Knowledge Generation and Distillation for Road Segmentation in Intelligent Transportation Systems

Jianyong Wang, Mingliang Gao, *Senior Member, IEEE*, Wenzhe Zhai, Imad Rida, Xianxun Zhu, and Qilei Li, *Member, IEEE*

***Abstract*— The rapid development of generative AI has opened a new era in Intelligent Transportation Systems (ITS). However, deploying high-performance segmentation models on resource-constrained edge devices remains challenging due to their substantial computational demands. To address this problem, in this work, we propose a lightweight road segmentation framework termed Knowledge Generation and Distillation (KGD). In the KGD, a lightweight Student model learns from a high-precision Teacher model. This approach balances accuracy and computational efficiency. To further enhance the knowledge transfer process, we introduce a knowledge distillation loss to better supervise the discrepancy between the Teacher and Student models. Meanwhile, we incorporate graph convolution to capture complex spatial dependencies. This can effectively enhance the understanding of road structure and irregular boundaries. Additionally, we built a Multi-scale Lightweight Spatial Attention (MS-LSA) module to focus on multi-scale spatial road information. Experimental results demonstrate that the proposed KGD achieves 96.33% and 94.02% in Max F1-measure and average precision(AP) on KITTI-Road, with only 1.17M in parameters and scores 6.73ms/frame in inference speed. It achieves a superior balance between accuracy and efficiency compared to mainstream deep networks. These advantages make KGD suitable for real-time use in large-scale ITS applications, such as smart traffic monitoring, autonomous vehicle perception, and adaptive traffic control systems.**

***Index Terms*— Road segmentation, graph convolutional neural network, tiny machine learning, knowledge distillation.**

I. Introduction

IN RECENT years, autonomous driving has received growing attention in Intelligent Transportation Systems (ITS). Road segmentation serves as a fundamental component of autonomous driving systems and plays a vital role in environmental perception. Earlier segmentation models [1] achieved breakthroughs in accuracy with the introduction of deep learning methods, which significantly advanced the development of road segmentation tasks.

However, deep learning-based road segmentation models are often highly complex, and traditional Convolutional Neural Networks (CNNs) demand substantial computational resources, which can be a barrier to their practical application in ITS environments [2]. To improve deployment on autonomous driving platforms, researchers have proposed lightweight model architectures. Oliveira et al. [3] introduced a multi-scale dilation module to achieve road segmentation with reduced computational overhead. Teichmann et al. [4] improved model efficiency by using a unified classification, detection, and semantic segmentation architecture. By using a shared encoder across tasks to achieve over 23 FPS for inference. Gong et al. [5] used an efficient encoder-decoder architecture to achieve 135 FPS on a single NVIDIA TITAN Xp GPU. However, with the continuous development of ITS, road segmentation models deployed on embedded systems increasingly demand lightweight models with high segmentation accuracy. Striking a balance between model accuracy and efficiency remains a significant challenge for current lightweight road segmentation models. In addition, road segmentation performance often shows noticeable degradation in complex real-world environments. For example, urban environments typically involve dense traffic, irregular lane layouts, and occlusion from vehicles and pedestrians. Various weather conditions cause occlusion of lane boundaries and variation in illumination, which creates difficulty in accurate segmentation.

To address these limitations, we propose a lightweight road segmentation framework called Knowledge Generation and Distillation (KGD). The KGD enables a small model (Student model) to learn from a larger and high-precision model (Teacher model). Additionally, during the training of the Student model, we proposed KDLoss to supervise the Student model's output. KDLoss is primarily obtained by a weighted combination of FSOhemCELoss [6] and DistillationLoss. Moreover, we design a Multi-scale Lightweight Spatial Attention (MS-LSA) module to capture spatial features at different scales. This design aids in road segmentation by

Received 24 March 2025; revised 6 May 2025 and 24 May 2025; accepted 4 June 2025. This work was supported by Shandong Province Undergraduate Teaching Reform Project under Grant Z2024184. The Associate Editor for this article was R. Jhaveri. *(Corresponding authors: Mingliang Gao; Xianxun Zhu.)*

Jianyong Wang, Mingliang Gao, and Wenzhe Zhai are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: 23404020559@stumail.sdut.edu.cn; mlgao@sdut.edu.cn; wenzhezhai@outlook.com).

Imad Rida is with the Laboratoire Biomécanique et Bioingénierie UMR 7338, Centre de Recherches de Royallieu, Université de Technologie de Compiègne, 60200 Compiègne, France (e-mail: imad.rida@utc.fr).

Xianxun Zhu is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: zhuxianxun@shu.edu.cn).

Qilei Li is with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China (e-mail: qilei.li@outlook.com).

Digital Object Identifier 10.1109/TITS.2025.3577794

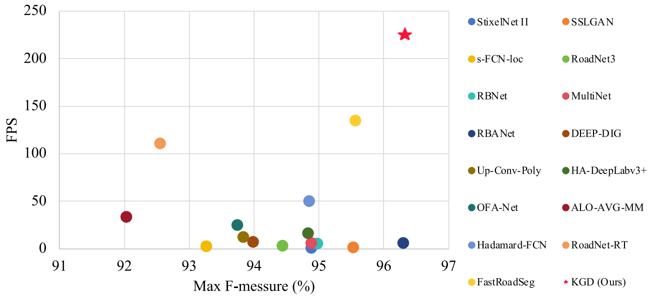


Fig. 1. Comparative results of accuracy and efficiency (Frames/second) for various road segmentation methods.

capturing various levels of detail (*e.g.,* lane markings, edges, and road surfaces) and semantic cues (such as the relative positioning of distant objects). Road structures, like lanes, frequently exhibit intricate geometric shapes and topologies. This complexity poses challenges for conventional Convolutional Neural Networks (CNNs). To overcome this limitation, we adopt graph convolution to flexibly model the topological structure of road regions to improve segmentation accuracy. Graph convolution facilitates information exchange among nodes (*i.e.,* pixels or regions) and enables broader contextual understanding through node connections. In road segmentation tasks, graph convolution can build closer relationships between distant pixels that share similar semantics. This helps the model recognize distant but coherent road areas and results in more precise segmentation.

To evaluate the effectiveness of KGD, we compare its performance with state-of-the-art road segmentation models on the KITTI-Road benchmark [7]. As shown in Fig. 1, KGD demonstrates a superior balance between accuracy (Max F-measure) and efficiency (Frames Per Second). The results highlight that while most models face difficulty achieving both high segmentation accuracy and real-time inference, KGD achieves a compelling trade-off, which makes it particularly suitable for deployment in real-world, time-sensitive autonomous driving scenarios. The main contributions of this work are as follows,

- We explore the application of knowledge distillation and propose a KGD for road segmentation. It achieves high accuracy while maintaining a low parameter count to meet the constraints of Intelligent Transportation Systems (ITS).
- We design a knowledge distillation loss to better supervise the discrepancy between the teacher and Student models.
- We build an MS-LSA module and adopt graph convolution to effectively capture semantic information in complex scenes.

The remainder of this paper is organized as follows. Section II reviews related works on road segmentation, lightweight neural networks, knowledge distillation, and graph neural networks. Section III introduces the proposed KGD architecture and provides a detailed description of its components. Section IV presents the implementation details, datasets and evaluation metrics, experimental results and analysis. Section V concludes the paper and discusses future research directions.

II. Related Works

This section provides a review of the literature relevant to the proposed approach for road segmentation. We first discuss existing methods in road segmentation (Section II-A) to contextualize our work. Subsequently, we examine advancements in lightweight neural networks (Section II-B), which are crucial for deployment in resource-constrained ITS. We then explore the principles and applications of knowledge distillation in visual tasks (Section II-C), a core component of our framework. Finally, we review the use of graph neural networks (Section II-D) for capturing complex spatial dependencies, which we leverage to enhance segmentation performance.

A. Road Segmentation

With the rapid development of autonomous driving technology, road segmentation has become an essential component for scene understanding and environmental perception. Early approaches relied on handcrafted features and traditional classifiers, such as random forests [8] and support vector machines [9], which often struggled with generalization in diverse environments.

The emergence of deep learning brought significant improvements. Fully Convolutional Networks (FCNs) introduced by Shelhamer et al. [10] enabled pixel-wise classification by replacing fully connected layers with convolutional layers, which laid the foundation for semantic segmentation. Chang et al. [11] proposed an uncertainty-aware symmetric network. The network includes an uncertainty-aware fusion module to balance the trade-off between segmentation speed and accuracy. It enables the model to adapt more effectively to variations in input data while maintaining fast inference speed. Ravishankar et al. [12] proposed a hypercolumn-based random forest of local experts for unstructured road segmentation, which combines CNN features with superpixel pooling to achieve efficient and accurate detection.

