IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 December 2023, accepted 18 December 2023, date of publication 21 December 2023,
date of current version 27 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3345790

RESEARCH ARTICLE

Lightweight Semantic Segmentation Network Leveraging Class-Aware Contextual Information

XUETIAN XU, SHAORONG HUANG, AND HELANG LAI
Department of Information Administration, Guangdong Justice Police Vocational College, Guangzhou 510520, China

Corresponding author: Xuetian Xu (hmilyxxt@163.com)

This work was supported in part by the China University Industry-Research-Innovation Fund for New Generation Information Technology Innovation Project under Grant 2022IT072, in part by the Guangdong Provincial Ordinary University Featured Innovation Projects under Grant 2023KTSCX295, and in part by the 5th Academic Level Project of Guangdong Judicial Police Officer Vocational College under Grant 2023YB02.

ABSTRACT Balancing model size, segmentation accuracy, and inference speed is a key challenge in image semantic segmentation. This paper introduces a novel lightweight semantic segmentation network, CACNet (Class-Aware Context Network), featuring the innovative Class-Aware Context Enhancement Module (CACEM). CACEM is designed to explicitly intertwine category and context information, addressing the shortcomings of traditional convolutional networks in capturing and encoding inter-category relationships. It operates by normalizing pixel probability distributions via softmax, mapping pixels to categories, and generating new feature maps that accurately encapsulate these relationships. Additionally, the network utilizes multi-scale context information and employs dilated convolutions, followed by upsampling to blend this context with single-channel category information. This process, enhanced by Fourier adaptive attention mechanisms, allows CACNet to capture intricate feature structures and manipulate features in the frequency domain for improved segmentation accuracy. On the Cityscapes and CamVid datasets, CACNet demonstrates competitive accuracies of 70.8 and 74.6 respectively, with a compact model size of 0.52M and an inference speed over 58FPS on GTX 2080Ti GPU platform. This blend of compactness, speed, and accuracy positions CACNet as an efficient choice in resource-constrained environments.

INDEX TERMS Semantic segmentation, class-aware context enhancement module, statistical multi-branch convolution network, Fourier adaptive attention mechanism.

I. INTRODUCTION

Semantic Segmentation involves assigning semantic labels to every pixel in an image, creating a semantically rich image. With the advent of deep learning techniques, deep learning-based semantic segmentation algorithms [1], [2] have progressively replaced traditional algorithms, enhancing segmentation accuracy significantly. The advancement of deep learning frameworks, coupled with the increase in GPU computing power, has allowed for the development of deeper neural network layers. This has significantly bolstered the learning and generalization capabilities of the models, thereby markedly improving the accuracy of semantic segmentation models. However, as the network layers increase, the quantity of model parameters and computation also escalates rapidly, posing challenges for model deployment in resource-constrained devices such as automobiles, mobile phones, and robots. In driving scenarios, to ensure that the driving system can promptly identify potential road safety hazards, the inference speed of the semantic segmentation model must meet real-time requirements [3]. Although deep learning-based semantic segmentation models can achieve satisfactory results in terms of accuracy, they often overlook the numerous limitations of computational and storage resources during the deployment process, which hampers the application of semantic segmentation algorithms in engineering practice. Therefore, the development of semantic segmentation models that simultaneously meet the needs of segmentation accuracy, model size, and inference speed has become a problem that an increasing

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

number of researchers are focusing on. Recent advancements in this field have seen the emergence of lightweight models, such as SqueezeNet [25] and MobileNet [10], which are specifically designed to be efficient in terms of computational resources while maintaining reasonable accuracy. Moreover, accelerated strategies like quantization [26] and network pruning [27] have been proposed to further optimize the inference speed and model size, making them more suitable for deployment in resource-constrained environments. However, these methods often involve trade-offs between accuracy and efficiency, highlighting the complexity of achieving an optimal balance.

The existing lightweight models have made significant strides in processing efficiency and storage management. However, these models still face certain limitations when dealing with complex semantic segmentation tasks, particularly in capturing and encoding intricate relationships between categories. Often prioritizing the reduction of parameter count and computational complexity, these models may compromise the precise capture of contextual information and inter-category interactions. To overcome these limitations, we designed CACNet (Class-Aware Context Network), aiming not only to enhance computational and storage efficiency but also to improve the precision of semantic segmentation and the utilization of contextual information while maintaining a lightweight framework. Guided by multi-level context information as a priori, it optimizes the segmentation results of each category of targets. The experiment demonstrates that this module further significantly improves the model segmentation accuracy. The segmentation accuracy of CACNet on the urban landscape dataset reached 70.2%, the model parameters were only 0.53M, and the inference speed reached 58FPS, meeting real-time requirements. The main contributions of this research are as follows:

(1)Designing the Statistical Multi-Branch Convolution Network (SMBCN) as the backbone network. This network balances the advantages of multi-branch and single-branch network structures, reducing the complexity of the model and improving the segmentation accuracy of the model.

(2)Proposing a novel Class-Aware Context Enhancement Module, which explicitly fuses class information and context information together. This module can better represent the relationship between classes and improve the segmentation accuracy of the model. The module designs a Fourier adaptive attention mechanism to process features in the frequency domain, which can more finely control various components in the features, thereby enhancing the performance of the model.

(3)The CACNet model has achieved superior performance compared to existing methods on the Cityscapes and CamVid datasets.

II. RELATED WORKS

A. DEEP LEARNING-BASED SEMANTIC SEGMENTATION

Within the computer vision domain, convolutional neural networks (CNNs) [2] have become an essential tool due to their suitability in local feature extraction, which is particularly advantageous given the distinct structure of image data. The introduction of the fully convolutional neural network (FCNN) [4] marked a significant milestone, pioneering the use of CNNs in end-to-end training for image semantic segmentation tasks. Subsequent deep learning-based image semantic segmentation methodologies have largely drawn upon the design principles of the FCNN, making incremental improvements in areas such as basic convolution modules, output modules, and feature extraction modules. UNet [5], PSPNet [6], CCNet [7], and the Deeplab [8] have become important benchmark models in the field of image semantic segmentation. In DeepLab [8], deep convolutional neural networks are combined with probabilistic graphical models (CRF), and the use of fully connected CRFs allows for the capture of global information in images, thereby enhancing segmentation precision. DeepLab v2 [28] adopts atrous convolution, expanding the receptive field of the convolution kernel. DeepLab v3 [29] and DeepLab v3+ [30] abandoned probabilistic graphical models in favor of the ASPP module, achieving end-to-end training and testing with deep learning. DeepLab v3+ made several innovations and improvements to the ASPP module.

In recent years, the computer vision landscape has seen Visual Transformers increasingly outperform earlier CNNs in tasks such as image classification and object detection. They have also shown promise in semantic segmentation tasks [9], though some limitations persist. The need to partition the image into smaller blocks for processing with Transformers can inhibit the acquisition of precise segmentation results, as the network is essentially predicting a sequence of image blocks instead of directly predicting pixels. Furthermore, due to architectural constraints, Transformer models typically possess parameters in the order of hundreds of megabytes, a scale far from the lightweight networks.

Although existing deep learning-based semantic segmentation models have greatly improved in accuracy, the problems of high computational load and a large number of parameters persist. This hinders real-time processing of the models and complicates deployment on mobile platforms. Therefore, research on real-time semantic segmentation networks has become a focal point.

