\boldsymbol{P}}^{\mathcal{S}}\right) \mathbf{1}_C, \tag{6}$$

where $\odot$ represents element-wise dot product between two matrices, $\mathbf{1}_C$ is a C-entry column vector of ones, and $\hat{\boldsymbol{P}}^{\mathcal{S}} \in \mathbb{R}^{HW\times C}$ is a matrix calculated by applying the softmax function to $\boldsymbol{Z}^{\mathcal{S}}$ and reshaping it into a two-dimensional matrix.

Subsequently, we assign logits using an adaptive weight matrix $\boldsymbol{\beta}^{\mathcal{S}} \in \mathbb{R}^{HW\times C}$, which is constructed based on $\hat{\boldsymbol{P}}^{\mathcal{S}}$ and $\boldsymbol{c}^{\mathcal{S}}$. $\boldsymbol{\beta}^{\mathcal{S}}$ is subsequently obtained as follows:

$$\boldsymbol{\beta}^{\mathcal{S}} = (\beta_{\max} - \beta_{\min}) \operatorname{Sigmoid}\left(\operatorname{MLP}(\boldsymbol{\Omega}^{\mathcal{S}})\right) + \beta_{\min} \mathbf{1}_{HW} \mathbf{1}_C{}^{\top}, \tag{7}$$

where $\boldsymbol{\beta}^{\mathcal{S}}$ stores the NCLD weights with fixed range $[\beta_{\min}, \beta_{\max}]$. We empirically set $\beta_{\min}$ to 1 and $\beta_{\max}$ to 10, the selection of which has been discussed in DKD [25]. The comparison experiments for the fixed $\boldsymbol{\beta}^{\mathcal{S}}$ in this interval are detailed in Sect. IV-C. $\boldsymbol{\Omega}^{\mathcal{S}}$ in (7) is obtained as follows:

$$\boldsymbol{\Omega}^{\mathcal{S}} = \operatorname{Concat}\left(\hat{\boldsymbol{P}}^{\mathcal{S}}, \boldsymbol{c}^{\mathcal{S}}\right), \tag{8}$$

where "Concat" represents the operation to concatenate the matrix $\hat{\boldsymbol{P}}^{\mathcal{S}} \in \mathbb{R}^{HW\times C}$ and the confidence vector $\boldsymbol{c}^{\mathcal{S}} \in \mathbb{R}^{HW}$ to obtain $\boldsymbol{\Omega}^{\mathcal{S}} \in \mathbb{R}^{HW\times(C+1)}$. Through extensive experiments detailed in Sect. IV-C, we first validate the effectiveness of our proposed DWC, when $\boldsymbol{\Omega}^{\mathcal{S}} = \hat{\boldsymbol{P}}^{\mathcal{S}}$. Moreover, it achieves improved performance, when further incorporating $\boldsymbol{c}^{\mathcal{S}}$ into $\boldsymbol{\Omega}^{\mathcal{S}}$. Therefore, extending (4) to the entire image results in the logit distillation loss as follows:

$$\begin{aligned}\mathcal{L}_L = {} & \alpha \underbrace{\mathbf{1}_{(HW)}^{\top}\left(\boldsymbol{B}^{\mathcal{S}} \odot \log\left(\boldsymbol{B}^{\mathcal{T}} \oslash \boldsymbol{B}^{\mathcal{S}}\right)\right)\mathbf{1}_2}_{\text{TCLD}} \\ & + \underbrace{\mathbf{1}_{(HW)}^{\top}\left(\boldsymbol{\beta}^{\mathcal{S}} \odot \hat{\boldsymbol{P}}^{\mathcal{S}}_{\backslash k} \odot \log\left(\hat{\boldsymbol{P}}^{\mathcal{T}}_{\backslash k} \oslash \hat{\boldsymbol{P}}^{\mathcal{S}}_{\backslash k}\right)\right)\mathbf{1}_C}_{\text{logit-wise-weighted NCLD}},\end{aligned} \tag{9}$$

where $\oslash$ denotes element-wise division. Let $\boldsymbol{T}^{\mathcal{S}} \in \mathbb{R}^{HW\times C}$ and $\boldsymbol{N}^{\mathcal{S}} = \mathbf{1}_{HW}\mathbf{1}_C^{\top} - \boldsymbol{T}^{\mathcal{S}} \in \mathbb{R}^{HW\times C}$ be the target and non-target ground-truth matrices in a boolean format, where each row is a binary vector corresponding to the given segmentation ground truth. Then, $\boldsymbol{B}^{\mathcal{T},\mathcal{S}} \in \mathbb{R}^{HW\times 2}$, the matrices that store the binary probabilities at each pixel, are obtained as follows:

$$\boldsymbol{B}^{\mathcal{T},\mathcal{S}} = \operatorname{Concat}\left(\hat{\boldsymbol{P}}^{\mathcal{T},\mathcal{S}} \odot \boldsymbol{T}^{\mathcal{T},\mathcal{S}}\mathbf{1}_C, \hat{\boldsymbol{P}}^{\mathcal{T},\mathcal{S}} \odot \boldsymbol{N}^{\mathcal{T},\mathcal{S}}\mathbf{1}_C\right), \tag{10}$$

where

$$\hat{\boldsymbol{P}}^{\mathcal{T},\mathcal{S}}_{\backslash k} = \frac{\exp\left(\operatorname{Reshape}\left(\boldsymbol{Z}^{\mathcal{T},\mathcal{S}}\right) - \delta \boldsymbol{T}^{\mathcal{T},\mathcal{S}}\right)}{\exp\left(\operatorname{Reshape}\left(\boldsymbol{Z}^{\mathcal{T},\mathcal{S}}\right) - \delta \boldsymbol{T}^{\mathcal{T},\mathcal{S}}\right)\mathbf{1}_C\mathbf{1}_C{}^{\top}} \tag{11}$$

is a matrix of size $\mathbb{R}^{HW\times C}$ storing the independently modeled probabilities among non-target classes at each pixel, "Reshape" operation denotes expanding a three-dimensional tensor into a two-dimensional matrix, and δ is a large value approaching infinity. Compared with (4), the inclusion of the variable $\boldsymbol{\beta}^{\mathcal{S}}$ in (9) allows each non-target class logit in the output to receive an adaptive weight that adjusts during the training process, thereby enhancing the overall performance of LD. The comprehensive quantitative comparison between DKD [25] and our proposed DWLD is provided in Sect. IV-C.

C. Adaptively-Recalibrated Feature Distillation

The above-mentioned LD method relies exclusively on the output of the last layer, overlooking the significance of intermediate features within both teacher and student networks. It has nevertheless been demonstrated that these features play a critical role in effective representation learning, especially in deep neural networks [42]. Therefore, we design an adaptively-recalibrated feature distillation approach to further boost the transfer of spatial geometric prior knowledge from the teacher network to the student network. The remainder of this subsection details a feature recalibration process using kernel regression and a comprehensive feature consistency measurement method leveraging CKA [26].

*1) **Feature Recalibration via Kernel Regression**:* Let the feature maps produced by teacher and student networks be the tensors $\boldsymbol{F}^{\mathcal{T}} \in \mathbb{R}^{C^{\mathcal{T}}\times H^{\mathcal{T}}\times W^{\mathcal{T}}}$ and $\boldsymbol{F}^{\mathcal{S}} \in \mathbb{R}^{C^{\mathcal{S}}\times H^{\mathcal{S}}\times W^{\mathcal{S}}}$, respectively. Following the previous studies [47], [48], [49], we first align tensors $\boldsymbol{F}^{\mathcal{T}}$ and $\boldsymbol{F}^{\mathcal{S}}$ and produce matrix $\boldsymbol{F}_a^{\mathcal{T},\mathcal{S}}$ through the following process:

$$\boldsymbol{F}_a^{\mathcal{T},\mathcal{S}} = \operatorname{ReLU}\left(\operatorname{Norm}\left(\underset{3\times 3}{\operatorname{Conv}}\left(\operatorname{Reshape}(\boldsymbol{F}^{\mathcal{T},\mathcal{S}})\right)\right)\right), \tag{12}$$

where $\boldsymbol{F}_a^{\mathcal{T},\mathcal{S}} \in \mathbb{R}^{C^{\mathcal{T}}\times HW}$, "$\operatorname{Conv}_{3\times 3}$" denotes a convolutional layer built upon 3×3 kernels, and $HW = \min H^{\mathcal{T},\mathcal{S}} \min W^{\mathcal{T},\mathcal{S}}$. We take into account both spatial and channel alignment. As for features with mismatched resolutions, the "Reshape" operation adjusts the larger feature map to match the smaller one. Given the specific nature of our task, where feature maps from data-fusion and single-modal networks differ significantly in the channel dimension, the convolutional layer aligns the channels between the teacher and student features. Simultaneously, the activated and normalized features from the teacher and student networks share a more similar feature representation space.