The aforementioned methods have excelled in road segmentation tasks. However, with the rapid progress in autonomous driving technology. The requirements of segmentation speed and accuracy for road segmentation networks are becoming increasingly stringent. In this work, we adopt a knowledge distillation framework in the road segmentation task to tackle this issue. The proposed framework leverages a larger model to produce an effective segmentation model, which serves as a pre-trained model and ground truth for the joint training of a lightweight segmentation network.

B. Lightweight Neural Network

In resource-constrained scenarios such as autonomous driving and edge deployment, the efficiency of neural networks becomes a critical design consideration. Conventional high-capacity networks, although achieving high accuracy, often incur prohibitive computational costs that make real-time inference impractical. As a result, significant research efforts have focused on developing lightweight models that maintain competitive performance with reduced computational complexity.

To address this problem, researchers introduced efficient models. Howard [13] proposed MobileNets, an efficient CNN designed for mobile and embedded vision applications. This model introduces two global hyperparameters: width multiplier and resolution multiplier. The width multiplier adjusts the number of channels in each layer, while the resolution multiplier controls the input image size. These parameters allow users to balance model size, computational cost, and accuracy based on resource constraints. Zhang et al. [14] developed ShuffleNet, which reduces computational complexity. This model applies group convolutions to divide feature maps into smaller groups and decrease the number of operations. It also uses channel shuffling to ensure information flows across groups. These techniques deduce computational demand while preserving segmentation accuracy. Redmon et al. [15] proposed a single-stage object detection framework that unifies localization and classification within a single neural network. This design enables real-time and efficient detection with competitive accuracy by processing each image through a single forward pass. Liu et al. [16] proposed a one-stage object detection framework that discretizes predictions across multiple feature maps and scales. It achieves high accuracy in real-time detection without relying on region proposal generation. Iandola et al. [17] introduced SqueezeNet, a lightweight convolutional neural network that achieves AlexNet-level accuracy while using 50 times fewer parameters and a model size under 0.5 MB.

Although these approaches have optimized parameter efficiency while retaining accuracy, they still fall behind larger networks. To bridge this gap, we transfer knowledge from larger models to smaller ones via distillation, which better balances model accuracy and efficiency.

C. Knowledge Distillation in Visual Tasks

Knowledge Distillation (KD) [18] is a technique used for model compression and knowledge transfer. In this approach, a larger model (Teacher model) provides guidance to train a smaller model (Student model). In the KD framework, the Teacher model encapsulates rich knowledge that can be transferred to the Student model appropriately. This process enhances the Student model's performance while reducing computational overhead.

Specifically, Zhu et al. [19] transferred feature knowledge from a high-resolution Teacher model to a low-resolution Student model using joint optimization of feature distillation loss and recognition loss. Wu et al. [20] extracted global feature distribution information from the Teacher model using a combination of modified knowledge distillation and relational embedding loss. This approach effectively reduces computational resources while enhancing accuracy in face recognition tasks. Zhang et al. [21] integrated KD mechanisms with cross-branch integration modules and action knowledge graphs to effectively combine human and scene knowledge. Chawla et al. [22] enhanced KD by generating diverse target object images using improved data augmentation and automated bounding box sampling. This method enables effective KD in scenarios with no original data. Chen et al. [23] introduced a network that combined multi-task learning with a mean Teacher model. By leveraging labeled and a large amount of unlabeled data, this method improved shadow detection performance through joint optimization of supervised and consistency losses.

In this work, we propose a KDLoss by combining the road segmentation task and the knowledge distillation technique. Different from the existing knowledge distillation methods, KDLoss not only combines the online hard example mining strategy (FSOhemCELoss) to enhance the learning ability of the model for complex regions (such as road boundaries and shadow regions), but also aligns the output distribution of the teacher model and the student model by KL divergence loss (DistillationLoss).

D. Graph Neural Network

Real-world applications often involve non-Euclidean data, in which traditional CNNs are difficult to handle. Graph Convolutional Networks (GCNs) offer a powerful solution to this problem. GCNs play important roles in image segmentation. Meng et al. [24] introduced a multi-level aggregation network that combines convolutional neural networks (CNNs) with attention refinement modules (ARMs) and graph convolutional networks (GCNs). This approach extracts rich semantic information from images and leverages GCNs to emphasize object contours. Soberanis-Mukul et al. [25] proposed a segmentation refinement method based on uncertainty analysis and graph convolutional networks (GCNs). This method applies GCNs to analyze uncertainty and address a semi-supervised graph learning problem, and it shows strong performance in medical image segmentation.

Although some graph convolution algorithms address vanishing gradients and limited receptive fields, these methods still face challenges with complex image topologies. In this work, the proposed graph convolution method maximizes relative differences in the neighborhood aggregation operation.

III. Methodology

A. Overview

The architecture of the proposed Knowledge Generation and Distillation Modeling (KGD) is illustrated in Fig. 2. The input is the road image and the output is the road segmentation result. The top section shows the knowledge distillation framework. The Teacher network (ResNet50) is pre-trained specifically for the road segmentation task. It guides a Student network (ResNet18) using KDLoss. The bottom section illustrates the Teacher and Student Network with two branches: a high-resolution branch and a low-resolution branch. In the low-resolution branch, the downsampling scales of width and height are set to 1/4 and 1/2, respectively. The reason behind this design strategy lies that roads usually stretch horizontally in view of the vehicle. A higher downsampling rate

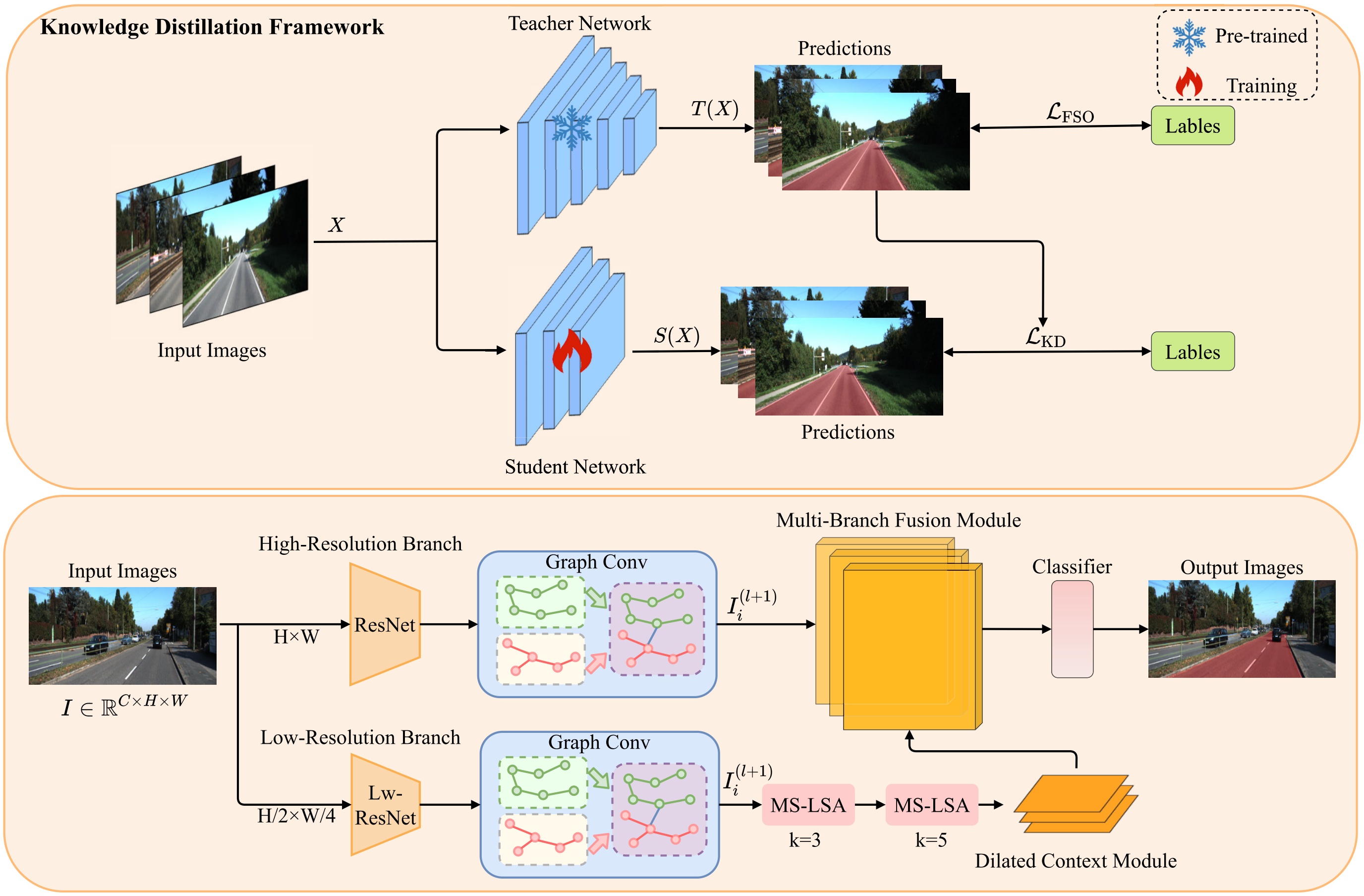


Fig. 2. Architecture of KGD for road segmentation.