B. LIGHTWEIGHT MODELS

Aiming for efficient model deployment and real-time application, researchers have proposed several lightweight real-time semantic segmentation networks. These often employ decomposed convolution and depthwise separable convolution [10] as replacements for standard convolution, thus reducing parameter count and computational load. The depth and width of the network are meticulously regulated to avoid the substantial computational burden that originates from operating multi-channel feature maps and convolution kernels. To counteract the decreased learning capability and generalization due to the reduced model parameters, these networks

necessitate carefully engineered basic modules. Depending on their structure, lightweight semantic segmentation networks can be categorized as either single-path or multi-path.

Single-path networks entail one input end, where the image, once fed into the network, follows a single path until it reaches the output end. A representative of this type is the ENet [11], which features a lightweight design expressed in three facets: (1)Restricting the number of feature map channels to a maximum of 128, hence reducing the number of convolution kernels and computational requirements during inference; (2)Applying three downsampling operations in the initialization module and the front end of the encoder to rapidly diminish the feature map's resolution, thereby minimizing computational requirements during inference; (3) Employing decomposed convolution instead of standard convolution to further decrease the number of parameters. ERFNet [12], LEDNet [13], STDC [32], and NDNet [14] have developed unique approaches in single-path network design, further enriching the landscape. Moreover, certain studies have elaborately discussed the new backbone and attention Module, offering fresh perspectives for designing lightweight model structures [33], [34].

On the other hand, multi-path networks have multiple inputs, where the input image is simultaneously processed through multiple paths, which are then merged and outputted as a segmentation result. Examples of such networks include BiSeNet [15], ContextNet [16], ICNet [17], and BiSeNet V2 [31]. These multi-path networks use distinct path depths to extract different levels of image features, leveraging the advantages of each depth for improved accuracy and detail recognition.

In sum, both single-path and multi-path lightweight semantic segmentation networks present unique methodologies and innovations, thus contributing to advancements in the field of computer vision and semantic segmentation. Nonetheless, the quest for the optimal balance between accuracy and computational efficiency continues [19], [20].

Current lightweight semantic segmentation models, despite being designed for efficiency and reduced computational costs, face certain limitations when dealing with more complex semantic segmentation tasks. The primary issues with these models pertain to two aspects: (1) Insufficient capture of inter-class relationships: While optimized for parameter count and computational complexity, these models often overlook the precise capture of complex inter-class relations, which is crucial for the accuracy of semantic segmentation tasks. (2) Inadequate utilization of contextual information: In the effort to reduce computational resources, these models may compromise on the in-depth use of contextual information, which is particularly important for semantic segmentation in scenes with complex backgrounds and diverse settings. To address the aforementioned issues, we have designed the Class-Aware Context Network (CACNet). The core idea of CACNet is to integrate a Class Aware Context Enhancement Module(CACEM) with a Statistical Multi-Branch Convolutional Network (SMBCN), aiming to overcome the limitations of existing lightweight models and further improve the accuracy of semantic segmentation.

III. METHODOLOGY

A. OVERALL MODEL ARCHITECTURE

In the realm of semantic segmentation networks, the role of the top layer, the classifier, is to amalgamate the deep features learned via the convolutional network and to produce pixel-wise prediction outcomes. Contrary to the output of typical convolutional layers within the network, the output of the classifier signifies not only the segmentation result but also a feature map whose channels equal the number of classes, where each pixel point contains explicit class information [21]. In an endeavor to extract context information from intermediate feature maps, it is common to employ strategies for expanding the receptive field and for learning multi-scale features. By performing convolution operations to combine multiple local features, the resultant feature map amalgamates a plethora of context information to attain accurate segmentation. Nevertheless, considering that the feature maps yielded by the intermediate convolutional layers are not the final classification results, despite each pixel point amalgamating abundant information, the interpretation of this information remains unclear. In certain cases, this may lead to misdirection of the segmentation outcome. To make better use of the context information contained in the feature map, we can utilize purer class information to steer the context information, thereby providing the network with a more unambiguous learning direction. In pursuit of enhancing the model's segmentation outcomes directly, this work introduces the Class Aware Context Enhancement Module (CACEM), and designs a novel lightweight semantic segmentation network termed CACNet (Class Aware Context Network) on the basis of this module. By optimizing the output of the classifier on a channel-by-channel, class-by-class basis using the rich context information present in the network's high-level features, the learning objectives of the model become more focused.

Figure 1 provides a schematic representation of the CACNet network structure, demonstrating its asymmetric encoder-decoder architecture. The convolutional layers constituting the encoder comprise convolutional modules formed by the alternative stacking of dilated and decomposed convolutions. The input image, subjected to three successive downsampling operations, yields a resolution reduced to one-eighth of its original value. Concurrently, the number of channels progressively escalates to 16, 64, and 128, respectively.

As the depth of the encoder network increases and the number of channels amplifies, the information encompassed within the output feature map likewise expands. Furthermore, the range of adjacent pixel context information aggregated by each pixel point broadens. Consequently, the first context branch is derived from the deeper layer of the encoder network.

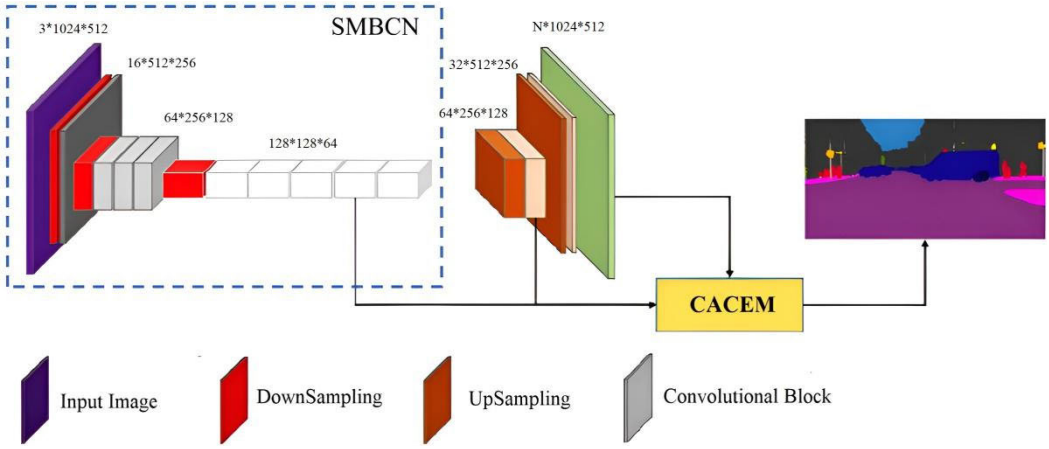

FIGURE 1. Schematic of CACNet network structure. The architecture features an input image processed by the statistical Multi-Branch convolutional network (SMBCN), which reduces dimensionality through sequential convolutions. The resulting features are refined and upsampled before being enhanced by the Class-Aware context enhancement module (CACEM) for semantic segmentation, leading to the final segmented output.