Before delving into feature distillation, a pertinent question arises: are the features, having undergone alignment to a common shape using (12), ready for utilization? Extensive quantitative and qualitative experimental results lead us to a negative conclusion. We attribute the unsatisfactory performance of feature distillation to the discrepancy in feature scales and distributions between the teacher and student feature maps. On one hand, features often exhibit variations in different orders of magnitude, making their direct comparisons difficult [16], [17]. Working directly with these aligned features $\boldsymbol{F}_a^{\mathcal{T}}$ and $\boldsymbol{F}_a^{\mathcal{S}}$ could result in consistency measures being dominated by the model producing features with larger magnitudes. On the other hand, the absolute density of the data may vary significantly, and the shape of feature clusters may also differ depending on the locality [24]. Thus, recalibrating feature map distributions becomes another critical aspect that

demands further attention. In particular, in our research problem, the teacher network learns heterogeneous features from both RGB images and spatial geometric information, while the student network is exclusively exposed to RGB images. Such a substantial difference in the learning data format exacerbates the knowledge transfer challenges. Taking inspiration from the knowledge transfer algorithm introduced in the study [15], which utilizes kernel regression to minimize the maximum mean discrepancy between teacher and student feature map probability distributions, we apply this technique to recalibrate the aligned features in our specific problem. The Laplace-based kernel regression process can be formulated as follows:

$$\boldsymbol{F}_k^{\mathcal{T},\mathcal{S}} = \mathrm{EXP}\left(-\frac{(\boldsymbol{F}_a^{\mathcal{T},\mathcal{S}} - \bar{\boldsymbol{F}}_a^{\mathcal{T},\mathcal{S}}) \odot (\boldsymbol{F}_a^{\mathcal{T},\mathcal{S}} - \bar{\boldsymbol{F}}_a^{\mathcal{T},\mathcal{S}})}{\sigma}\right), \quad (13)$$

where σ is the standard deviation of the squared distances between $\boldsymbol{F}_a^{\mathcal{T},\mathcal{S}}$ and $\bar{\boldsymbol{F}}_a^{\mathcal{T},\mathcal{S}}$, "EXP" represents the operation to take the exponent for each element in the tensor, and

$$\bar{\boldsymbol{F}}_a^{\mathcal{T},\mathcal{S}} = \frac{1}{H \times W \times C^{\mathcal{T}}} \mathbf{1}_{C^{\mathcal{T}}}^{\top} \boldsymbol{F}_a^{\mathcal{T},\mathcal{S}} \mathbf{1}_{HW}. \quad (14)$$

The comprehensive comparisons of different kernel regression methods are given in Sect. IV-D.

*2) **Feature Consistency Measurement**:* As mentioned above, measuring the consistency (or similarity) between intermediate features in teacher and student networks is the key to feature distillation. Although Euclidean distance, cosine similarity, and the Pearson correlation coefficient can all serve this purpose, it has been witnessed that CKA [26] based on the HSIC [27] provides a more comprehensive quantification of feature consistency, as demonstrated in several fundamental machine learning studies [26], [50], [51].

We begin by computing two Gram matrices $\boldsymbol{T}_k = \boldsymbol{F}_k^{\mathcal{T}} \boldsymbol{F}_k^{\mathcal{T}\top} \in \mathbb{R}^{C^{\mathcal{T}} \times C^{\mathcal{T}}}$ and $\boldsymbol{S}_k = \boldsymbol{F}_k^{\mathcal{S}} \boldsymbol{F}_k^{\mathcal{S}\top} \in \mathbb{R}^{C^{\mathcal{T}} \times C^{\mathcal{T}}}$, reflecting the similarities between pairs of examples based on the representations contained in $\boldsymbol{F}_k^{\mathcal{T}}$ and $\boldsymbol{F}_k^{\mathcal{S}}$ [26]. Their HSIC measure is subsequently computed as follows [27]:

$$\mathrm{HSIC}(\boldsymbol{T}_k, \boldsymbol{S}_k) = \frac{C^{\mathcal{T}2}}{(C^{\mathcal{T}}-1)^2}\Bigg(\mathrm{tr}(\boldsymbol{S}_k \boldsymbol{T}_k) + \frac{\mathbf{1}_{C^{\mathcal{T}}}^{\top} \boldsymbol{S} \mathbf{1}_{C^{\mathcal{T}}} \mathbf{1}_{C^{\mathcal{T}}}^{\top} \boldsymbol{T}_k \mathbf{1}_{C^{\mathcal{T}}}}{C^{\mathcal{T}2}} - \frac{2}{C^{\mathcal{T}2}} \mathbf{1}_{C^{\mathcal{T}}}^{\top} \boldsymbol{S}_k \boldsymbol{T}_k \mathbf{1}_{C^{\mathcal{T}}}\Bigg), \quad (15)$$

which statistically quantifies dependencies between $\boldsymbol{T}_k$ and $\boldsymbol{S}_k$. CKA further normalizes the HSIC measures using the following expression:

$$\mathrm{CKA}(\boldsymbol{T}_k, \boldsymbol{S}_k) = \frac{\mathrm{HSIC}(\boldsymbol{T}_k, \boldsymbol{S}_k)}{\sqrt{\mathrm{HSIC}(\boldsymbol{T}_k, \boldsymbol{T}_k)\mathrm{HSIC}(\boldsymbol{S}_k, \boldsymbol{S}_k)}}. \quad (16)$$

A CKA measure approaching 1 indicates that the intermediate features in teacher and student networks tend to be consistent. In our task, this suggests that the spatial geometric prior knowledge learned by the data-fusion teacher networks at the feature level is likely to have been implicitly infused into the single-modal student network. Therefore, the feature distillation loss can be formulated as follows:

$$\mathcal{L}_F = \sum_{n=1}^{N} (1 - \mathrm{CKA}(\boldsymbol{T}_{k,n}, \boldsymbol{S}_{k,n})), \quad (17)$$

where $\boldsymbol{T}_{k,n}$ and $\boldsymbol{S}_{k,n}$ are yielded using the n-th pair of recalibrated teacher and student feature maps, respectively, and N denotes the total number of feature maps. The quantitative comparisons of the above-mentioned feature consistency measurement methods are provided in Sect. IV-D.

D. Overall Loss

As depicted in Fig. 1, our proposed LIX framework distills an RGB-X data-fusion teacher network into a student network trained exclusively with RGB images by two critical KD strategies: DWLD and ARFD. DWLD first reformulates the conventional KD loss into two independent and decoupled terms: TCLD and NCLD. Subsequently, a dynamic weight controller utilizing MLP layers generates a weight for each logit adaptively. Therefore, TCLD and weighted NCLD terms together form the LD loss $\mathcal{L}_L$. As for FD, we first utilize adaptive feature alignment and recalibration approaches based on kernel regression, which recalibrate the feature map distributions of teacher and student networks across various dimensions. Subsequently, we measure the similarity between features to formulate the FD loss $\mathcal{L}_F$ in teacher and student networks based on the CKA algorithm, which quantifies the feature consistency between teacher and student networks.