in the width direction removes unnecessary details. In contrast, height is more important for road boundaries and lane markings. A smaller downsampling rate in this direction helps maintain segmentation accuracy. The two branches process multi-scale features via ResNet and lightweight ResNet (Lw-ResNet), graph convolutions, and the MS-LSA and dilated context modules (in the low-resolution branch). The features are fused and passed through a classifier to produce the segmented road map.

B. Knowledge Distillation for Road Segmentation

To achieve a lightweight road segmentation network, we designed a knowledge distillation framework tailored for road segmentation tasks. This method includes a pre-trained, large-scale Teacher model T and a lightweight Student model S. Once the training begins, the Teacher model remains frozen throughout the process. For an input image X, P_T and P_S represent the feature maps generated by the Teacher model and Student model, respectively. It is formulated as follows,

$$P_T = T(X), \quad P_S = S(X) \tag{1}$$

where T represents the Teacher model, S represents the Student model, and X is the input image.

The framework aims to minimize the difference between the feature maps P_T and P_S to ensure that the lightweight

TABLE I
MATHEMATICAL NOTATIONS

Notation	Description
$\mathbf{X}$	Input image
$\mathbf{T}$	Teacher model
$\mathbf{S}$	Student model
$\mathbf{a}, \mathbf{b}$	The pixels in the image
$\mathcal{L}_{\text{FSO}}$	The FSOhemCELoss
$\mathcal{L}_{\text{ce}}$	The binary cross-entropy loss
$\mathcal{L}_{\text{KL}}$	The DistillationLoss
$\mathcal{L}_{\text{KD}}$	The Knowledge Distillation Loss
$I \in \mathbb{R}^{C\times H\times W}$	The input feature map
$\mathbf{I}_i^{(l)}$	The feature at position i in the feature map
$\mathbf{I}_i^{(l+1)}$	The output of graph convolution
$\mathcal{N}(i)$	The set of neighboring pixels around position i
$\mathcal{W}^{(l)}$	The trainable weight used to concat features
$\bar{F}^C$	The output of depthwise separable convolution
$\mathbf{X}^C$	The output of small-scale depthwise separable convolutions
$\mathbf{B}^C$	The output of dilated convolution
$\mathbf{D}^C$	The output of small-scale dilated convolution

Student model achieves satisfactory road segmentation performance. In this work, to effectively supervise the discrepancy between P_T and P_S, we proposed the Knowledge Distillation Loss (KDLoss), which consists of FSOhemCELoss [6] and DistillationLoss. The FSOhemCELoss incorporates the Online Hard Example Mining (OHEM) strategy along with the commonly used cross-entropy loss in segmentation tasks. This combination enables the model to focus on the pixels that are difficult to classify, such as road boundaries, sidewalks, and shadowed areas. The loss function for FSOhemCELoss is expressed as follows,

$$\mathcal{L}_{\text{FSO}} = \frac{1}{N}\sum_{a,b}\mathbb{O}\left(P^{a,b} < \lambda_b\right)\mathcal{L}_{\text{ce}}^{a,b}, \tag{2}$$

where N represents the total number of pixels, P is the predicted value, a and b are the pixels in the image, λ_b is the confidence threshold. This loss only performs gradient backpropagation for pixels whose predicted confidence P is less than the threshold λ_b, and $\mathbb{O}(\cdot)$ is the indicator function. $\mathcal{L}_{\text{ce}}$ represents the binary cross-entropy loss.

The DistillationLoss is used to measure the similarity between the output distributions of the Student model and the Teacher model using the KL divergence loss. A temperature parameter t is introduced to smooth the probability distribution, making it easier for the Student model to learn the soft labels from the Teacher model. For each pixel, the DistillationLoss (based on KL divergence) is defined as follows,

$$\begin{aligned}&\mathcal{L}_{\text{KL}}(S^{a,b}, T^{a,b})\\ &\quad = T^2 \cdot \sum_{c=1}^{C} P\left(T^{a,b} = c \mid t\right)\log\frac{P(T^{a,b} = c \mid t)}{P(S^{a,b} = c \mid t)},\end{aligned} \tag{3}$$

where $\cdot$ denotes element-wise multiplication. $T^{a,b}$ represents the prediction of the Teacher model at the pixel (a, b). $S^{a,b}$ is the prediction made by the Student model at the same pixel. The temperature parameter is denoted as t, and it controls the smoothing of the probability distribution. $P(T^{a,b} = c \mid t)$ is the probability distribution generated by the Teacher model at temperature t, and $P(S^{a,b} = c \mid t)$ is the probability distribution generated by the Student model at the same temperature.

We performed a weighted sum of the two losses based on specific weights α. The final Knowledge Distillation Loss (KDLoss) function $\mathcal{L}_{\text{KD}}$ is expressed as follows,

$$\mathcal{L}_{\text{KD}} = \alpha \cdot \mathcal{L}_{\text{FSO}} + (1-\alpha)\cdot\mathcal{L}_{\text{KL}}(\mathbf{S}, \mathbf{T}), \tag{4}$$

where $\mathcal{L}_{\text{FSO}}$ is the FSOhemCELoss, $\mathcal{L}_{\text{KL}}$ is the DistillationLoss, $\mathbf{S}$ represents the output of Student network, and $\mathbf{T}$ is the output of Teacher network.

In summary, the FSOhemCELoss component adopts Online Hard Example Mining (OHEM) with spatial constraints and compels the Student model to focus on geometrically complex areas (*e.g.,* intersections with discontinuous lane markings) and ensures spatial coherence. Meanwhile, DistillationLoss reduces the divergence between the Student and Teacher models and aligns feature distributions, so that the Student model can learn the Teacher's long-range dependency patterns, such as the correlation between parallel lanes and the connectivity of intersections.

C. Graph Convolutional Module

The road segmentation task requires the model to capture the spatial relationships of each pixel in the image, particularly the connectivity and continuity of road boundaries. Traditional convolutions can only capture features from local neighborhoods when processing regular grid data (*e.g.,* images). In contrast, graph convolution operates on graph structures and can model irregular spatial dependencies. Moreover, graph convolution enhances the model's perception of irregular boundaries. This ensures more precise segmentation results, especially along road edges. This is particularly effective for segmenting roads with asymmetric shapes in complex scenes. By applying graph convolution to both high-resolution and low-resolution branches, we can enhance the model's understanding of the overall road structure and avoid false segmentation that may arise from relying solely on local features.

In road segmentation tasks, feature maps can be viewed as a grid, with each pixel representing a node in the graph. When applying graph convolution to the feature map, we can define the neighborhood of each pixel and establish irregular spatial dependencies, as illustrated by the Graph-Conv module in Fig. 2.

Given an input feature map, $I \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of input channels, H and W denote the height and width of the image, respectively. We defined a neighborhood aggregation operation and constructed a relative graph convolution structure by using the maximum relative difference. The graph convolution can be expressed as,

$$I_i^{(l+1)} = \sigma\left(W^{(l)}\left(I_i^{(l)} \oplus \max_{j\in\mathcal{N}(i)}\left(I_j^{(l)} - I_i^{(l)}\right)\right)\right), \tag{5}$$

where $I_i^{(l)}$ represents the feature at position i in the feature map, $\mathcal{N}(i)$ denotes the set of neighboring pixels around position i, and $W^{(l)}$ is a trainable weight used to perform a linear transformation on the concatenated features. The $\oplus$ operation indicates feature concatenation, where the node's feature $I_i^{(l)}$ is concatenated with the maximum relative feature difference. The $\max_{j\in\mathcal{N}(i)}\left(I_j^{(l)} - I_i^{(l)}\right)$ function represents the maximum feature difference between the node and its neighbor.