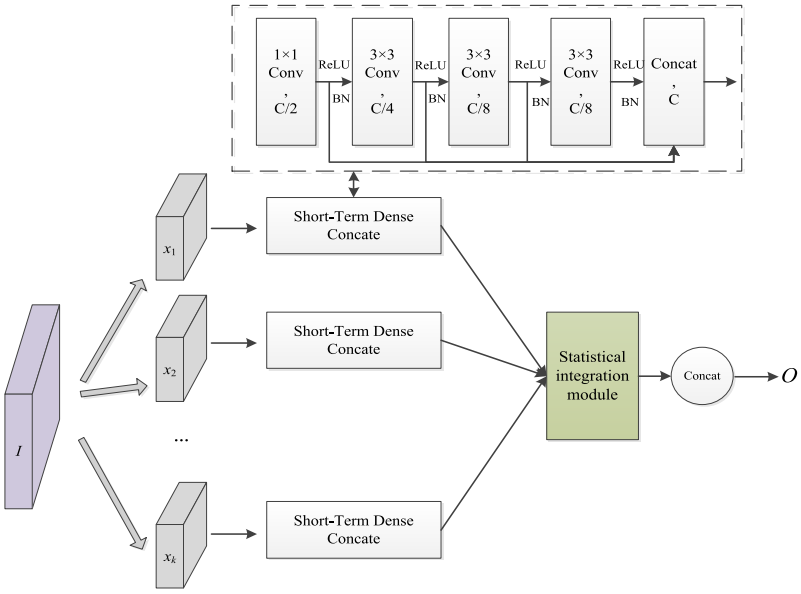

FIGURE 2. Structure of the statistical multi-branch convolution network. The input feature map I is divided into K branches. Each branch processes through a series of convolutions, reducing the channel dimensions successively and producing intermediate feature maps $x_1, x_2, \ldots, x_K$. The exact shapes of these intermediate features x_1 through x_K are detailed in Table 1, showing the dimensionality reduction at each convolutional step. These feature maps are subsequently integrated using a statistical module to compute the final output feature map O that delivers enhanced segmentation detail and accuracy.

The second branch emanates from the first upsampling module of the decoder. Utilizing the feature map from this branch proffers two advantages: 1) the introduction of the feature fusion module within the first upsampling module yields output features with more discernible edge details, and 2) the feature map of the initial upsampling exhibits a degree of channel compression. Given that the resolution of the feature map is not excessively high, the computational load it imparts is relatively manageable, striking a balance between detail and efficiency.

B. BACKBONE NETWORK BASED ON STATISTICAL MULTI-BRANCH CONVOLUTIONAL NETWORK

Addressing the issue of large model sizes prevalent in multi-branch network structures and low segmentation precision inherent in single-branch network structures, this work presents a statistical multi-branch convolution network (SMBCN) as the backbone network.

The Statistical Multi-branch Convolution Network (SMBCN) begins processing the input feature map I, which has dimensions $3 \times H \times W$, by dividing it into K distinct

branches for parallel processing. The processing methods for different branches are identical, using the same layers and parameters (such as kernel size, stride, padding, etc.) in their design. Even though each branch performs similar operations, they can process different parts of the feature map simultaneously. This parallel processing allows for increased efficiency and diversity in feature extraction while maintaining a moderate model size. Moreover, although each branch performs similar operations, they can focus on features of different scales or types. This multi-scale feature representation is crucial for enhancing the detail and accuracy of image segmentation.

Focusing on a single branch as an example, the transformation begins with a 1×1 convolution that contracts the channel dimension from C to $C/2$, thus producing an intermediate feature map x_{k1}. This map is further processed by a series of two 3×3 convolutions, which sequentially reduce the number of channels first to C/4, resulting in x_{k2}, and subsequently to $C/8$, creating x_{k3}. The final step in the branch involves a 3×3 convolution that maintains the channel dimension at C/8, completing the series of transformations for this branch with x_k. The operations for each branch can be described by the following equations:

$$x_{k1} = \delta(\beta(w_{k1} * I)) \tag{1}$$

$$x_{k2} = \delta(\beta(w_{k2} * x_{k1})) \tag{2}$$

$$x_{k3} = \delta(\beta(w_{k3} * x_{k2})) \tag{3}$$

$$x_k = \delta(\beta(w_k * x_{k3})) \tag{4}$$

where δ is the *ReLU* activation function [22], β is batch normalization [23]. Each branch's convolutional operations are designed to capture a range of features from the input, providing a rich, multi-scale feature representation.

Following the short-term dense concatenate, each branch's feature map x_k is concatenated in the statistical integration module. This module processes the features from all branches to produce the final output feature map O, which is then used for precise semantic segmentation.Within the statistical integration module, we employ mean pooling, variance pooling, and max pooling submodules to calculate the mean, variance, and max values of the series $x_1, x_2, \ldots, x_K$ respectively. Upon concatenating these three results, the final output feature O is obtained:

$$O = Concat[\mu, \sigma^2, m] \tag{5}$$

The formula of the mean pooling submodule is:

$$\mu = \frac{1}{K} \sum_{k=1}^{K} x_k \tag{6}$$

The formula of the variance pooling submodule is:

$$\sigma^2 = \frac{1}{K} \sum_{k=1}^{K} (x_k - \mu)^2 \tag{7}$$

The formula of the max pooling submodule is:

$$m = \max_{k=1}^{K} x_k \tag{8}$$

TABLE 1. Detailed parameters of SMBCN.

Layer	Input Size	Output Size	Number of Parameters
Input	3x1024x512	3x1024x512	0
Branch operation	3x1024x512	K branches, each branch size is 3x1024x512	0
K branches, each branch size is	3x1024x512	K branches, each branch size is16x512x256	(113*16+16)
3x3 Convolution (Reducing to C/4)	16x512x256	K branches, each branch size is64x256x128	(3316*64+64)
3x3 Convolution (Reducing to C/8)	64x256x128	K branches, each branch size is128x128x64	(3364*128+128)
3x3 Convolution (Channel number remains the same)	128x128x64	K branches, each branch size is128x128x64	(33128*128+128)
Statistical Integration Module	K branches, each branch size is128x128x64	128x128x64	0

The design ethos underlying the SMBCN backbone network is encapsulated by four core characteristics: (1) The network navigates features in a wider space by allowing each branch's convolution operation to have diverse parameters, thereby enabling the network to probe features within a larger space. (2) Through the statistical integration module, the network extracts a greater quantity of useful features, by statistically analyzing the results of multi-branch convolution operations, thus enhancing the network's representation capability and image segmentation precision. (3) The network maintains a moderate size, circumventing the issues of low segmentation precision characteristic of original single-branch network structures, and the problem of large model sizes inherent in original multi-branch network structures. (4) The network augments the model's segmentation accuracy by harnessing the statistical integration module to extract more beneficial features, thereby bolstering the network's representational power and the precision of image segmentation.

Table 1 provides a detailed description of the SMBCN backbone network.

C. CLASS-AWARE CONTEXT ENHANCEMENT MODULE (CACEM)

In semantic segmentation tasks, the fusion of context information and category-specific details is vital. Traditional convolutional neural networks, however, face a significant challenge—they lack an explicit mechanism for capturing and encoding inter-category relationships, leading to an output feature map that does not explicitly differentiate between categories.