In summary, the overall loss function can be formulated as a combination of the initial loss $\mathcal{L}_H$ of the hard labels (ground truth) and distillation losses, consisting of a logit distillation loss $\mathcal{L}_L$ and a feature distillation loss $\mathcal{L}_F$, as follows:

$$\mathcal{L} = \mathcal{L}_H + \lambda_L \mathcal{L}_L + \lambda_F \mathcal{L}_F, \quad (18)$$

where λ_L and λ_F are hyper-parameters to balance distillation losses. By minimizing (18), the single-model student network can be implicitly infused with the spatial geometric knowledge learned by the data-fusion teacher network.

IV. Experiments

A. Experimental Setup

*1) **Datasets**:* We evaluate the performance of KD methods on three public datasets: the vKITTI2 dataset [54] (synthetic yet large-scale), the KITTI Semantics dataset [52] (real-world yet modest-scale), and the nuImage dataset [53] (real-world and large-scale). Their details are as follows:

- The **vKITTI2 dataset** contains virtual replicas of five sequences from the KITTI dataset and provides semantic annotations for 15 different classes. Dense ground-truth depth maps are acquired through depth rendering using a virtual engine. In our experiments, we randomly select 700 images from this dataset, along with their semantic and depth annotations, to validate the effectiveness of the proposed approaches. These images are randomly divided into a training set and a validation set with a ratio of 5:2.
- The **KITTI Semantics dataset** contains 200 real-world images captured in various driving scenarios. It provides ground-truth semantic annotations for 19 different classes (in alignment with the Cityscapes [55] dataset). Sparse disparity ground truth is obtained using a Velodyne HDL-64E LiDAR. We generate dense depth maps using a well-trained ViTAStereo [56] network in the experiments.

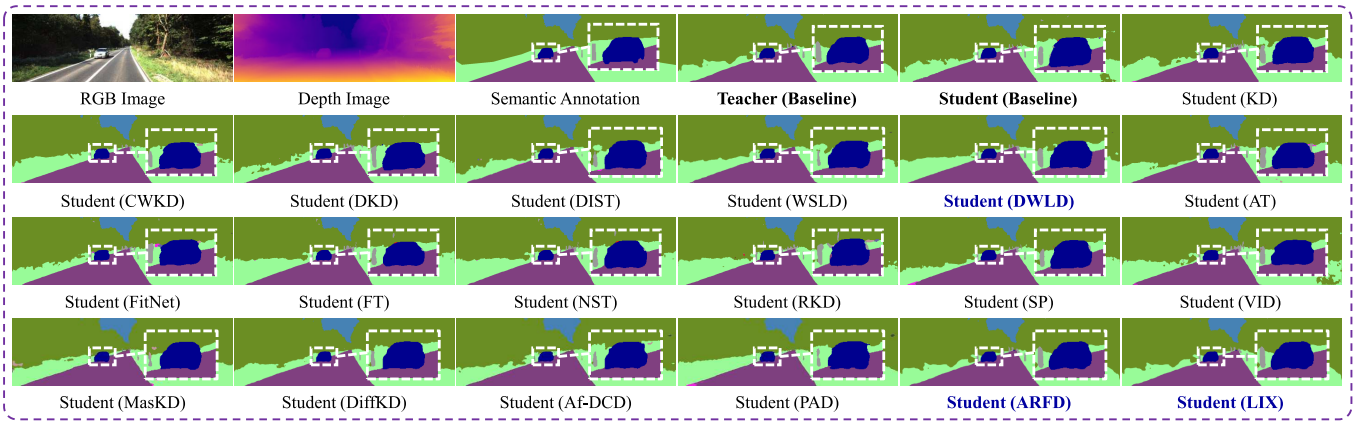

Fig. 2. Qualitative comparison with other SoTA KD approaches on the KITTI Semantics dataset [52] using SNE-RoadSeg, where significantly improved areas are highlighted with white dashed boxes.

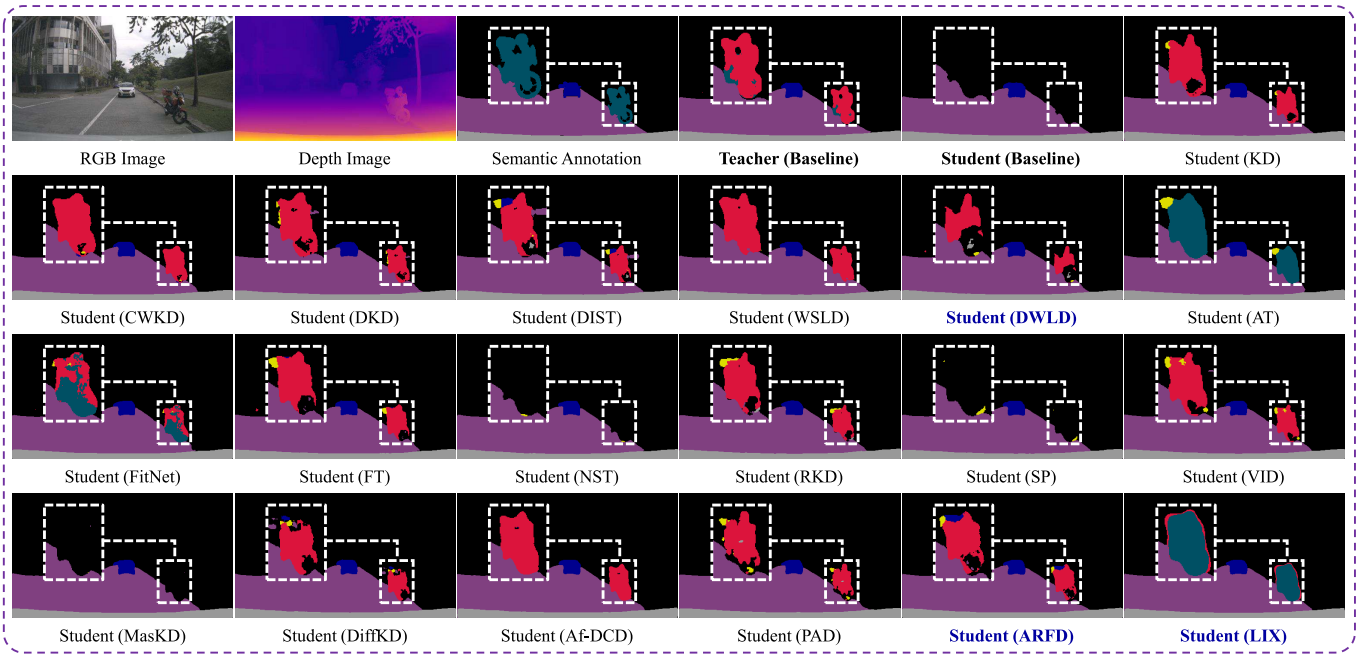

Fig. 3. Qualitative comparison with other SoTA KD approaches on the nuImage dataset [53] using SNE-RoadSeg, where significantly improved areas are highlighted with white dashed boxes.

These images are randomly divided into a training set and a validation set with a ratio of 3:1.

- The **nuImage dataset** is a large-scale, real-world dataset designed for autonomous driving perception. It consists of 93,476 images with 2M labeled objects. In our experiments, we randomly select 800 images with semantic annotations and generate dense depth maps using a pre-trained Depth Anything network [57]. These images are then randomly split into a training set and a validation set with a ratio of 7:3.

*2) **Implementation Details and Evaluation Metrics**:* Our experiments are conducted on an NVIDIA RTX 3090 GPU. All images in the vKITTI2 dataset and the KITTI Semantics dataset are resized to $1,248 \times 384$ pixels, and images in the nuImage dataset are resized to 512×288 pixels. We utilize the stochastic gradient descent [59] optimizer for network training, with momentum and weight decay parameters set to 0.9 and 5×10^{-4}, respectively. The initial learning rate is set to 5×10^{-3}, and training is conducted for 500 epochs with early stopping used to prevent over-fitting. The batch size is set to 3 for SNE-RoadSeg on the vKITTI2 and KITTI Semantics datasets, 8 for MFNet on the vKITTI2 and KITTI Semantics datasets, and 8 for both networks on the nuImage dataset. To enhance the model's robustness, we employ standard data augmentation techniques, such as random color adjustment, photometric distortion, rescaling, and flipping. We employ four metrics to quantify the performance of KD algorithms: the mean and frequency-weighted F1-score (abbreviated as mFsc and fwFsc, respectively) as well as the mean and frequency-weighted intersection over union (abbreviated as mIoU and fwIoU, respectively).