D. Multi-Scale Lightweight Spatial Attention Module

In road segmentation tasks, objects such as lane markings, road edges, and intersections often exhibit variations in scale and spatial distribution. Standard convolutional layers with fixed receptive fields fail to capture these multi-scale patterns effectively, especially under complex road conditions. Attention mechanisms can improve spatial feature extraction, but many existing methods rely on large-scale convolutions or self-attention, which increase computational cost and limit deployment on edge devices. To address this issue, we introduce a Multi-scale Lightweight Spatial Attention (MS-LSA) module that enhances the model's ability to focus on relevant features at different spatial scales while maintaining low computational overhead.

The schematic of the MS-LSA module is delineated in Fig. 3. This module is designed as a lightweight multi-scale spatial attention mechanism that can be configured with different convolution kernel sizes to capture features at various scales. The input feature map $I \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of input channels, H represents the height of the feature map, and W is the width. The motivation

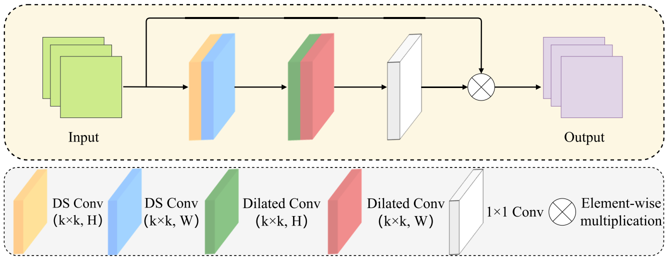


Fig. 3. Architecture of MS-LSA module. The DS Conv is denoted as depthwise separable convolution. The Dilated Conv denotes dilated convolution.

behind developing MS-LSA stems from the inherent limitations of existing approaches. While traditional attention mechanisms [26], [27] and large-scale deep convolutions offer powerful feature extraction capabilities, they often come with significant drawbacks. For instance, traditional global self-attention mechanisms suffer from quadratic computational complexity with respect to input resolution, which makes them impractical for high-resolution feature maps. Meanwhile, non-local operations introduce significant memory overhead when modeling long-range dependencies. Similarly, for large-scale convolutions, their parameter redundancy and fixed receptive fields limit both computational efficiency and adaptability to multi-scale spatial relationships.

To address the issue of high computational cost, we replaced the large-scale convolutions with two smaller-depth separable convolutions. Meanwhile, to maintain good spatial characteristics, we decomposed these convolutions into horizontal and vertical separable convolutions. Additionally, we decomposed the dilated convolutions into two small-scale convolutions: one horizontal and one vertical. This improvement helps the model capture long-distance spatial dependencies. The output of the MS-LSA can be computed as,

$$\bar{F}^C = \sum_W X^C_{k\times 1} * (\sum_H X^C_{1\times k} * I^C), \tag{6}$$

where $*$ denotes element-wise multiplication, $\bar{F}^C$ is the output of depthwise separable convolution, $\sum_W X^C_{k\times 1}$ and $\sum_H X^C_{1\times k}$ denote small-scale depthwise separable convolutions in two spatial directions, respectively. k denotes the convolutions kernel size. And then the output of the dilated convolution can be formulated as,

$$B^C = \sum_W D^C_{k\times 1} * (\sum_H D^C_{1\times k} * \bar{F}^C), \tag{7}$$

where $\sum_W D^C_{k\times 1}$ and $\sum_H D^C_{1\times k}$ denote small-scale dilated convolutions in two spatial directions, respectively. The final output can be expressed as follows,

$$A^C = (n_{1\times 1} * B^C) \otimes I^C, \tag{8}$$

where $n_{1\times 1}$ denote 1×1 convolution, and $\otimes$ is the Hadamard product.

In the low-resolution branch, we configured MS-LSA modules with kernel sizes $k = 3$ and $k = 5$. This setup allows the extraction of context information at different scales to enhance the features in the spatial dimension.

IV. Experiments

A. Implementation Details

In the experimental setup, the feature extraction backbones are ResNet50 (Teacher network) and ResNet18 (Student network), with parameters initialized from ImageNet pre-training. The training starts with a learning rate of 0.01 and it is decayed to 1e-5 using a cosine annealing schedule. We used a batch size of 6 and trained KGD for 500 epochs to ensure model convergence. We used a single NVIDIA 3090 GPU for model training. For model efficiency computation (inference times, FPS), we used a single NVIDIA 2080 SUPER GPU.

The initial learning rate was set to 0.01 based on standard practice in training residual networks with moderate depth, which balances convergence speed and training stability. We employed cosine annealing to gradually reduce the learning rate during training to avoid sharp drops and ensure smoother convergence. The batch size was empirically set to 6 to accommodate the high-resolution input images (375×1240) and the limited memory of the GPU used (NVIDIA RTX 3090).

B. Datasets and Evaluation Metrics

1) Datasets: To evaluate the proposed methods, we used the KITTI-Road dataset [7], which contains 289 training images and 290 testing images. Road scenes are classified into three types, namely Urban Unmarked (UU), Urban Marked (UM), and Urban Multiple Marked (UMM). The URBAN category (UM/UMM/UU) includes multi-lane roads, intersections, and unmarked roads, which have irregular shapes and complex topological structures. For example, multi-lane roads in UMM require capturing long-range spatial dependencies, whereas unmarked roads in UU rely on contextual inference to determine boundaries. Road markings (such as thin lane lines in UM), shadow occlusions (*e.g.,* tree shadows in UU), and low-resolution distant regions require consideration of both fine-grained details and high-level semantics. Furthermore, the dataset covers variations in illumination, perspective distortion, and scene complexity. These characteristics collectively make the KITTI-Road dataset a rigorous benchmark for assessing the robustness and generalization ability of road segmentation models. The resolution of training images varies from 370×1224 to 375×1242, and we standardized it to 375×1240 through padding as a data preprocessing step. The experiments used 5-fold cross-validation on the training set, with results presented as mean values for consistency with previous approaches.

2) Evaluation Metrics: In this work, six metrics, namely average precision (AP), maximum F1-measure (MaxF), recall (REC), precision (PRE), false negative rate (FNR), and false positive rate (FPR) are adopted for performance evaluation. MaxF was chosen as the main accuracy metric because it balances precision and recall. Metrics were calculated in Birds Eye View (BEV), the standard format for the KITTI-Road dataset. Additionally, we analyzed computational costs, including network Parameters, MACs (Multiply-Accumulate Operations), and Inference time. To minimize the effects

TABLE II

COMPARISON WITH PRIOR NOTABLE ROAD SEGMENTATION METHODS ON THE KITTI-ROAD DATASET. FOR THE FIRST AND SECOND GROUP, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SUB-BEST RESULTS ARE HIGHLIGHTED IN UNDERLINE. "↑" MEANS THE HIGHER THE BETTER. "↓" MEANS THE LOWER THE BETTER

Method	Input shape	MaxF(%)↑	AP(%)↑	PRE(%)↑	REC(%)↑	FPR(%)↓	FNR(%)↓	Time(ms)↓	Device
StixelNet II [28]	370 × 800	94.88	87.75	92.97	96.87	4.04	3.13	1200	Quadro M6000
SSLGAN [29]	375 × 1242	95.53	90.35	95.84	95.24	2.28	4.76	700	TITAN X
s-FCN-loc [30]	500 × 500	93.26	-	94.16	92.39	3.16	7.61	400	Tesla K80
RoadNet3 [31]	160 × 600	94.44	93.45	94.69	94.18	2.91	5.82	300	GTX 950M
RoadFormer [32]	352 × 640	**97.50**	**93.85**	97.16	**97.84**	1.57	**2.16**	210	GTX 3090
RBNet [33]	300 × 900	94.97	91.49	94.94	95.01	2.79	4.99	180	Tesla K20c
MultiNet [4]	384 × 1248	94.88	93.71	94.84	94.91	2.85	5.09	170	GTX 1080
RBANet [34]	360 × 720	96.30	89.72	95.14	97.50	2.75	2.50	160	TITAN Xp
SkipcrossNets [35]	1280 × 384	96.85	90.15	**97.45**	97.14	**0.57**	2.84	146	GTX 1080 Ti
DEEP-DIG [36]	-	93.98	93.65	94.26	93.69	3.14	6.31	**140**	TITAN X
Up-Conv-Poly [3]	500 × 500	93.83	90.47	94.00	93.67	3.29	6.33	80	TITAN X
HA-DeepLabv3+ [37]	-	94.83	93.24	94.77	94.89	2.88	5.11	60	-
OFA-Net [38]	-	93.74	85.37	90.36	**97.38**	5.72	**2.62**	40	-
ALO-AVG-MM [39]	192 × 624	92.03	85.64	90.65	93.45	5.31	6.55	30	GTX 1080
Hadamard-FCN [40]	375 × 1242	94.85	91.48	94.81	94.89	2.86	5.11	20	TITAN X
RoadNet-RT [41]	280 × 960	92.55	93.21	92.94	92.16	3.86	7.84	9	GTX 1080
FastRoadSeg [5]	375 × 1240	95.56	93.89	95.53	95.59	2.47	4.41	7.4	TITAN Xp
TEDNet [42]	-	94.62	93.05	94.28	94.96	3.17	5.04	90	GTX 2080 Ti
KGD(Ours)	375 × 1240	**96.33**	**94.02**	**96.69**	95.97	**1.88**	4.03	**6.73**	GTX 2080 SUPER

of model initialization, we averaged inference times over 1000 forward passes, with an input resolution of 375 × 1240.