To address this limitation, we propose the Class-Aware Context Enhancement Module (CACEM), which is engineered to intertwine category and context information explicitly. The CACEM operates by first normalizing the

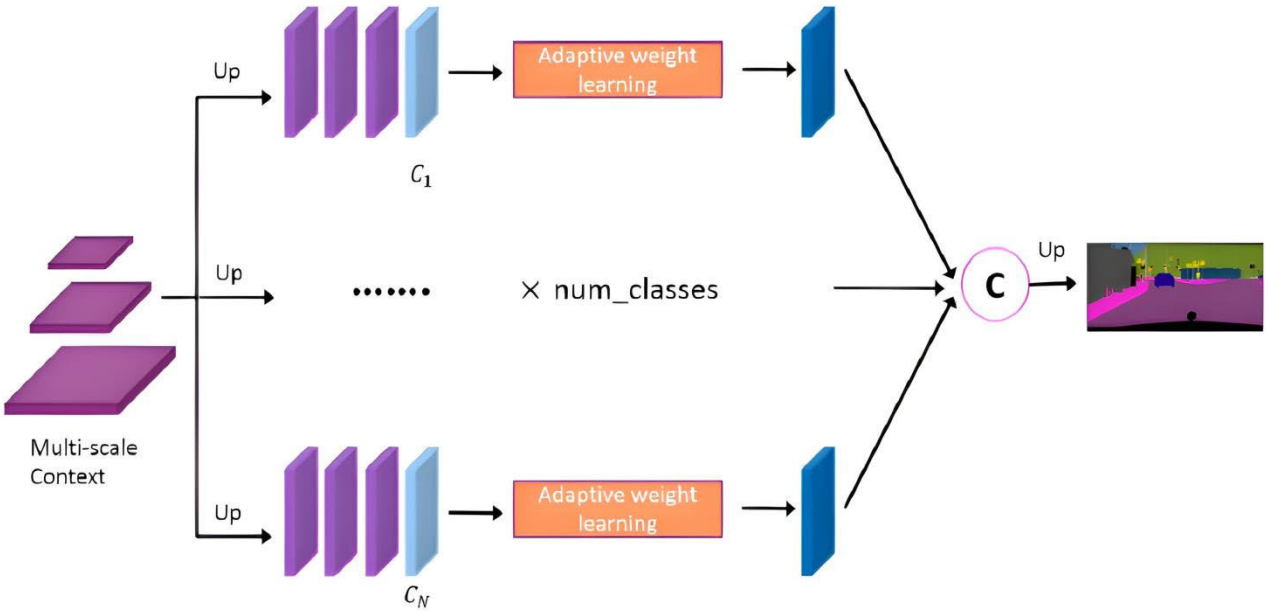

FIGURE 3. Schematic diagram of CACEM Module. The diagram showcases the CACEM in action, processing multi-scale context information through adaptive weight learning for each category to generate category-specific feature maps. These are then integrated and upsampled to produce a semantically enriched segmentation output, demonstrating CACEM's capability for enhancing segmentation accuracy by leveraging category-discriminative features.

probability distribution of each pixel's network output via the softmax function, mapping each pixel to its respective category based on the derived probability distribution. Each category's pixels utilize the associated feature map as input, and this input is subsequently processed through a compact convolutional network to yield a new feature map that more accurately encapsulates inter-category relationships. Such a strategy permits better utilization of category information, thus enhancing the model's segmentation accuracy.

As depicted in Figure 3, let us assume a current classification task consisting of N categories. The preliminary segmentation outcomes of the classifier output are broken down into N single-channel feature maps. Since each pixel's category is determined by the index of the maximum value along the channel dimension, every segmentation result channel contains information for a single category. By employing these single-channel feature maps replete with category information, the model is guided to extract more targeted information from the context, specifically, information that contributes to accurate classification of the current category.

To access multi-scale context information, multi-level features are extracted from the backbone network. Dilated convolution is subsequently employed to further encode these features, which are then upsampled to match the resolution of the category information. The result is divided into N branches and combined with the N single-channel features possessing category information. Through adaptive weight learning, this procedure outputs the enhanced segmentation result for the current category. Finally, N enhanced feature maps are generated, representing the segmentation result.

Let's assume the output of the segmentation model is a tensor O of shape $H \times W \times N$, where H and W represent the image's height and width, and N is the number of categories. The category of each pixel is derived via the argmax function:

$$O' = \text{argmax (O,axis} = 2) \tag{9}$$

Subsequently, O' is expanded into N binary maps, each representing a unique category. Let O_i denote the binary map of the i-th category:

$$O_i = O' == i \tag{10}$$

This equation produces a binary map O_i in which only pixels of category i are marked as 1, while all other pixels are assigned a value of 0.

Assuming the output of the backbone network is a list, with each element being a feature map F_i of shape $H_i \times W_i \times C_i$. A dilated convolution operation is performed on feature F_i, followed by an upsampling to match the resolution of O_i:

$$F_i' = Upsample\,(DilatedConv\,(F_i)) \tag{11}$$

Subsequently, for each category i, we integrate its corresponding binary map O_i and feature map F_i' through concatenation:

$$Z_i = Concat\,\left(O_i, F_i'\right) \tag{12}$$

Z_i is then processed via an adaptive weight learning module to derive the enhanced feature map E_i. To thoroughly integrate the features corresponding to each category, we utilize the Fourier transform to convert features from the spatial to the frequency domain. This conversion facilitates the capturing of patterns of varying scales and orientations on a

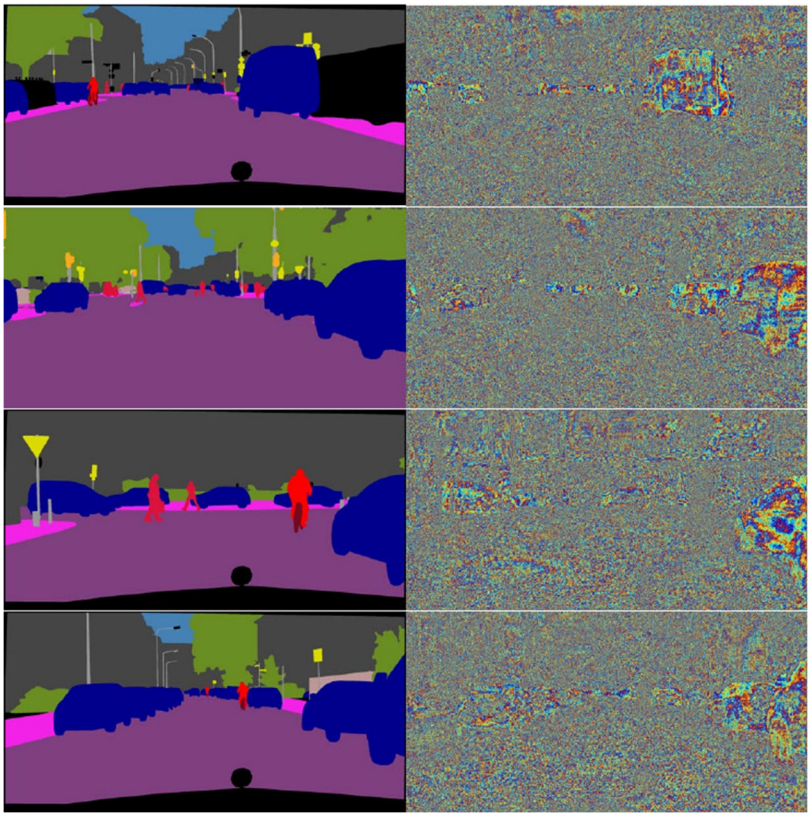

FIGURE 4. Visualization of car category segmentation and feature activation. The figure contrasts semantic segmentation results with the activation maps for the car category, illustrating the Class-Aware context enhancement module (CACEM)'s accuracy in segmenting and highlighting cars within the images.

global scale, thereby enriching the understanding of complex structures in the image. The rationale for this approach is that high-frequency components signify details and boundaries in the image, while low-frequency components depict global and smooth parts of the image. Therefore, by manipulating features in the frequency domain, we can exert finer control over the various components in the features, thereby boosting the model's performance.