TABLE I

COMPARISON WITH SOTA KNOWLEDGE DISTILLATION APPROACHES ON THE VKITTI2 DATASET

KD Type	Algorithm	SNE-RoadSeg				MFNet			
		mFsc (%) ↑	fwFsc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑	mFsc (%) ↑	fwFsc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑
/	Teacher (Baseline w/o KD)	97.72	99.12	95.63	98.27	92.65	97.68	86.96	95.57
	Student (Baseline w/o KD)	94.51	97.83	89.91	95.82	91.22	96.93	84.59	94.19
LD	KD [12]	94.89	97.90	90.59	95.94	91.93	97.06	85.70	94.42
	CWKD [11]	95.02	97.95	90.82	96.03	91.63	97.02	85.29	94.36
	DKD [25]	94.97	97.94	90.75	95.97	91.68	97.06	85.29	94.42
	DIST [40]	95.17	97.97	91.08	96.06	84.21	95.05	74.58	91.01
	WSLD [39]	94.86	97.95	90.52	96.04	85.85	95.38	76.69	91.53
	DWLD (Ours)	**95.42**	**98.06**	**91.50**	**96.24**	**92.29**	**97.22**	**85.94**	**94.68**
FD	AT [13]	94.61	97.87	90.12	95.89	91.97	97.13	85.78	94.55
	FitNet [42]	94.89	97.92	90.59	95.99	91.36	96.93	84.83	94.19
	FT [14]	94.75	97.91	90.34	95.97	91.47	96.96	84.96	94.25
	NST [15]	83.19	90.94	73.27	84.55	89.39	96.58	81.88	93.58
	RKD [16]	94.69	97.81	90.25	95.78	85.33	95.44	76.19	91.66
	SP [17]	95.08	97.96	90.93	96.06	91.98	97.10	85.78	94.50
	VID [18]	95.05	97.95	90.87	96.03	91.78	97.01	85.54	94.34
	MasKD [43]	81.26	92.07	70.95	86.34	80.92	94.08	70.45	89.44
	DiffKD [58]	94.87	97.92	90.54	95.98	84.67	94.93	75.17	90.75
	Af-DCD [44]	93.81	97.72	88.76	95.62	80.34	94.21	70.00	89.76
	PAD [45]	95.11	97.97	90.96	96.07	85.63	95.07	76.30	90.96
	ARFD (Ours)	**95.20**	**98.01**	**90.97**	**96.16**	**92.46**	**97.27**	**86.52**	**94.80**
LD+FD	**LIX (Ours)**	**95.79**	**98.23**	**91.95**	**96.33**	**92.50**	**97.32**	**86.88**	**95.10**

TABLE II

COMPARISON WITH SOTA KNOWLEDGE DISTILLATION APPROACHES ON THE KITTI SEMANTICS DATASET

KD Type	Algorithm	SNE-RoadSeg				MFNet			
		mFsc (%) ↑	fwFsc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑	mFsc (%) ↑	fwFsc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑
/	Teacher (Baseline w/o KD)	70.85	93.02	61.34	87.76	45.17	89.22	37.86	82.37
	Student (Baseline w/o KD)	59.43	90.48	49.17	83.86	35.91	86.52	30.68	79.19
LD	KD [12]	**63.71**	90.88	52.75	84.39	37.38	87.14	31.79	80.01
	CWKD [11]	60.84	90.49	50.15	83.88	38.04	87.14	32.25	79.88
	DKD [25]	61.61	90.57	51.28	83.90	37.62	86.35	31.61	78.65
	DIST [40]	55.26	91.71	47.71	85.88	37.94	86.74	32.05	79.31
	WSLD [39]	53.24	91.16	45.66	85.09	40.57	87.77	33.91	80.49
	DWLD (Ours)	62.78	**92.36**	**53.54**	**86.62**	39.82	**88.41**	**33.93**	**81.62**
FD	AT [13]	61.81	91.24	51.57	84.98	37.02	86.36	31.31	78.80
	FitNet [42]	57.88	90.46	48.18	83.82	38.70	87.29	32.88	80.02
	FT [14]	60.61	90.41	50.16	83.71	38.24	86.89	32.37	79.52
	NST [15]	55.41	87.77	45.56	79.93	40.53	**88.08**	34.05	**81.06**
	RKD [16]	58.52	89.55	47.74	82.50	33.11	84.26	28.10	76.21
	SP [17]	**62.89**	90.52	52.11	83.91	38.35	86.92	32.49	79.51
	VID [18]	62.45	90.51	51.66	83.89	38.16	87.07	32.32	79.78
	MasKD [43]	51.41	88.58	42.32	81.56	40.85	87.27	34.10	79.90
	DiffKD [58]	55.94	91.97	48.37	86.29	**41.48**	87.73	34.45	80.29
	Af-DCD [44]	51.04	91.65	43.60	85.81	41.02	86.75	33.74	78.82
	PAD [45]	48.65	90.84	41.09	84.65	36.99	86.70	31.65	79.40
	ARFD (Ours)	62.25	**92.34**	**52.92**	**86.59**	41.32	87.76	**34.46**	80.56
LD+FD	**LIX (Ours)**	**63.70**	**92.47**	**54.65**	**86.80**	**43.25**	**88.59**	**36.30**	**81.63**

B. Comparison With State-of-The-Art Methods

The quantitative and qualitative results on the vKITTI2, KITTI Semantics, and nuImage datasets are presented in Tables I, II, and III as well as Figs. 2 and 3. In our baseline experiments, the teacher network utilizes a duplex encoder for RGB-Depth (RGB-D) semantic segmentation, while the student network uses a single encoder to learn semantic clues exclusively from RGB images. These results suggest that our proposed LIX framework enables the student network to retain over 90% of the teacher network's performance in terms of mFsc, and in some cases, achieves nearly comparable performance. Moreover, when DWLD is utilized solely in this specific task, we observe significant improvements compared to the baseline student network. These results provide strong evidence for the effectiveness and superiority of our proposed DWLD algorithm. A similar conclusion is reached for ARFD, which also yields significant performance gains. LIX, the combined use of these two algorithms, enables the single-

TABLE III
COMPARISON WITH SOTA KNOWLEDGE DISTILLATION APPROACHES ON THE NUIMAGE DATASET

KD Type	Algorithm	SNE-RoadSeg				MFNet			
		mFsc (%) ↑	fwFsc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑	mFsc (%) ↑	fwFsc (%) ↑	mIoU (%) ↑	fwIoU (%) ↑
/	Teacher (Baseline w/o KD)	74.42	96.83	65.80	94.16	59.98	94.59	54.06	90.40
	Student (Baseline w/o KD)	60.83	95.30	55.27	91.66	55.97	92.51	49.03	87.03
LD	KD [12]	68.20	95.61	60.46	92.07	56.11	92.98	49.35	87.79
	CWKD [11]	67.17	95.62	59.48	92.11	56.49	93.01	49.72	87.84
	DKD [25]	67.66	95.60	60.06	92.04	56.38	92.92	49.59	87.72
	DIST [40]	67.11	95.56	59.50	91.99	56.38	92.97	49.61	87.77
	WSLD [39]	67.16	95.53	59.46	91.96	57.52	93.18	50.00	88.12
	DWLD (Ours)	**68.70**	**95.83**	**61.11**	**92.45**	**57.55**	**93.21**	**50.67**	**88.15**
FD	AT [13]	66.06	95.42	58.69	91.79	57.02	93.16	50.15	**88.09**
	FitNet [42]	**70.54**	95.66	61.39	92.11	56.28	92.92	49.46	87.71
	FT [14]	67.21	95.76	59.52	92.36	56.33	92.99	49.56	87.80
	NST [15]	62.24	**95.88**	56.21	**92.68**	**57.39**	**93.26**	50.06	87.26
	RKD [16]	64.87	95.02	57.07	91.11	53.37	91.59	46.45	85.62
	SP [17]	61.45	95.41	55.90	91.83	56.27	92.93	49.45	87.70
	VID [18]	68.29	95.65	60.64	92.13	56.38	92.90	49.45	87.65
	MasKD [43]	62.48	95.68	57.08	92.32	57.39	92.26	50.05	87.22
	DiffKD [58]	66.73	95.32	58.92	91.59	56.78	93.02	50.07	87.85
	Af-DCD [44]	66.14	94.99	58.18	91.01	57.53	93.31	50.08	87.34
	PAD [45]	66.42	95.21	58.52	91.38	56.87	92.85	49.94	87.57
	ARFD (Ours)	69.49	**95.88**	**61.84**	92.52	56.94	93.04	**50.26**	87.89
LD+FD	**LIX (Ours)**	**72.90**	**96.06**	**63.54**	**92.79**	**57.91**	**93.34**	**51.06**	**88.38**