This work uses MACs as the primary metric for analyzing computational cost, instead of Floating Point Operations (FLOPs). The reason lies in that both MACs and FLOPs estimate computational complexity, but MACs provide a more hardware-relevant and consistent measure. Each MAC corresponds to one multiplication followed by one addition, which reflects the actual arithmetic operations required by convolutional and fully connected layers.

FLOPs may include a broader set of floating-point operations, such as multiplications, additions, divisions, and other operations that vary across implementations. This inconsistency reduces the comparability of FLOPs across different platforms. In contrast, MACs remain consistent regardless of implementation and correlate more directly with inference latency and energy consumption on common deployment hardware, which includes GPUs, FPGAs, and NPUs. For this reason, MACs serve as a more practical and interpretable metric for evaluating model efficiency in intelligent transportation systems.

C. Comparison With State-of-the-Art Methods

A comparison of segmentation performance and efficiency between KGD and the SOTA methods on the KITTI-Road dataset is presented in Table II. The methods were evaluated on different hardware platforms, which may influence inference time due to variations in computational capacity, memory bandwidth, and architecture. In general, the comparison of runtime performance across different devices may not offer a fully standardized benchmark. To ensure a fair evaluation of the proposed method, we validate all timing results using devices (NVIDIA 2080 SUPER GPU) similar to those used in other methods whenever possible. While absolute inference times may vary from device to device, this work focuses on the relative balance between accuracy and efficiency, which remains consistent when comparing models on the same platform. In future work, we plan to provide device-independent metrics such as throughput or energy-delay product under normalized MACs to further support fair comparison.

We organized methods based on time efficiency. Methods requiring more than 100ms were in the upper half (Group 1), while those under 100ms were in the lower half (Group 2). Group 2 focused on high-efficiency methods for primary comparison with SOTA. Although the methods in the first group have longer inference times, data analysis reveals that some methods in the second group still outperform the first group in certain metrics.

We used KITTI benchmark data and efficiency information from the listed SOTA methods to validate the efficiency and accuracy of each model. KGD surpassed all SOTA methods in Group 2 in terms of the primary accuracy metric (MaxF) and achieved a score of 96.33%. It also achieves the optimal average precision (AP) and precision (PRE). Regarding segmentation efficiency, the proposed method scores 6.73 ms,

TABLE III

Model Complexity and Precision Comparison. The Best Results Are Highlighted in **Bold**, and the Sub-Best Results Are Highlighted in Underline. "↑" Means the Higher the Better. "↓" Means the Lower the Better

Method	MaxF(%)↑	#MACs (G)↓	#Params (M) ↓
RoadNet-RT [41]	92.55	-	13.39
SNE-RoadSeg [43]	96.27	78.44	13.62
UNET3+ [44]	95.64	164.63	14.70
LRDNet+ [45]	95.10	-	19.5
CFECA [46]	85.02	716.22	26.11
SkipcrossNets [35]	**96.85**	38.39	2.33
FastRoadSeg [5]	95.56	18.32	11.33
KGD(Ours)	96.33	**11.46**	**1.17**

surpassing all the SOTA methods. Compared to the second-best method, FastRoadSeg [5], KGD achieved a 0.81% MaxF improvement and a 10% improvement in segmentation speed. With only 1.17M parameters versus FastRoadSeg's 11.33M, KGD reduced model size by 89.67%. Additionally, compared to the dense networks in Group 1, KGD is also competitive. Some models in the first group have higher accuracy but operate significantly slower. For example, RoadFormer [32] achieves 97.50% MaxF but takes 210 ms per image, which is 31.2 times slower than KGD. Its dual-encoder design improves accuracy but increases computation, which limits its practicality for real-time applications. KGD balances accuracy and speed through three key designs. The knowledge distillation framework reduces parameters while preserving important features. The graph convolution module captures spatial relationships, which enhances the segmentation of irregular boundaries. The MS-LSA module enhances multi-scale feature extraction without high computational cost. These designs allow KGD to achieve high accuracy with a compact structure.

In terms of precision (PRE) and recall (REC), KGD achieves 96.69% and 95.97%, respectively. The high precision indicates that the proposed model effectively reduces false positive predictions, which is crucial for autonomous driving applications. Additionally, the false positive rate (FPR) is only 1.88%. It demonstrates the robustness of the model in distinguishing road regions from non-road areas. Compared to methods such as FastRoadSeg [5] and TEDNet [42], KGD consistently maintains a lower false negative rate (FNR) of 4.03%. It ensures reliable segmentation even in challenging environments such as shadowed or unmarked roads.

The comparisons of model complexity and key performance metrics (MaxF) between the KGD and SOTA methods are shown in Table III. The experimental analysis showed that although KGD achieves the second-best segmentation accuracy, it achieves the fastest segmentation speed. Moreover, KGD also delivers the best results in terms of model complexity. As shown in Table III, KGD has only 1.17M parameters, which represents a 49.79% reduction compared to SkipcrossNets. In terms of model computation (MACs), KGD scores 11.46 which is reduced by 70.15% compared to SkipcrossNets [5]. In summary, KGD not only reduces model complexity but also maintains the lowest model complexity while performing excellently on other evaluation metrics. This improvement can be attributed to the effective use of knowledge distillation. It reduces model complexity and computational cost. Additionally, graph convolution and MS-LSA modules enhance the ability of the model to capture semantic information and improve segmentation accuracy.

To assess the practicality of deployment on edge devices, we analyzed the computational complexity and model size of KGD. The proposed model contains 1.17M parameters and requires 11.46G MACs, which are significantly lower than most existing approaches. This level of efficiency allows KGD to deploy in real time on resource-constrained platforms. Therefore, KGD satisfies the key requirements for integration into edge devices, such as achieving low latency, maintaining a compact model size, and operating under limited memory consumption.

Fig. 4 illustrates the visualization results of KGD on the KITTI-Road dataset [7]. The dataset is divided into three categories: Urban Marked (UM), Urban Multiple Marked (UMM), and Urban Unmarked (UU). In the UM category, roads typically represent urban streets with standardized lane markings. These roads are generally straight, clearly marked, and are primarily affected by vehicles of varying scales. It shows that the KGD accurately segments marked lane areas, with precise annotations and clear boundaries. The segmented area covers the majority of the road, and the model maintains high detection accuracy, even at distant viewpoints. UMM includes roads with multiple lane markings, such as dual-lane or multi-lane roads, or areas with intersecting lanes. As depicted in the figure, KGD performs highly accurate lane segmentation, especially in bifurcation areas, where lane structures are fully captured. The segmented areas show no significant detection omissions, and the model handles the complex lane markings appropriately. The combination of Graph-Conv and MS-LSA modules enables the model to understand both the global layout and the details of the lanes. UU refers to roads lacking distinct lane markings or road signs. These roads typically represent streets without markings in urban areas.