In particular, the Fourier Adaptive Attention mechanism is utilized to process Z_i, producing the frequency domain feature W_{ij}. Initially, a discrete Fourier transform [24] is performed on the input feature Z_i to obtain its frequency domain representation F_{ij}:

$$F_{ij} = FFT\ (Z_i) \tag{13}$$

We then design an attention mechanism to procure three representations—query (Q), key (K), and value (V)—via a linear transformation of the frequency domain representation F_{ij}:

$$\begin{aligned} Q_{ij} &= W_Q * F_{ij} \\ K_{ij} &= W_K * F_{ij} \\ V_{ij} &= W_V * F_{ij} \end{aligned} \tag{14}$$

where W_Q, W_K, and W_V are learnable weight matrices.

The intricate interaction between the query and the key is captured through the Gaussian kernel function:

$$O_{ij} = \exp\left(-\left\|Q_{ij} - K_{ij}\right\|^2 \Big/ \left(2 * \sigma^2\right)\right) \tag{15}$$

where σ is a learnable parameter.

The softmax function is employed to normalize the result of this Gaussian kernel function, and the Fourier inverse transform is performed to convert the frequency domain features back to the spatial domain:

$$W_{ij} = IFFT\left(Soft\max\left(O_{ij}\right) * V_{ij}\right) \tag{16}$$

Ultimately, the enhanced feature maps of all categories are weighted and stacked to deliver the final segmentation output.

Figure 4 provides a visualization of the single-channel result of the category representing a car in the feature map of the classifier output. It can be observed that aside from the pixels located at the car's position, the brightness of pixels elsewhere is relatively low. The congruence of the bright spots in this channel map with the shape and position of the car in the label map directly determines the model's prediction accuracy for the car category. From a theoretical perspective, the visualization phenomenon presented in Figure 4 is primarily attributed to the design and operation of the Class-Aware Context Enhancement Module (CACEM).

CACEM, by mapping each pixel to its associated probability distribution, enables an explicit fusion of category-specific and contextual information. This design ensures that the model, when performing semantic segmentation, can rely not just on the inherent information from each pixel, but also on the category information from surrounding pixels. Throughout this process, category information is explicitly encoded into single-channel feature maps corresponding to each category. Consequently, this empowers the model to extract more precise contextual information beneficial for the accurate classification of the current category.

IV. EXPERIMENTS AND ANALYSIS

The efficacy of the CACEM and the superior performance of CACNet were corroborated through a series of experiments conducted on two widely accepted public datasets, namely Cityscapes and CamVid.

A. EXPERIMENTAL SETUP

Training of the CACNet was conducted on a hardware platform comprising dual NVIDIA GTX 2080Ti GPUs, complemented by an Intel Xeon CPU at 2.20GHz and 128GB of system RAM. The models were implemented and trained using the TensorFlow and Keras frameworks.

Initially, the backbone network, SMBCN, was pre-trained on the Cityscapes dataset for 300 epochs, which equates to approximately 32 hours of computational time. This initial training phase was crucial to establish a baseline for feature extraction from the dataset. The training process was governed by the following update rule for the weights $\mathbf{W}$ at each epoch t:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \nabla L\left(\mathbf{W}_t\right) \tag{17}$$

where η_t is the learning rate at epoch t and $\nabla \mathrm{L}\left(\mathbf{W}_t\right)$ is the gradient of the loss function L with respect to the weights $\mathbf{W}_t$.

Subsequently, we integrated the Class-Aware Context Enhancement Module (CACEM) into the pre-trained ERFNet. During this phase, only the parameters of the CACEM module were trained, ensuring that the module could effectively learn to capture and utilize the context information specific to each category. The learning rate for this stage was set at $1e^{-4}$, following the.update rule:

$$\mathbf{W}_{CACEM_{t+1}} = \mathbf{W}_{CACEM_t} - \eta \nabla L\left(\mathbf{W}_{CACEM_t}\right) \tag{18}$$

Finally, the entire CACNet architecture underwent a fine-tuning process. The learning rate was sustained at 1^{e-4}, and the network was trained for an additional 100 epochs. This fine-tuning step was instrumental in optimizing the interaction between the backbone and the CACEM, thereby enhancing the network's overall segmentation accuracy.

Poly learning rate decay strategy was applied across all phases, with the learning rate η_t adjusted according to:

$$\eta_t = \eta_{initial} \times \left(1 - \frac{t}{T}\right)^{power} \tag{19}$$

TABLE 2. Ablation experiment results on the Cityscapes dataset.

Model	mIoU(%)	Params(M)	FloPs(G)	FPS
ERFNet	65.5	0.56	4.09	57
+SMBCN	67.3	0.48	4.11	60
+CACEM	71.4	0.52	4.66	58

TABLE 3. Ablation experiment results on the CamVid dataset.

Model	mIoU(%)	Params(M)	FloPs(G)	FPS
ERFNet	68.2	0.52	3.71	54
+SMBCN	70.5	0.49	3.53	63
+CACEM	71.7	0.50	3.61	59

where $\eta_{initial}$ is the initial learning rate, t is the current epoch, T is the maximum number of epochs, and *power* is set to 0.9, dictating the decay rate. This strategy ensures that the learning rate decreases polynomially with the number of epochs, enabling finer adjustments to the weights as the network converges.

Training focused exclusively on the finely annotated images within the Cityscapes dataset, deliberately excluding the 20,000 coarsely annotated images to ensure high-quality data input. Prior to feeding these images into the network, rigorous data preprocessing steps were undertaken. This included normalization of image pixel values, augmentation techniques like rotation and scaling to enhance generalization, and cropping to maintain consistency in image size and aspect ratio. The overall network training, enriched by these preprocessing strategies, was conducted over 300 epochs. Each epoch meticulously processed input images with a resolution of 1024×512 pixels, ensuring that the network received well-prepared and standardized data at each step. This approach maximized the efficacy of the training process, allowing the network to learn more effectively from the intricately annotated data.

B. ABLATION EXPERIMENTS

Employing ERFNet as the baseline backbone network, we tested the efficacy of our primary network, SMBCN, and the CACEM module. Metrics for evaluation included segmentation accuracy, model parameters, computational requirements, and inference speed. The experimental findings are delineated in the subsequent table:

Remarkable improvements were observed with the SMBCN backbone compared to the ERFNet backbone across both Cityscapes and CamVid datasets. In terms of the Cityscapes dataset, mIoU experienced an enhancement by 1.8 percentage points, model parameters were reduced by 0.08M, computations increased by 0.02G, and inference speed advanced by 3FPS. For the CamVid dataset, mIoU improved by 2.3 percentage points, model parameters

TABLE 4. Ablation study results on the cityscapes Dataset.

Model	mIoU(%)	Params(M)	FloPs(G)
Baseline (SMBCN only)	67.3	0.48	4.11
Category Mapping Only	68.5	0.49	4.20
Feature Integration w/o Adaptive Learning	69.8	0.50	4.25
Adaptive Learning w/o Fourier Transform	70.2	0.51	4.30
Without Fourier Transform	70.6	0.52	4.35
High-Frequency Components Only	69.5	0.52	4.35
Low-Frequency Components Only	68.9	0.52	4.35
Self-Attention (SA)	70.9	0.53	4.40
Convolutional Block Attention Module (CBAM)	71.0	0.53	4.42
Squeeze-and-Excitation (SE) Block	70.8	0.53	4.41

decreased by 0.03M, computation decreased by 0.18G, and the inference speed increased by 9FPS.