modal student network to achieve comparable performance to that of the data-fusion teacher network. This demonstrates the effectiveness of our proposed KD strategy for the implicit infusion of spatial geometric prior knowledge.

Moreover, we observe that KD's performance is significantly influenced by the network parameters as well as the difficulty level of the dataset. The encoders of the teacher and student SNE-RoadSeg networks have 116.28 M and 58.14 M parameters, respectively, while the encoders of the teacher and student MFNet networks have 0.60 M and 0.52 M parameters, respectively. As illustrated in Table I, since the vKITTI2 dataset is larger and less challenging, DWLD, ARFD, and LIX all achieve the SoTA performance regardless of the network parameters. When our method is applied to MFNet, its performance is on par with that of the teacher network on the vKITTI2 dataset, demonstrating superior effectiveness compared to SNE-RoadSeg with LIX. We attribute this superior performance not only to the comparable parameter numbers between the teacher and student MFNet networks but also to the lower difficulty level of the dataset. It is possible that the lightweight MFNet is adequately capable of learning semantic segmentation on this less challenging dataset.

On the other hand, the experimental results on the KITTI Semantics dataset have exceeded our expectations, when using MFNet in conjunction with our method. As illustrated in Table II, ARFD outperforms other algorithms only in mIoU and achieves performance comparable to that of the SoTA algorithms when evaluated using other metrics. We attribute this unexpected outcome to a potential "mismatch" between MFNet and the KITTI Semantics dataset, where a lightweight network may struggle to handle a challenging dataset. Additionally, while the student SNE-RoadSeg network trained via LIX achieves superior performance, its mIoU reaches only 89% of the teacher network's mIoU. This suggests that infusing spatial geometric priors into the student network is generally effective but not universally reliable. In particular, when the teacher and student networks differ significantly in parameter numbers, the student network may struggle to maintain performance on more challenging datasets.

As depicted in Table III, the experimental results on the nuImage dataset highlight LIX's adaptability to real-world scenarios with rich contextual diversity. Notably, LIX demonstrates superior performance with a 12% increase in mFsc over the baseline student network. This success can be attributed to the large-scale and diverse nature of the nuImage dataset, which provides sufficient samples for the student SNE-RoadSeg network to effectively absorb the teacher's spatial geometric prior knowledge. Nevertheless, as shown in Fig. 3, only student networks trained via LIX and AT [13] can recognize the motorcycle, while the remaining networks misclassify it as a person. This failure may stem from the teacher network that misleads the student network during the distillation process. Notably, LIX successfully achieves accurate segmentation results, demonstrating the collaborative effect of combining logit and feature distillation techniques in a challenging large-scale dataset. Additionally, the improvements achieved by KD methods using MFNet are relatively modest. With fewer parameters to handle extensive data, the teacher's knowledge is inherently limited, leaving less room for the student to improve.

While existing KD methods can be effectively applied to solve this problem, their performance generally falls short of our newly proposed LIX framework, which is specifically developed for this task. Among LD methods, classical KD [12] and CWKD [11] demonstrate moderate improvements over the student baseline but lag behind more advanced approaches

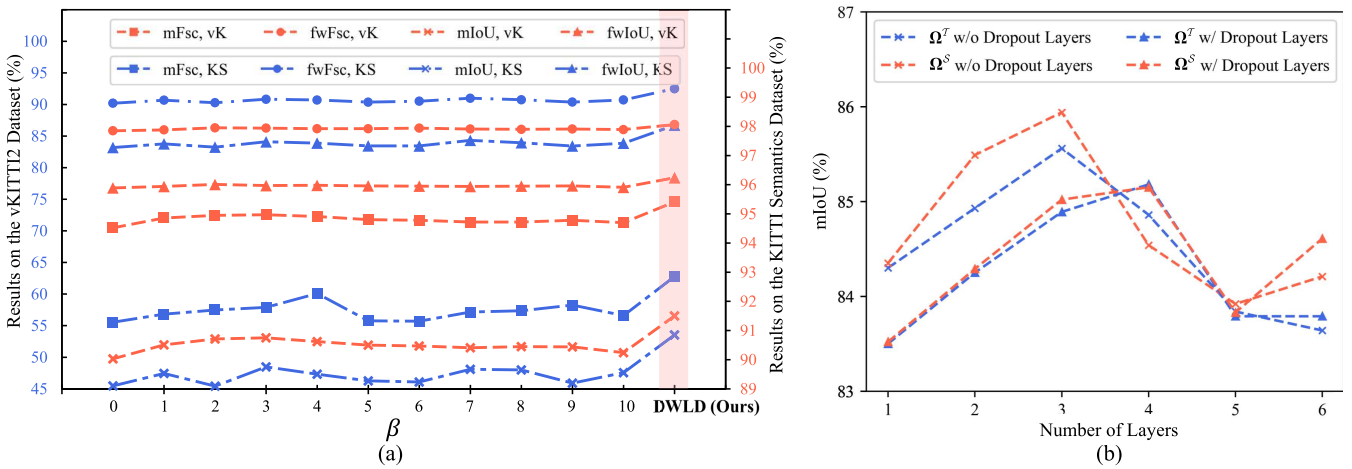

Fig. 4. Ablation studies on LD: (a) comparison between DKD [25] and DWLD with respect to different β, where "vK" and "KS" are the abbreviations of "vKITTI2" and "KITTI Semantics", respectively; (b) comparison among various designs of $\mathbf{\Omega}^{\mathcal{S}}$ on the vKITTI2 dataset.

TABLE IV
COMPARISON AMONG VARIOUS DESIGNS OF $\mathbf{\Omega}^{\mathcal{S}}$ ON THE VKITTI2 DATASET

$\mathbf{\Omega}^{\mathcal{S}}$ Type	mFsc (%)	fwFsc (%)	mIoU (%)	fwIoU (%)
$\hat{\boldsymbol{P}}^{\mathcal{S}}$	90.90	96.84	84.12	94.03
$\boldsymbol{c}^{\mathcal{S}}$	91.42	96.93	84.96	94.22
Concat $\left(\hat{\boldsymbol{P}}^{\mathcal{S}}, \boldsymbol{c}^{\mathcal{S}}\right)$	**92.29**	**97.22**	**85.94**	**94.68**

TABLE V
COMPARISONS OF FEATURE RECALIBRATION AND CONSISTENCY MEASUREMENT METHODS IN MFSC ON THE VKITTI2 DATASET

Feature Recalibration \ Feature Consistency	Cosine Similarity	Euclidean Distance	Pearson Correlation Coefficient	CKA
Laplace-Based Kernel Regression	88.69	90.64	89.15	**92.46**
Gaussian-Based Kernel Regression	91.63	91.26	88.93	91.33
Linear-Based Kernel Regression	86.36	90.50	89.68	91.25
w/o Kernel Regression	86.39	89.81	88.51	90.02

[25]. Furthermore, our proposed DWLD focuses on uncertain logits by assigning an appropriate weight to each logit, leading to a 0.02-9.25% higher IoU compared to WSLD, which prioritizes soft label regularization. As for FD methods, AT [13] and FitNet [42] yield only marginal improvements, highlighting the challenge of direct feature alignment in complex scenarios. Methods such as SP [17] and VID [18] achieve better results by preserving pairwise feature similarities. However, their reliance on dimensionality reduction limits their effectiveness in fine-grained scene parsing.