As depicted in Fig. 4, the absence of clear lane markings makes segmentation more challenging. This is particularly evident as these roads often have vehicles parked along their sides. Nevertheless, KGD still achieves clear road boundary segmentation, even in these challenging scenarios. For example, in the first and second rows, the absence of clear lane markings (*e.g.,* worn-out paint or occlusions by parked vehicles) creates ambiguity in road geometry. Notably, vehicles parked along the curbs (visible in the "Road Image" column) obscure traditional boundary cues. Nevertheless, the "Segmentation Mask" column reveals sharp and continuous boundaries (*e.g.,* red curves outlining curbsides adjacent to parked cars), which demonstrates the graph convolution's ability to infer structural relationships between fragmented road edges and contextual objects (*e.g.,* vehicle wheels aligning with inferred curbs). In the second row of the "Bird's Eye View" column, the segmented road maintains a smooth width transition despite perspective distortion in the original

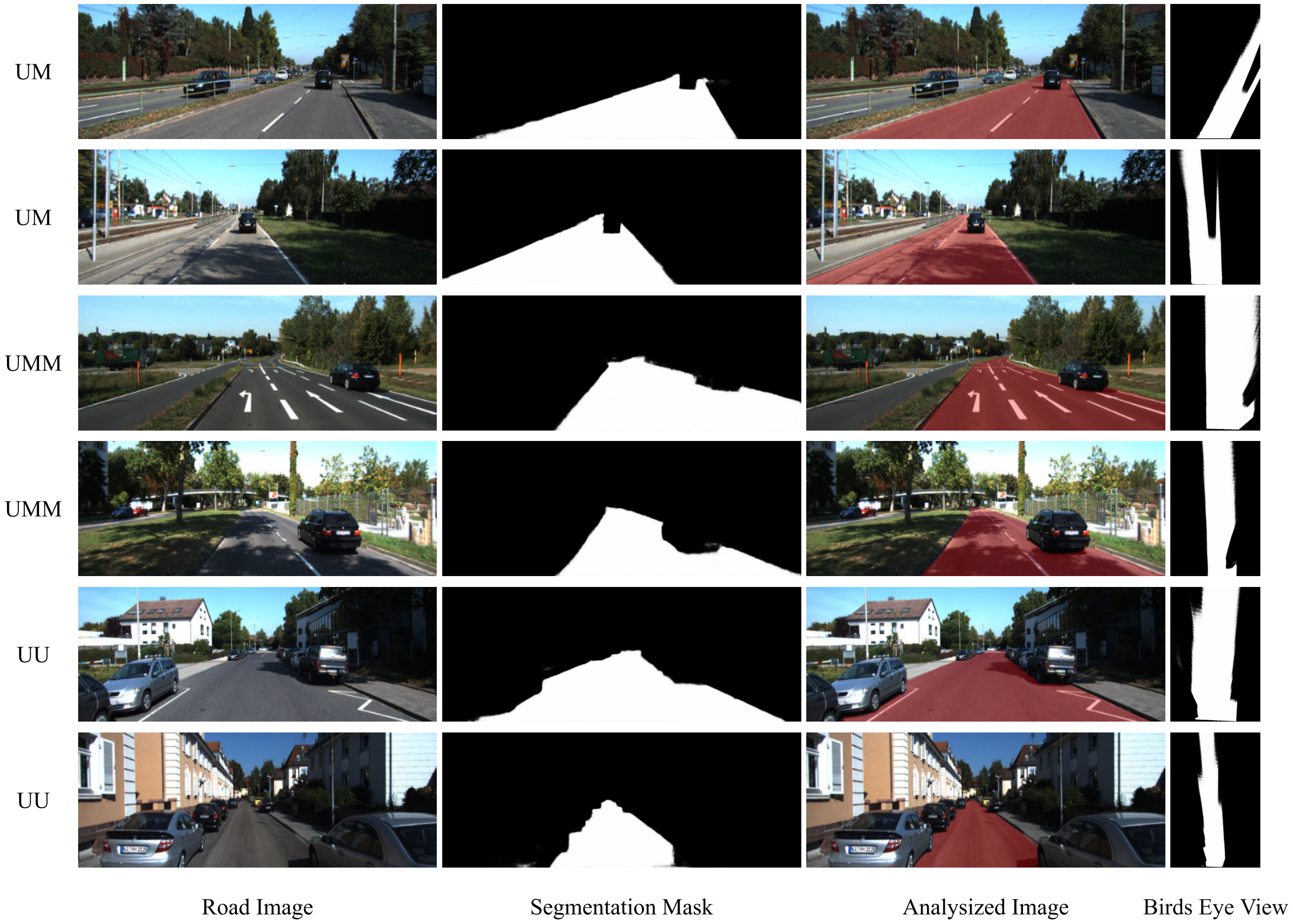


Fig. 4. Visualization result of the proposed KGD on KITTI-Road datasets.

image. This is primarily attributed to the graph convolution's ability to enhance the understanding of road structures and irregular boundaries. Meanwhile, the MS-LSA module effectively emphasizes multi-scale spatial road information, which enables the KGD to accurately segment roads with irregular boundaries across multiple scales.

Compared to existing state-of-the-art methods, KGD achieves a superior balance between segmentation accuracy and computational efficiency. Specifically, KGD surpasses most lightweight models, such as FastRoadSeg [5] and RoadNet-RT [41], by delivering a higher Max F1-measure while maintaining significantly fewer parameters and faster inference speed. The multi-scale lightweight spatial attention module improves the model's ability to handle complex spatial patterns, especially in urban unmarked scenes where lane boundaries are unclear. Furthermore, the integration of graph convolution strengthens the segmentation performance around irregular road structures and occluded areas, which traditional CNN-based methods often misclassify. However, KGD still exhibits slight performance drops in scenes with heavy shadow occlusions or extreme lighting variations, where more global context modeling could further improve robustness. Nevertheless, the proposed framework demonstrates a practical and efficient solution for real-time road segmentation in intelligent transportation systems.

TABLE IV

PERFORMANCE COMPARISON OF DIFFERENT TEACHER MODELS ON KITTI-ROAD DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. "↑" MEANS THE HIGHER THE BETTER. "↓" MEANS THE LOWER THE BETTER

Backbone	MaxF(%)↑	#MACs (G)↓	#Params (M) ↓
VGG16 [47]	96.02	158.69	31.5
MobileNetV2 [48]	96.14	**4.14**	14.75
HRNet [49]	96.56	62.6	**0.94**
ResNet50 [50]	**96.65**	15.26	2.09

D. Ablation Study

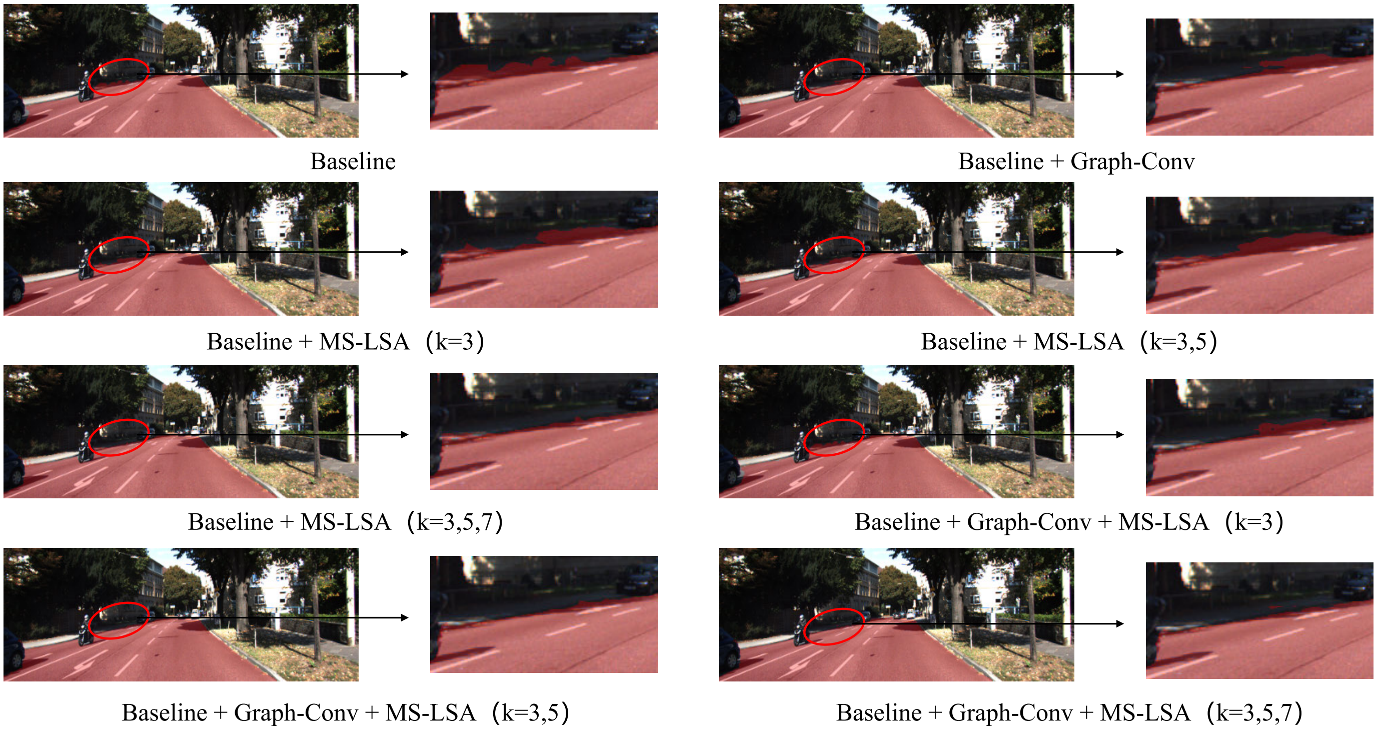
1) Impact of Teacher Model: We analyzed the performance of Teacher models with different backbones. Table IV shows that VGG16 [47] has high computational complexity and a large parameter count, yet its performance is suboptimal. MobileNetV2 [48] has the lowest MACs and is limited by low parameter counts and poor accuracy. HRNet [49] performs well in terms of parameter count but has relatively high MACs, which is unsuitable for resource-constrained environments. However, ResNet50 [50] strikes an ideal balance between MaxF, MACs, and Params.