Further accuracy enhancements were attained by appending the CACEM module to the SMBCN backbone network, with minimal impact on model size. For the Cityscapes dataset, the integration of the CACEM module led to an mIoU increase by 4.1 percentage points, model parameters increment by 0.04M, computation augmentation by 0.55G, and an inference speed decrement of 2FPS. Nevertheless, the inference speed still achieved 58FPS, fulfilling the real-time requirement of over 30FPS for driving scenarios. Similarly, for the CamVid dataset, the CACEM module's addition led to an mIoU increase by 1.2 percentage points, model parameters increment by 0.01M, computation increment by 0.08G, and an inference speed decrement of 4FPS. However, the inference speed still reached 59FPS, meeting the real-time requirement of over 30FPS for driving scenarios.

The CACEM module's parameters predominantly reside within the context information extraction module, and the number of feature map channels is contingent upon the backbone network's feature map channels. During the experiments, the two context information branches derived from the backbone network had feature map channel numbers of 128 and 64, and resolutions of 1/8 and 1/4, respectively. The context information extraction module executes feature compression operations on the two features, employing 3×3 and 1×1 convolutions to reduce the number of feature map channels to 32 and 16, subsequently combining them with single-channel category features to form one input in the category feature enhancement branch of the CACEM module. Despite the presence of multiple convolutional layers, the increased parameters were minimal since each convolutional layer processes a limited number of feature map channels, thereby ensuring the model remains lightweight.

We further designed ablation experiments to assess the function of the submodules within CACEM, and to comprehend the effects of the Fourier transform and attention mechanisms on the overall performance of our semantic segmentation model. The Statistical Multi-Branch Convolution Network (SMBCN) served as our baseline, onto which we incrementally introduced components of CACEM to evaluate their individual and combined impacts. The experimental results are presented in Table 4.

In the comprehensive ablation studies conducted to dissect the Class-Aware Context Enhancement Module (CACEM) and its constituents, we witnessed a nuanced interplay of performance impacts across the different configurations tested on the Cityscapes Dataset. The baseline model, which operates solely on the Statistical Multi-Branch Convolution Network (SMBCN), achieved a mean Intersection over Union (mIoU) of 67.3%.

When the Category Mapping Only approach was adopted, where pixels were mapped to categories based solely on softmax-normalized probabilities, we observed a modest improvement in mIoU to 68.5%. This increment highlights the significance of precise category mapping in enhancing segmentation accuracy. Further complexity was added with Feature Integration without Adaptive Learning, leading to an mIoU of 69.8%. This suggests that integrating multi-scale context information, even without adaptive learning, contributes positively to the model's discernment capabilities. The Adaptive Learning without Fourier Transform variant marked an mIoU of 70.2%, underscoring the value of adaptively weighted features in capturing the intricacies of the semantic classes, despite the absence of frequency domain processing.

Interestingly, the exclusion of the Fourier Transform from the CACEM led to a higher mIoU of 70.6%. This could indicate that while the Fourier Transform contributes to feature enhancement, its absence does not significantly detract from the model's performance, possibly due to the redundancy of information in the spatial domain. Isolating frequency components revealed differential effects: High-Frequency Components Only registered a decrease in mIoU to 69.5%, while Low-Frequency Components Only further declined to 68.9%. This delineates the high-frequency components' role in capturing fine-grained details that low-frequency components may miss, thus being more consequential for segmentation accuracy.

TABLE 5. Comparative experimental results on the cityscapes dataset.

Method	Road	Sid	Bui	Wal	Fen	Pol	Tli	Tsi	Veg	Ter	Sky
ENet	92.7	78.1	85.3	44.3	41.5	52.1	61.8	64.2	85.9	63.7	88.9
CGNet	93.5	79.4	86.7	45.6	43.2	53.5	62.9	65.7	87.2	64.9	90.2
ERFNet	94.3	80.6	87.9	46.9	44.8	54.7	64.1	67.1	88.5	65.9	91.4
NDNet	94.9	81.7	88.9	47.9	46.2	55.8	65.1	68.3	89.6	66.7	92.5
BiseNet	95.3	82.3	89.5	48.6	47.1	**58.6**	66.0	69.2	90.3	67.3	93.1
ICNet	96.0	82.9	90.0	49.2	47.9	57.1	67.7	70.0	91.0	68.0	93.7
BiSeNet V2	97.8	82.9	90.5	50.2	48.7	56.9	**68.5**	68.7	92.0	69.8	93.6
PIDNet	97.5	83.0	**91.5**	49.8	49.1	57.0	68.0	69.0	91.8	**70.5**	93.9
RegSeg	**98.4**	83.2	90.7	50.5	48.9	56.7	69.0	68.9	92.3	69.9	93.8
CACNet	98.2	**83.4**	91.2	**50.8**	**49.1**	57.4	67.2	**70.3**	**92.4**	70.2	**94.8**

Method	Per	Rid	Car	Tru	Bus	Tra	Mot	Bic	MIoU
ENet	74.8	59.1	89.1	40.6	56.2	44.1	52.4	64.2	65.2
CGNet	75.6	60.3	90.1	41.7	57.3	45.2	53.5	65.3	66.4
ERFNet	76.3	61.5	91.1	42.7	58.3	46.2	54.6	66.4	67.5
NDNet	77.0	62.6	92.0	43.6	59.3	47.1	55.6	67.4	68.5
BiseNet	77.6	63.6	92.8	44.4	60.2	45.9	56.5	68.3	69.3
ICNet	78.3	64.5	93.5	**45.1**	61.0	**48.6**	56.3	69.1	70.0
BiSeNet V2	78.4	**65.2**	94.7	44.2	60.6	46.3	57.0	69.0	70.3
PIDNet	**80.1**	64.3	93.8	44.5	60.2	47.4	55.8	**69.8**	70.4
RegSeg	77.9	65.0	94.6	43.5	59.8	48.3	56.1	68.5	70.3
CACNet	79.8	64.2	**95.1**	44.7	**61.3**	48.2	**57.5**	69.3	**70.8**

In the ablation experiments, we evaluated the role of attention mechanisms relative to the performance of CACE). The Self-Attention (SA) mechanism resulted in an mIoU of 70.9%, highlighting its effectiveness in capturing global dependencies within the data. The Convolutional Block Attention Module (CBAM) showed a slight improvement with an mIoU of 71.0%, suggesting the benefits of addressing both channel and spatial features for a comprehensive feature analysis. The Squeeze-and-Excitation (SE) Block attained an mIoU of 70.8%, which supports the significance of channel-wise feature adjustments. Although these attention mechanisms each have their strengths, our CACEM's use of the Fourier Adaptive Attention mechanism outperforms these standard approaches. CACEM's strategy to process features in both spatial and frequency domains allows for a more detailed and context-aware segmentation. This innovative approach provides a clear advantage, as it not only enhances the distinctiveness of category-specific features but also maintains a high segmentation accuracy, thereby reinforcing the effectiveness of CACEM in complex segmentation tasks.