C. Ablation Study on Logit Distillation

We first compare DWLD with the baseline algorithm DKD [25], which requires a manually-set, fixed β. As shown in Fig. 4(a), DWLD consistently demonstrates superior performance over DKD across different values of β. These results suggest that, compared to DKD, which sometimes struggles to find a proper single, fixed weight β, our proposed DWLD offers a preferable option for both achieving better performance and simplifying the deployment process.

Moreover, we validate the effectiveness of DWC both when $\mathbf{\Omega}^{\mathcal{S}}$ is set to $\hat{\boldsymbol{P}}^{\mathcal{S}}$ and when $\boldsymbol{c}^{\mathcal{S}}$ is additionally incorporated into $\mathbf{\Omega}^{\mathcal{S}}$. As depicted in Table IV, when $\mathbf{\Omega}^{\mathcal{S}} = \text{Concat}\left(\hat{\boldsymbol{P}}^{\mathcal{S}}, \boldsymbol{c}^{\mathcal{S}}\right)$, the student network achieves a increase by over 1.0% in mIoU, supporting our claim regarding the design of $\mathbf{\Omega}^{\mathcal{S}}$.

In theory, greater confidence in the teacher network should lead to a better distillation of non-target class information into the student network. While the design of the DWC draws inspiration from differentiating DKD [25] concerning $z_k^{\mathcal{S}}$, there remains an open question whether the confidence of the teacher network also exerts a significant influence on DWC. Therefore, we further validate the effectiveness of DWC using both $\mathbf{\Omega}^{\mathcal{S}}$ and $\mathbf{\Omega}^{\mathcal{T}}$ as inputs, with and without the incorporation of dropout layers, as shown in Fig. 4(b). As anticipated, DWC yields superior performance when utilizing $\mathbf{\Omega}^{\mathcal{S}}$ as input, confirming the fundamental practicability of the core concept underlying our DWC design. Furthermore, Fig. 4(b) also provides readers with the quantitative results on the selection of MLP layers, which is another key aspect of our DWC design. It is evident that as the number of MLP layers increases, the performance of the student network shows a gradual improvement until reaching a saturation point, after which its performance degrades due to over-fitting. Additionally, the inclusion of dropout layers does not appear to enhance the overall performance of DWC. Therefore, our DWC utilizes three MLP layers without dropout.

D. Ablation Study on Feature Distillation

As presented in Table V, remarkable improvements are achieved through feature recalibration via kernel regression. These improvements can primarily be attributed to the effectiveness of kernel regression in reducing the gap between features in teacher and student networks across multiple dimensions. As expected, both Euclidean distance and the Pearson correlation coefficient prove effective for feature

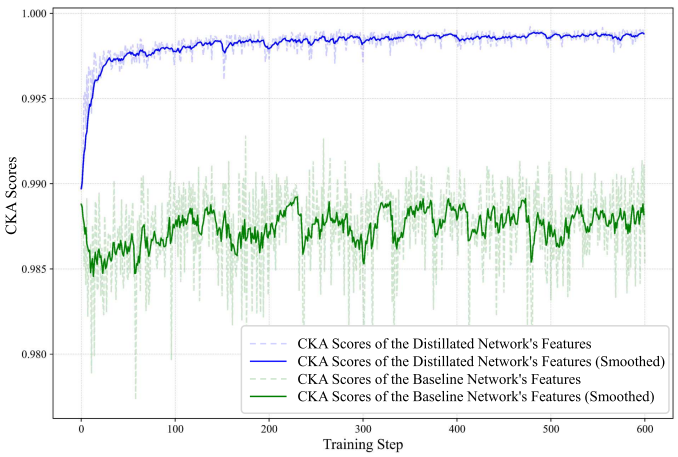

Fig. 5. Comparison of features' CKA scores obtained from baseline and distilled student networks.

consistency measurement. However, when cosine similarity is used in conjunction with a linear kernel, it demonstrates even poorer performance compared to the cases where no kernel regression is employed. We posit that conventional similarity measurement methods, such as cosine similarity, have the opposite effect by focusing on the differences in attribute values. Consequently, these methods are often misled by irrelevant details, such as image backgrounds, while overlooking the essential features. In contrast, our ARFD, which leverages the CKA strategy, offers a more comprehensive quantification of feature consistency. The optimal performance is achieved when combining Laplace-based kernel regression with CKA-based feature consistency measurement.

Furthermore, Fig. 5 provides readers with a quantitative comparison of the CKA scores between the features of the baseline and distilled student networks. As for the distilled student network, the CKA score of the features steadily increases and stabilizes as training progresses. In contrast, the CKA score of the features from the baseline student network exhibits continuous fluctuations throughout the training process. This comparison highlights that, without feature consistency supervision, it is challenging for the single-model student network to achieve stable feature alignment with the teacher network. This further highlights the effectiveness of the ARFD in facilitating consistent knowledge transfer.

We also provide a qualitative comparison between our proposed ARFD and other LD approaches. As depicted in Fig. 6, ARFD generates more confident predictions. Notably, it leads to a more uniform distribution of probabilities for the "traffic sign" class, indicating the effective infusion of spatial geometric prior knowledge through ARFD.

E. Additional Experiments

LD has been widely adopted in dense prediction tasks due to its simplicity and direct supervision on output probabilities [70]. To further demonstrate the general applicability of the proposed ARFD strategy, we extend our experiments beyond semantic segmentation to other computer vision tasks. As presented in Table VI, for the object tracking on the LaSOT [71] dataset, we adopt MixFormerV2-B [60] as the teacher network and replace its original feature distillation with ARFD. As expected, the distilled student network achieves competitive performance compared with other SoTA trackers [61], [64], [67]. Additionally, Tables VII and VIII present quantitative results for image classification and object detection. In both tasks, the student network distilled with ARFD consistently outperforms the baseline student networks, highlighting its broad applicability across diverse computer vision tasks. Further experimental details are provided in the supplementary material.

Furthermore, we recently developed a high-performing vision foundation model (VFM)-based RGB-D semantic segmentation network, referred to as heterogeneous feature integration Transformer (HFIT) [72]. It utilizes a side adapter to extract multi-scale spatial pyramid features from RGB-D pairs and adaptively integrates them with prior features from VFMs. In this subsection, we leverage this network to further validate the compatibility of LIX within a VFM-based

TABLE VI

COMPARISON OF LIX WITH OTHER SOTA OBJECT TRACKERS, WHERE "T" REPRESENTS THE BASELINE TEACHER NETWORK. RESULTS OF THE COMPETING METHODS ARE REPORTED IN THE STUDY [60]

Algorithm	AUC (%)	P_{Norm} (%)	P (%)
MixFormerV2-B (T) [60]	70.6	80.8	76.2
MixFormerV2-S [60]	60.6	69.9	60.4
FEAR-L [61]	57.9	68.6	60.9
FEAR-XS [61]	53.5	64.1	54.5
HCAT [62]	59.0	68.3	60.5
E.T.Track [63]	59.1	-	-
LightTrack-LargeA [64]	55.5	-	56.1
LightTrack-Mobile [64]	53.8	-	53.7
STARK-Lightning [65]	58.6	69.0	57.9
DiMP [66]	56.9	65.0	56.7
SiamFC++ [67]	54.4	62.3	54.7
ARFD (Ours)	55.2	64.3	55.8