2) Effectiveness of KGD Components: Additionally, we evaluated the individual components in the KGD.

TABLE V

ABLATION STUDIES ON THE CRITICAL MODULES IN THE KGD. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. "↑" MEANS THE HIGHER THE BETTER. "↓" MEANS THE LOWER THE BETTER

Method	MaxF↑	AP↑	PRE↑	REC↑	FPR↓	FNR↓
Baseline	95.21	93.71	95.35	95.08	2.56	4.92
Baseline+MS-LSA($k = 3$)	96.14	93.94	96.37	95.91	2.07	4.09
Baseline+MS-LSA($k = 3, 5$)	96.11	93.96	96.35	95.87	2.09	4.13
Baseline+MS-LSA($k = 3, 5, 7$)	96.12	93.94	96.34	95.91	2.09	4.09
Baseline+Graph-Conv	96.24	93.99	96.48	**96.00**	2.01	**4.00**
Baseline+Graph-Conv+MS-LSA($k = 3$)	96.22	93.98	96.51	95.93	1.99	4.07
Baseline+Graph-Conv+MS-LSA($k = 3, 5$)	**96.33**	**94.02**	**96.69**	95.97	**1.88**	4.03
Baseline+Graph-Conv+MS-LSA($k = 3, 5, 7$)	96.18	93.99	96.48	95.88	2.01	4.12



Baseline

Baseline + Graph-Conv

Baseline + MS-LSA （k=3）

Baseline + MS-LSA （k=3,5）

Baseline + MS-LSA （k=3,5,7）

Baseline + Graph-Conv + MS-LSA （k=3）

Baseline + Graph-Conv + MS-LSA （k=3,5）

Baseline + Graph-Conv + MS-LSA （k=3,5,7）

Fig. 5. Visualization result of the critical modules in the KGD on KITTI-Road datasets.

As shown in Table V, we analyzed the effects of the Graph-Conv and MS-LSA modules. Firstly, we explored the MS-LSA module and found that the performance improvement varied depending on the kernel size. This variation arises from the capability of the MS-LSA module to extract multi-scale spatial features. It can improve the model's attention to different levels of information, especially in capturing detailed features like lane markings and edges. Secondly, we examined the impact of the Graph-Conv module. The results indicate that adding the Graph-Conv module led to performance gains. This is because graph convolution promotes the exchange of information between distant pixels with similar semantics. This helps achieve coherent road segmentation results. Graph convolution helps the model better understand road topology in segmentation tasks and enhances segmentation accuracy. Finally, we conducted an ablation experiment by combining both modules. The combination "Baseline + Graph-Conv + MS-LSA ($k = 3$, $k = 5$)" achieved the best performance. This improvement is attributed to the MS-LSA module's enhancement of multi-scale detail extraction and the Graph-Conv module's ability to ensure road topology consistency.

Fig. 5 illustrates the visualization results of key modules in KGD. It shows that the Baseline model can extract some road regions. However, it exhibits noticeable detection omissions, particularly in areas with irregular boundaries and complex environments, such as regions shaded by trees (highlighted by the red box). After incorporating MS-LSA (k=3), the model's ability to capture road details is enhanced, particularly in segmentation performance in occluded areas. As the scale increases to k=3,5 or 3,5,7, the model further focuses on multi-scale spatial information. This significantly reduces boundary errors and improves its understanding of road structures. When Graph-Conv is equipped, the model demonstrates stronger boundary detection capabilities. This is particularly evident in complex scenarios with irregular boundaries and noisy backgrounds. The road segmentation in the red-boxed areas is more complete, with a significant reduction

TABLE VI

ABLATION STUDIES WITH DIFFERENT WEIGHTS (α) IN THE LOSS FUNCTION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. "↑" MEANS THE HIGHER THE BETTER. "↓" MEANS THE LOWER THE BETTER

Weight	MaxF↑	AP↑	PRE↑	REC↑	FPR↓	FNR↓
α=0.3	96.20	93.92	96.37	96.04	2.08	3.96
α=0.4	96.23	93.92	96.41	**96.05**	2.05	**3.95**
α=0.5	96.22	93.95	96.52	95.93	1.98	4.07
α=0.6	96.22	93.95	96.53	95.90	1.97	4.10
α=0.7	**96.33**	**94.02**	**96.69**	95.97	**1.88**	4.03
α=0.8	96.21	93.97	96.53	95.89	1.98	4.11
α=0.9	96.20	93.96	96.53	95.88	1.98	4.12

in detection omissions. When combining the Graph-Conv and MS-LSA modules (*i.e.,* k=3,5), the model achieves optimal performance in handling irregular boundary areas. As observed, the road segmentation results in the red-boxed areas are closest to the real-world scenario. When larger scales (k=3,5,7) are used, more details are captured. However, this may lead to slight overfitting or boundary blurring, as seen in the increased redundant labeling of non-road areas.

3) Analysis of KDLoss Component Weighting: In the loss function, we combined FSOhemCELoss [6] and DistillationLoss with specific weights to effectively supervise the model. We tested different values of α to find the optimal weight. As shown in Table VI, we tested α values ranging from 0.3 to 0.9 to determine the optimal weight. When α is low (*i.e.,* FSOhemCELoss has a lower weight), the model tends to focus more on knowledge distillation. In this case, both MaxF and AP are lower, which suggests that the model is not fully using the segmentation loss to optimize performance directly. When α increases to 0.7, MaxF reaches its highest value of 96.33, and the AP, PRE, and FPR also achieve their best values. It indicates that the balance between FSOhemCELoss and DistillationLoss is optimal at this point. However, when α is increased further to 0.8 or 0.9, MaxF and AP slightly decrease, and FPR and FNR increase. This suggests that an excessively high α value causes the model to rely too much on foreground segmentation loss and ignore the knowledge distillation component from the Teacher model.

V. CONCLUSION

In this paper, we proposed a road segmentation network, termed KGD, to balance high segmentation accuracy with computational efficiency in Intelligent Transportation Systems (ITS). The KGD integrates a knowledge distillation framework, Graph-Conv modules, and Multi-scale Lightweight Spatial Attention (MS-LSA) module. The knowledge distillation framework enables a lightweight student network to learn from a pre-trained high-precision teacher network. This process reduces computational requirements while preserving segmentation accuracy. Additionally, to optimize the model's output, we employed the proposed KDLoss to supervise the training process. By incorporating graph convolutional, the proposed network can capture intricate spatial dependencies and accurately model the irregular boundaries of road structures. Furthermore, the MS-LSA module enhances feature representation by focusing on multi-scale spatial information. It captures both local and global contexts necessary for complex segmentation tasks. Experimental results validate the efficacy of the proposed KGD model. It surpasses the traditional road segmentation networks by achieving superior performance in both accuracy and computational efficiency. These advantages make KGD suitable for real-time ITS applications, which include autonomous navigation, smart traffic monitoring, and adaptive traffic management. Given the good performance of the proposed method, it still has room to improve. For example, although KGD performs on standard benchmarks such as KITTI-Road, the generalization capability in extreme scenarios (*e.g.,* heavy occlusion, extreme weather conditions, or poorly illuminated environments) requires further validation. In future work, we will improve robustness by using multi-modal data fusion to handle adverse conditions.