C. COMPARATIVE EXPERIMENTAL RESULTS

To verify the superior performance of CACNet, experiments were conducted on the Cityscapes and CamVid street scene datasets, and comparisons were made with other lightweight semantic segmentation networks. The detailed results are as follows:

1) CITYSCAPES DATASET

On the Cityscapes dataset, comparisons were made with advanced methods such as ENet [9], CGNet [16], ERFNet [10], NDNet [12], BiseNet [13], ICNet [17], BiSeNet V2 [31], PIDNet [35], RegSeg [35], etc. The specific results are as follows TABLE 4:

In the evaluation of semantic segmentation models on the Cityscapes dataset, as presented in Table 5, the Class-Aware Context Network (CACNet) exhibited exemplary performance, achieving the highest mean Intersection over Union (mIoU) score of 70.8%. This performance was not only consistent across the majority of individual categories but also demonstrated superior precision in 9 out of the 19 classes, highlighting the model's robustness and adaptability across diverse urban scenes. Particularly notable is the model's performance in the 'Sky' and 'Motorcycle' categories, where CACNet achieved 70.3% and 61.3% accuracy, respectively, surpassing the next best-performing model by 0.9% and 0.5%. Such fine-grained differentiation is crucial in urban landscape parsing where structures like buildings and vegetation are prevalent.

Moreover, in direct comparison to the intricate multi-branch structures of ICNet and BiseNet, which command a considerable parameter volume of over 10M, CACNet's more modest parameter count of approximately 0.52M represents a breakthrough in efficient network design. This significant reduction in complexity does not come at the cost of performance; rather, it emphasizes the strategic design choices made in CACNet that avoid redundant computations and focus on feature quality over quantity.

These quantifiable advances are a testament to the efficacy of the model's architecture, particularly the integration of the CACEM, which facilitates a nuanced understanding of class-specific context. Such context awareness is pivotal in

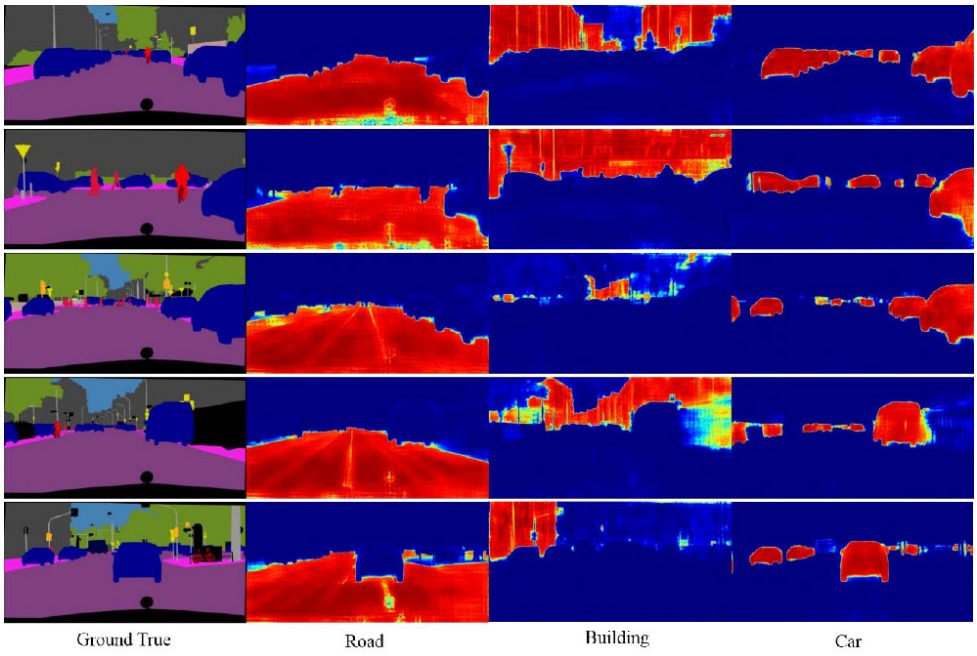

FIGURE 5. Visualization of category features on the cityscapes dataset.

TABLE 6. Comparative experimental results on the camvid dataset.

Method	Bui	Tre	Sky	Car	Sig	Roa	Ped	Fen	Pol	Sid	Bic	mIoU
ENet	82.1	75.3	89.1	86.9	35.2	92.5	64.1	50.1	42.2	83.1	60.8	69.2
CGNet	83.6	76.5	90.0	88.0	36.3	93.3	65.6	51.3	43.3	84.2	62.3	70.4
ERFNet	84.9	77.6	90.7	88.9	37.3	93.9	66.9	52.4	44.3	85.2	63.7	71.4
NDNet	85.9	78.6	91.3	89.7	38.2	94.4	69.9	53.4	43.2	86.1	64.8	72.3
BiseNet	87.1	79.6	92.7	90.8	39.2	95.7	68.8	53.9	46.1	87.0	65.8	73.3
ICNet	87.8	80.4	92.5	91.6	39.8	96.3	69.1	**55.0**	46.7	87.9	**66.4**	74.0
BiSeNet V2	87.9	80.6	92.9	91.8	40.2	96.1	**70.2**	55.1	**47.6**	88.1	66.0	74.2
PIDNet	87.5	80.2	92.6	91.5	39.7	95.8	69.7	54.8	46.8	87.5	65.6	73.8
RegSeg	87.7	80.5	92.8	91.7	39.9	95.9	69.9	55	46.9	87.8	65.9	74.0
CACNet	**88.6**	**81.2**	**94.2**	**93.0**	**40.3**	**97.7**	69.3	54.2	47.3	**89.2**	65.1	**74.6**

semantic segmentation tasks and is evidently reflected in the enhanced precision of CACNet's output.

Figure 5 displays the schematic post-feature enhancement for the three categories of roads, buildings, and cars in CACNet. Unlike the raw and chaotic intensity distribution in Figure 2, one can clearly discern the edges, shapes, and positions of each category here. Utilizing enhanced category features simplifies outputting precise segmentation results.

2) CAMVID DATASET

The results in Table 6 highlight the performance of various semantic segmentation models on the CamVid dataset. The CACNet model, our proposed approach, outperforms other models with an mIoU of 74.6%. It shows notable improvements in specific categories like 'Sky' (94.2%) and 'Car' (93.0%), which are critical for applications in autonomous navigation and urban planning. Compared to the BiSeNet V2 and RegSeg models, which show mIoUs of 74.2% and 74.0% respectively, CACNet demonstrates a clear advantage, particularly in complex urban environments. The model achieves this high accuracy while maintaining a smaller parameter size, indicating its efficiency and effectiveness for real-time processing. The performance gain with CACNet can be largely attributed to its innovative Class-Aware Context Enhancement Module, which significantly improves the representation of inter-class relationships and context understanding. This enhancement is pivotal in scenarios where the precise delineation of objects from their surroundings is essential.