TABLE VII

COMPARISON BETWEEN LIX AND OTHER SOTA KD METHODS ON THE IMAGENET [68] DATASET FOR IMAGE CLASSIFICATION

KD Type	Algorithm	Acc (%)
/	Teacher (Baseline w/o KD)	73.6
	Student (Baseline w/o KD)	69.9
LD	KD [12]	71.8
	WSLD [39]	71.5
FD	RKD [16]	70.2
	ARFD (Ours)	72.0

TABLE VIII

COMPARISON BETWEEN LIX AND CWKD [11] ON THE COCO [69] DATASET FOR OBJECT DETECTION

KD Type	Algorithm	mAP (%)
/	Teacher (Baseline w/o KD)	44.7
	Student (Baseline w/o KD)	40.2
FD	CWKD [11]	41.9
	ARFD (Ours)	42.1

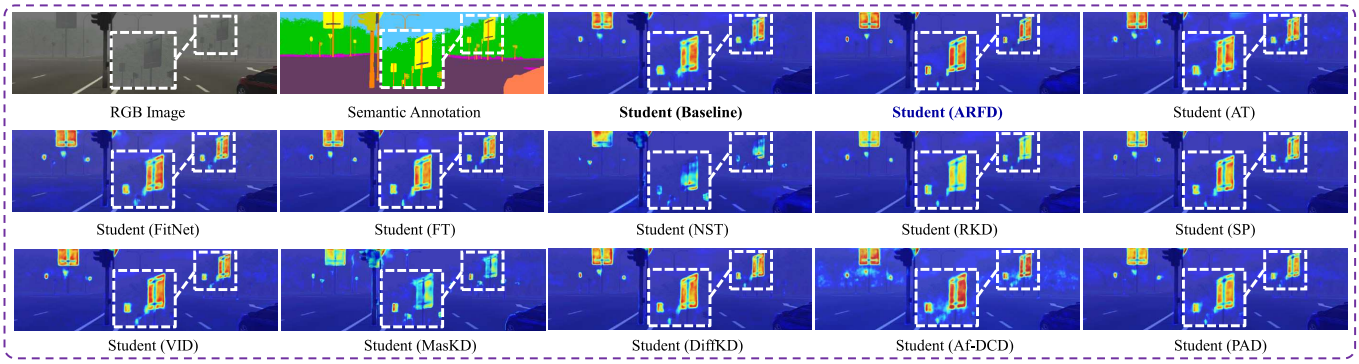

Fig. 6. Probabilities of the "traffic sign" class produced by student networks, where the red and blue colors correspond to the high and low probability of the predictions, respectively. Significantly improved areas are highlighted with white dashed boxes.

TABLE IX
COMPARISON BETWEEN LIX AND OTHER KD METHODS WHEN USING HFIT AS THE BASELINE NETWORK

KD Type	Algorithm	mIoU (%)
/	Teacher (Baseline w/o KD)	84.7
	Student (Baseline w/o KD)	79.1
LD	KD [12]	80.3
	CWKD [11]	79.5
FD	FitNet [42]	80.3
LD+FD	**LIX (Ours)**	81.2

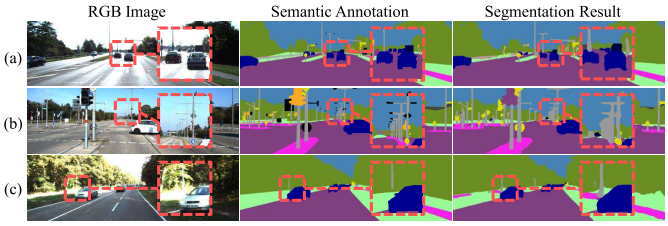

Fig. 7. Unsatisfactory results on the KITTI Semantics dataset: (a) distant areas; (b) congested objects; (c) over-exposed areas. Challenging areas are highlighted with red dashed boxes.

architecture. The teacher network adopts the HFIT architecture with the side adapter, while the student network employs the HFIT architecture without the side adapter. Quantitative experimental results on the Cityscapes [55] dataset are reported in Table IX, indicating that our proposed LIX outperforms other SoTA logit and feature distillation algorithms. These findings further validate the effectiveness and superiority of LIX when applied to VFM-based architectures. A promising direction for future research is the development of dedicated knowledge distillation methods tailored for VFM-based networks.

To comprehensively validate the effectiveness of the proposed LIX framework, we conduct a series of additional experiments, as detailed in the supplement material. Specifically, we evaluate LIX's generalizability by training the student network on one dataset and testing it on another. Furthermore, we utilize Transformer-based architectures, including OFF-Net [73] and HFIT [72], along with real-world datasets, including CityScapes [55] and ADE20K [68], to further demonstrate the compatibility of our proposed LIX across diverse architectures and datasets. Additionally, we compare LIX's training overhead with that of other KD methods and conduct a per-category performance evaluation to quantify improvements across different categories. We further provide visual analyses of the KD loss distribution, the NCLD weight $\boldsymbol{\beta}^{\mathcal{S}}$, and the confidence map to enhance interpretability. These results provide additional quantitative and qualitative evidence, highlighting not only LIX's superior performance and generalizability but also its limitations.

V. DISCUSSION

While LIX demonstrates significant progress, it still faces challenges in several scenarios. As shown in Fig. 7(a), the segmentation of distant objects remains unsatisfactory due to ambiguous representations in low-resolution feature maps. As shown in Figs. 7(b) and (c), inaccurate depth estimation, particularly in areas with object congestion or over-exposure, can impair the teacher network's ability to transfer spatial geometric knowledge. Moreover, while our LIX framework improves the network's generalizability to some extent, the zero-shot performance of the student network remains suboptimal. This limitation may stem from substantial variations in spatial geometric information across datasets, such as differing depth ranges. Addressing these limitations presents promising directions for future research and optimization.

Additionally, since the LIX framework is inherently modular, it is extendable beyond autonomous driving. In particular, DWLD is applicable to tasks with softmax-based outputs, such as medical image segmentation [70] and remote sensing [74]. ARFD, with its feature recalibration and CKA-based consistency measurement, can be adapted to diverse vision tasks such as object tracking [38], [60], [75], [76], image classification [16], [39], [42], and object detection [11]. As a versatile distillation framework, LIX's ability to handle diverse data modalities makes it a promising solution for applications such as multi-modal learning and cross-domain adaptation. For instance, in the crop monitoring task, LIX can infuse prior knowledge from a multi-spectral teacher network into a student network trained exclusively with RGB images. This flexibility highlights LIX's potential as a general framework for knowledge distillation across various fields.

VI. CONCLUSION AND FUTURE WORK

This article discussed a new computer vision problem: the implicit infusion of spatial geometric prior knowledge

acquired by a data-fusion teacher network into a single-modal student network. We contributed to both logit distillation and feature distillation by introducing the DWLD and the ARFD algorithms, respectively. We extended the DKD algorithm by introducing a logit-wise dynamic weight controller, which assigns an appropriate weight to each logit. As for FD, we introduced two novel techniques: feature recalibration via kernel regression and feature consistency quantification via CKA. Through extensive experiments conducted with representative data-fusion semantic segmentation networks on public autonomous driving datasets, we validated the effectiveness and superior performance of our developed LIX framework. Our future work will primarily concentrate on refining the designs of LIX for greater feasibility and generalizability.

VII. Acknowledgment

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of European Union-European Commission. Neither European Commission nor European Union can be held responsible for them.

References

[1] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[2] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.

[3] J. Zhang et al., "Delivering arbitrary-modal semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1136–1147.

[4] N. Wang et al., "SegNet4D: Efficient instance-aware 4D semantic segmentation for LiDAR point cloud," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 15339–15350, 2025.

[5] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024.

[6] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 340–356.

[7] Z. Wu, Y. Feng, C.-W. Liu, F. Yu, Q. Chen, and R. Fan, "S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 2, pp. 3940–3951, Feb. 2024.

[8] Z. Huang, J. Li, P. Zhong, and R. Fan, "Environment-driven online LiDAR-camera extrinsic calibration," *IEEE Trans. Autom. Sci. Eng.*, 2025.