REFERENCES

[1] M. J. AlvarezTheo, G. LeCunAntonio, and M. Lopez, "Road scene segmentation from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 376–389, doi: 10.1007/978-3-642-33786-4_28.

[2] A. Bizzarri, M. Fraccaroli, E. Lamma, and F. Riguzzi, "Integration between constrained optimization and deep networks: A survey," *Frontiers Artif. Intell.*, vol. 7, Jun. 2024, Art. no. 1414707.

[3] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2016, pp. 4885–4891.

[4] M. Teichmann, M. Weber, M. Zöllner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.

[5] S. Gong, H. Zhou, F. Xue, C. Fang, Y. Li, and Y. Zhou, "FastRoadSeg: Fast monocular road segmentation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21505–21514, Nov. 2022.

[6] H. Zhou, F. Xue, Y. Li, S. Gong, Y. Li, and Y. Zhou, "Exploiting low-level representations for ultra-fast road segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9909–9919, Aug. 2024.

[7] J. Fritsch, T. Kühnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 1693–1700.

[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[9] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Aug. 1998.

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[11] Y. Chang, F. Xue, F. Sheng, W. Liang, and A. Ming, "Fast road segmentation via uncertainty-aware symmetric network," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 11124–11130.

[12] P. G. Ravishankar, A. M. Lopez, and G. M. Sanchez, "Unstructured road segmentation using hypercolumn based random forests of local experts," 2022, *arXiv:2207.11523*.

[13] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[16] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[17] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.

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

[19] M. Zhu, K. Han, C. Zhang, J. Lin, and Y. Wang, "Low-resolution visual recognition via deep feature distillation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3762–3766.

[20] X. Wu, R. He, Y. Hu, and Z. Sun, "Learning an evolutionary embedding via massive knowledge distillation," *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2089–2106, Sep. 2020.

[21] S. Zhang, S. Guo, L. Wang, W. Huang, and M. R. Scott, "Knowledge integration networks for action recognition," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12862–12869.

[22] A. Chawla, H. Yin, P. Molchanov, and J. Alvarez, "Data-free knowledge distillation for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3289–3298.

[23] Z. Chen, L. Zhu, L. Wan, S. Wang, W. Feng, and P.-A. Heng, "A multi-task mean teacher for semi-supervised shadow detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5611–5620.

[24] Y. Meng et al., "CNN-GCN aggregation enabled boundary regression for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 352–362.

[25] R. D. S. Mukul, N. Navab, and S. Albarqouni, "An uncertainty-driven GCN refinement strategy for organ segmentation," *Mach. Learn. Biomed. Imag.*, vol. 1, pp. 1–27, Dec. 2020.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[27] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020.

[28] N. Garnett et al., "Real-time category-based and general obstacle detection for autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 198–205.

[29] X. Han, J. Lu, C. Zhao, S. You, and H. Li, "Semisupervised and weakly supervised road detection based on generative adversarial networks," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 551–555, Apr. 2018.

[30] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 230–241, Jan. 2018.

[31] Y. Lyu, L. Bai, and X. Huang, "Road segmentation using CNN and distributed LSTM," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2019, pp. 1–5.

[32] 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.

[33] Z. Chen and Z. Chen, "RBNet: A deep neural network for unified road and road boundary detection," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, Nov. 2017, pp. 677–687.

[34] J.-Y. Sun, S.-W. Kim, S.-W. Lee, Y.-W. Kim, and S.-J. Ko, "Reverse and boundary attention network for road segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 876–885.

[35] Y. Gong et al., "SkipcrossNets: Adaptive skip-cross fusion for road detection," *Automot. Innov.*, pp. 1–17, Apr. 2025.

[36] J. Muñoz-Bulnes, C. Fernandez, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 366–371.

[37] R. Fan et al., "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 1, pp. 225–233, Feb. 2022.

[38] S. Zhang, Z. Zhang, L. Sun, and W. Qin, "One for all: A mutual enhancement method for object detection and semantic segmentation," *Appl. Sci.*, vol. 10, no. 1, p. 13, Dec. 2019.

[39] F. A. Reis et al., "Combining convolutional side-outputs for road image segmentation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8.

[40] M. Oeljeklaus, *An Integrated Approach for Traffic Scene Understanding From Monocular Cameras*. Düsseldorf, Germany: VDI Verlag, 2020.

[41] L. Bai, Y. Lyu, and X. Huang, "RoadNet-RT: High throughput CNN architecture and SoC design for real-time road segmentation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 2, pp. 704–714, Feb. 2021.

[42] M. Bayón-Gutiérrez, M. T. García-Ordás, H. A. Moretón, J. Aveleira-Mata, S. Rubio-Martín, and J. A. Benítez-Andrades, "TEDNet: Twin encoder decoder neural network for 2D camera and LiDAR road detection," *Log. J. IGPL*, May 2024, Art. no. jzae048.

[43] 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.* Cham, Switzerland: Springer, 2020, pp. 340–356.

[44] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[45] A. A. Khan, J. Shao, Y. Rao, L. She, and H. T. Shen, "LRD-Net: Lightweight LiDAR aided cascaded feature pools for free road space detection," *IEEE Trans. Multimedia*, vol. 27, pp. 652–664, 2022.

[46] X. Zhang, Z. Li, X. Gao, D. Jin, and J. Li, "Channel attention in LiDAR-camera fusion for lane line segmentation," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 108020.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2015.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[49] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

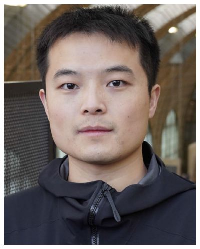
Jianyong Wang is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include crowd counting, computer vision, and deep learning.

Mingliang Gao (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from Sichuan University. He is currently an Associate Professor and the Vice Dean of the School of Electrical and Electronic Engineering, Shandong University of Technology. He was a Visiting Lecturer at the University of British Columbia from 2018 to 2019. He has been the Principal Investigator for a variety of research funding, including the National Natural Science Foundation, China Postdoctoral Foundation, and the National Key Research Development Project. He has published more than 200 journal/conference papers in IEEE, Springer, Elsevier, and Wiley. His research interests include computer vision, machine learning, and intelligent optimal control. He is an Associated Editor of *Expert Systems* and *Network Modeling Analysis in Health Informatics and Bioinformatics*.

Wenzhe Zhai is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo. His research interests include smart city systems, information fusion, crowd analysis, and deep learning. Additionally, he serves as a reviewer for numerous journals, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Neurocomputing*, EAAI, and *Multimedia Systems*.

Imad Rida received the Ph.D. degree in computer science from Normandy University, Rouen, France, in 2017. He is an Associate Professor at the Université de Technologie de Compiègne, Compiègne, France. He is affiliated with the Laboratoire Biomécanique et Bioingénierie UMR, CNRS. His work has contributed to advancements in several areas, including deep learning, feature learning, and denoising, with a particular focus on applications like muscle activity analysis and reducing power line interference in biomedical signals. His research interests lie in machine learning, pattern recognition, and signal/image processing.

Xianxun Zhu received the M.Sc. degree in 2021, through a collaborative program between Nokia Bell Labs Shanghai and Shanghai Second Polytechnic University, focusing on wireless sensing. From 2021 to 2025, he was a joint Ph.D. candidate with Shanghai University and Macquarie University, Australia, where he specialized in multimodal sentiment analysis. He is currently a Post-Doctoral Researcher at Shanghai University, China. He has authored over 20 peer-reviewed papers in leading venues such as AAAI, *Computer Methods and Programs in Biomedicine*, and IEEE COMMUNICATIONS LETTERS. He serves on the editorial boards of IEEE TRANSACTIONS ON MOBILE COMPUTING, the *International Journal of Computer Vision*, *Pattern Recognition*, The *Journal of Supercomputing*, *Computer Methods and Programs in Biomedicine, and Neural Computing and Applications*, and has chaired special sessions at IEEE conferences, notably the IEEE BESC series.



Qilei Li (Member, IEEE) received the M.S. degree from Sichuan University in 2020 and the Ph.D. degree in computer science from the Queen Mary University of London. From June 2022 to April 2024, he was a Machine Learning Scientist at Veritone Inc., where he focused on developing a scalable person search framework for retrieving individuals at different locations and times, as captured by various cameras. His current research interests lie in privacy-aware distributed machine learning, with a particular emphasis on learning domain-invariant knowledge representation from multimodal data captured in diverse environments. His research outcome has been recognized as ESI Highly Cited Paper (Top 1%). Additionally, he serves as an Evaluator for the ELLIS Ph.D. Program.