Figure 6 exhibits some segmentation results of CACNet on the Cityscapes dataset. Arranged from left to right are the original image, ground truth, ENet, ERFNet, and CACNet segmentation results. The red circles in the figure indicate

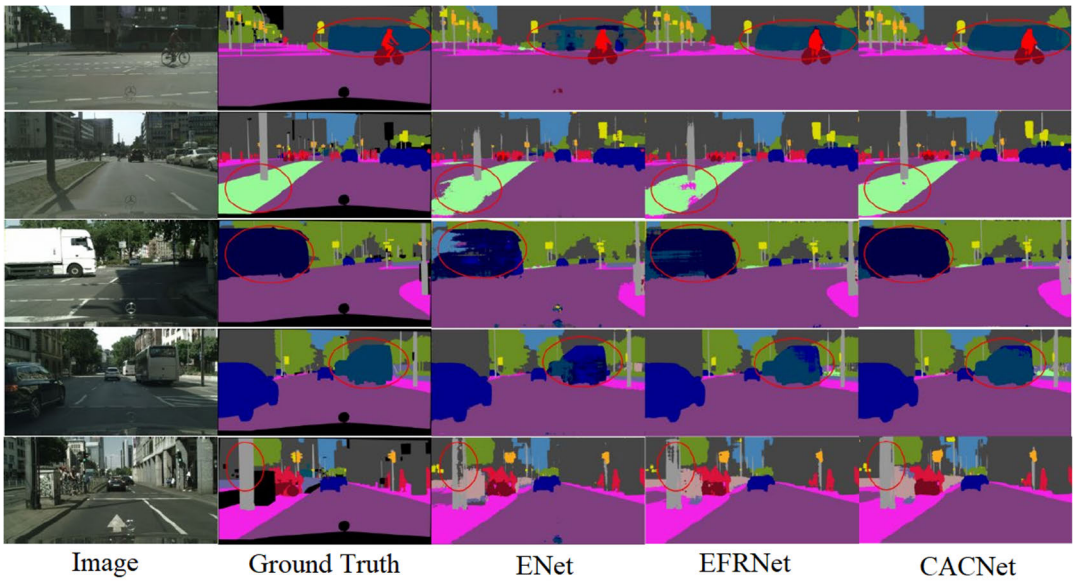

FIGURE 6. Visualization of some results on the CamVid dataset.

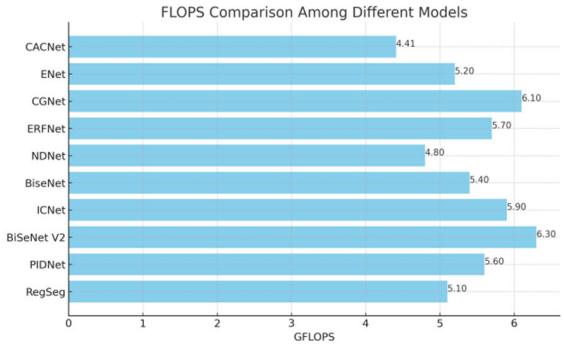

FIGURE 7. Comparison of FLOPS metrics for different models.

areas with notable improvements in the segmentation results. It can be observed that the segmentation outcomes of CACNet align more closely with the ground truth.

We further compared the FLOPs performance of different models, and the comparative results are shown in Figure 7. It can be observed that CACNet has the lowest FLOPs (4.41 GFLOPS). This indicates that CACNet is more computationally efficient than other models. In scenarios with limited computing resources, such as mobile devices or edge computing devices, CACNet is a more suitable choice because it requires fewer computing resources and may also have lower power consumption.

V. CONCLUSION

This paper introduced a lightweight feature enhancement module, the Class-Aware Context Enhancement Module (CACEM). Through visual observation and analysis of the coarse segmentation results yielded by the backbone network, along with drawing from previous research insights, a statistical multi-branch convolutional network is designed as the backbone network. Furthermore, a method to directly enhance the category feature map is proposed. The amalgamation of contextual information with category information benefits in two ways: (1) Guiding the application of contextual information using category information can mitigate misclassifications due to deficient prior knowledge; (1) Strengthening category feature information can directly influence segmentation outcomes, and integrating contextual information can considerably enhance segmentation accuracy. Experimental results indicate that the lightweight semantic segmentation network CACNet, based on the CACEM module, achieved segmentation accuracies of 70.8 and 74.6 on the Cityscapes and CamVid datasets, respectively. These results rank among the best in lightweight segmentation models. Simultaneously, the model size is only 0.52M, the inference speed exceeds 58FPS, and overall performance surpasses all methods in the comparison.

ACKNOWLEDGMENT

The assistance provided was instrumental in the successful completion of this study.

REFERENCES

[1] 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.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[3] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 393–402, Jan. 2020.

[4] 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.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[7] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.
[9] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.
[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[11] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.
[12] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.
[13] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, "Lednet: A lightweight encoder–decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1860–1864.
[14] Z. Yang, H. Yu, Q. Fu, W. Sun, W. Jia, M. Sun, and Z.-H. Mao, "NDNet: Narrow while deep network for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5508–5519, Sep. 2021.
[15] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[16] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: Exploring context and detail for semantic segmentation in real-time," 2018, *arXiv:1805.04554*.
[17] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.
[18] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.
[19] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 9.
[20] S. Vadera and S. Ameen, "Methods for pruning deep neural networks," *IEEE Access*, vol. 10, pp. 63280–63300, 2022.
[21] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, Jul. 2022.
[22] E. Boursier, L. Pillaud-Vivien, and N. Flammarion, "Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 20105–20118.
[23] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109115.
[24] K. Yi, Q. Zhang, L. Cao, S. Wang, G. Long, L. Hu, H. He, Z. Niu, W. Fan, and H. Xiong, "A survey on deep learning based time series analysis with frequency transformation," 2023, *arXiv:2302.02173*.
[25] 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*.
[26] Y. Zhou, S.-M. Moosavi-Dezfooli, N.-M. Cheung, and P. Frossard, "Adaptive quantization for deep neural network," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[27] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" in *Proc. Mach. Learn. Syst.*, vol. 2, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., 2020, pp. 129–146.
[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[30] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoderdecoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.
[31] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.
[32] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9711–9720.
[33] L. Fan, W.-C. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, pp. 71566–71580, 2018.
[34] G. Li, L. Li, and J. Zhang, "BiAttnNet: Bilateral attention for improving real-time semantic segmentation," *IEEE Signal Process. Lett.*, vol. 29, pp. 46–50, 2022.
[35] J. Xu, Z. Xiong, and S. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," 2022, *arXiv:2206.02066*.

XUETIAN XU received the B.S. degree in electronic information science and technology from South China Normal University, China, in 2007, and the M.S. degree in computer theory and software from Sun Yat-sen University, China, in 2009. He is currently an Associate Professor with the Department of Information Management, Guangdong Judicial Police Vocational College. He has authored a total of 25 research articles that have been published in reputable peer-reviewed journals. These articles primarily encompass his key areas of interest, such as digital signal processing, machine learning, and deep learning. His commitment to innovative research is also underscored by his successful filing of three patents of invention, which stand testament to his ingenuity and technical prowess.

SHAORONG HUANG received the M.S. degree in computer theory and software from Sun Yat-sen University, China, in 2003. She is currently a Professor with the Department of Information Management, Guangdong Judicial Police Vocational College. She is a leading talent in Guangdong higher vocational education and has authored over 20 research articles and one patent of invention. Her research interests include computational intelligence, swarm intelligence algorithms, and deep learning algorithms.

HELANG LAI received the B.S. degree in computer science and technology from Shaoguan University, China, in 2010, and the M.S. degree in computer application technology and the Ph.D. degree in theory and technology of service computing from South China Normal University, China, in 2013 and 2021, respectively. He is currently a Lecturer with the Department of Information Management, Guangdong Justice Police Vocational College, China. His current research interests include artificial intelligence, sentiment analysis, and network information security.