[9] Z. Huang, Y. Zhang, Q. Chen, and R. Fan, "Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching," *IEEE Trans. Intell. Vehicles*, vol. 10, no. 5, pp. 3531–3542, May 2025, doi: 10.1109/TIV.2024.3456299.

[10] C. Li, G. Cheng, and J. Han, "Boosting knowledge distillation via intra-class logit distribution smoothing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4190–4201, Jun. 2024.

[11] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5291–5300.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[13] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[14] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1–11.

[15] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," 2017, *arXiv:1707.01219*.

[16] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3962–3971.

[17] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1365–1374.

[18] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9163–9171.

[19] W. Zheng, M. Hong, L. Jiang, and C.-W. Fu, "Boosting 3D object detection by simulating multimodality on point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13628–13637.

[20] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5116–5125.

[21] J. Cen et al., "CMDFusion: Bidirectional fusion network with cross-modality knowledge distillation for LiDAR semantic segmentation," *IEEE Robot. Autom. Lett.*, vol. 9, no. 1, pp. 771–778, Jan. 2024.

[22] W. Zheng, L. Jiang, F. Lu, Y. Ye, and C.-W. Fu, "Boosting single-frame 3D object detection by simulating multi-frame point clouds," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4848–4856.

[23] S. Qiu, F. Jiang, H. Zhang, X. Xue, and J. Pu, "Multi-to-single knowledge distillation for point cloud semantic segmentation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 9303–9309.

[24] C. C. Aggarwal, *Data Mining: The Textbook*, vol. 1. Cham, Switzerland: Springer, 2015.

[25] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11953–11962.

[26] T. Nguyen, M. Raghu, and S. Kornblith, "Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth," 2020, *arXiv:2010.15327*.

[27] W.-D. Kurt, J. Lewis, and W. B. Kleijn, "The HSIC bottleneck: Deep learning without back-propagation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 4, pp. 5085–5092.

[28] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, Jan. 2021, Art. no. 104042.

[29] J. Huang et al., "RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Vehicles*, vol. 10, no. 5, pp. 3156–3165, May 2025, doi: 10.1109/TIV.2024.3448251.

[30] Y. Feng et al., "SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, 2025.

[31] J. Huang, J. Li, S. Vityazev, A. Dvorkovich, and R. Fan, "DepthMatch: Semi-supervised RGB-D scene parsing through depth-guided regularization," *IEEE Signal Process. Lett.*, vol. 32, pp. 2549–2553, 2025.

[32] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115.

[33] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "AdapNet: Adaptive semantic segmentation in adverse environmental conditions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 4644–4651.

[34] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3029–3037.

[35] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[36] G. Bang, K. Choi, J. Kim, D. Kum, and J. W. Choi, "RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15491–15500.

[37] W. Zhou, Y. Li, J. Huan, Y. Liu, and Q. Jiang, "MSTNet-KD: Multilevel transfer networks using knowledge distillation for the dense prediction of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

[38] J. Shen, Y. Liu, X. Dong, X. Lu, F. S. Khan, and S. Hoi, "Distilled Siamese networks for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8896–8909, Dec. 2022.

[39] H. Zhou et al., "Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–7.

[40] T. Huang, S. You, F. Wang, Q. Chen, and C. Xu, "Knowledge distillation from a stronger teacher," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 33716–33727.

[41] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2015, pp. 912–921.

[42] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[43] T. Huang et al., "Masked distillation with receptive tokens," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. -1–17.

[44] J. Fan, C. Li, X. Liu, M. Song, and A. Yao, "Augmentation-free dense contrastive knowledge distillation for efficient semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 51359–51370.

[45] Y. Zhang et al., "Prime-aware adaptive distillation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 658–674.

[46] Z. Guo, C. Zhang, Y. Fan, Y. Tian, C. Zhang, and N. V. Chawla, "Boosting graph neural networks via adaptive knowledge distillation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, vol. 37, no. 6, pp. 7793–7801.

[47] B. Peng et al., "Correlation congruence for knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5006–5015.

[48] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, vol. 33, no. 1, pp. 3779–3787.

[49] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7130–7138.

[50] S. Kornblith, M. Norouzi, H. Lee, and G. E. Hinton, "Similarity of neural network representations revisited," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 3519–3529.

[51] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for learning kernels based on centered alignment," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 795–828, 2012.

[52] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3061–3070.

[53] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[54] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," 2020, *arXiv:2001.10773*.

[55] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[56] C.-W. Liu, Q. Chen, and R. Fan, "Playing to vision foundation model's strengths in stereo matching," *IEEE Trans. Intell. Vehicles*, early access, Sep. 25, 2024, doi: 10.1109/TIV.2024.3467287.

[57] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10371–10381.

[58] T. Huang et al., "Knowledge diffusion for distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 65299–65316.

[59] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[60] Y. Cui, T. Song, G. Wu, and L. Wang, "MixFormerV2: Efficient fully transformer tracking," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 58736–58751.

[61] V. Borsuk, R. Vei, O. Kupyn, T. Martyniuk, I. Krashenyi, and J. Matas, "FEAR: Fast, efficient, accurate and robust visual tracker," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 644–663.

[62] X. Chen, B. Kang, D. Wang, D. Li, and H. Lu, "Efficient visual tracking via hierarchical cross-attention transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2023, pp. 461–477.

[63] P. Blatter, M. Kanakis, M. Danelljan, and L. V. Gool, "Efficient visual tracking with exemplar transformers," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1571–1581.

[64] B. Yan, H. Peng, K. Wu, D. Wang, J. Fu, and H. Lu, "LightTrack: Finding lightweight neural networks for object tracking via one-shot architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15180–15189.

[65] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10448–10457.

[66] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6181–6190.

[67] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC+$: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 12549–12556.

[68] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 633–641.

[69] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[70] D. Qin et al., "Efficient medical image segmentation based on knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3820–3831, Dec. 2021.

[71] H. Fan et al., "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5369–5378.

[72] S. Guo, T. Wen, C.-W. Liu, Q. Chen, and R. Fan, "Fully exploiting vision foundation model's profound prior knowledge for generalizable RGB-depth driving scene parsing," 2025, *arXiv:2502.06219*.

[73] C. Min et al., "ORFD: A dataset and benchmark for off-road freespace detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2532–2538.

[74] Y. Yang et al., "Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[75] W. Han, X. Dong, Y. Zhang, D. Crandall, C.-Z. Xu, and J. Shen, "Asymmetric convolution: An efficient and generalized method to fuse feature maps in multiple vision tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 11, pp. 7363–7376, Nov. 2024.

[76] X. Dong, J. Shen, F. Porikli, J. Luo, and L. Shao, "Adaptive Siamese tracking with a compact latent network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8049–8062, Jul. 2022.

Sicen Guo (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with Tongji University. Her research interests include computer vision and deep learning.

Ziwei Long has been a Research Assistant at Tongji University since 2024. His research interests include computer vision and deep learning.

Zhiyuan Wu is currently pursuing the B.E. degree with Tongji University. His research interests include computer vision and deep learning.

Qijun Chen (Senior Member, IEEE) is currently a Full Professor with the College of Electronic and Information Engineering, Tongji University. His research interests include robotics control, environmental perception, and the understanding of mobile robots and bioinspired control.

Ioannis Pitas (Life Fellow, IEEE) is currently a Professor with the Department of Informatics, AUTH, where he also serves as the Director of the Artificial Intelligence and Information Analysis Laboratory. His current research interests include computer vision, machine learning, autonomous systems, and image/video processing. He is an IEEE Distinguished Lecturer and a fellow of EURASIP.

Rui Fan (Senior Member, IEEE) is currently a tenured Professor with Tongji University. His research interests include computer vision, deep learning, and robotics. He was honored by being included in Stanford University's List of Top 2% Scientists Worldwide between 2022 and 2024, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, acknowledged as one of Xiaomi Young Talents in 2023, and awarded Shanghai Science and Technology 35 Under 35 Honor in 2024.