IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 January 2023, accepted 26 January 2023, date of publication 2 February 2023, date of current version 9 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241638

RESEARCH ARTICLE

Attention Based Quick Network With Optical Flow Estimation for Semantic Segmentation

JIAWEN CAI, YARONG LIU, AND PAN QIN
School of Control Science and Engineering, Dalian University of Technology, Dalian, Liaoning 116014, China

Corresponding author: Pan Qin (qp112cn@dlut.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 81872247.

ABSTRACT Video semantic segmentation is a challenging vision task since the temporal-spatial characteristics are difficult to model to satisfy the requirements of real-time and accuracy simultaneously. To tackle this problem, this paper proposes a novel optical flow based method. We propose an adaptive threshold key frame scheduling strategy to model the temporal information by estimating the inter-frame similarity. To ensure segmentation accuracy, we construct a convolutional neural network named Quick Network with attention (QNet-attention), a lightweight image semantic segmentation model with a spatial-pyramid-pooling-attention module. The proposed network is further combined with optical flow estimation to realize a semantic segmentation framework. The performance of the proposed method is verified with existing benchmark methods. The experimental results indicated that our method achieves excellent balanced performance on accuracy and speed.

INDEX TERMS Semantic segmentation, deep learning, video processing.

I. INTRODUCTION

Semantic segmentation, as a challenging subject in computer vision, classifies each pixel of images or videos with given semantic labels to achieve the purpose of object detection. Semantic segmentation can be widely applied in obstacle avoidance, tracking, and path planning in the intelligent scenes such as autopilot and unmanned aerial vehicle (UAV) [1]. In the past decades, with the development of deep learning and hardware equipment, deep learning models have been widely used in image and video semantic segmentation and achieved obviously better performance than the classical machine learning methods, leading to great progress in computer vision [3], [4].

Videos are essentially composed of a series of temporally continuous images. The abundant temporal information in the videos can be integrated into the image segmentation model by using special modules to extract effective features to improve the segmentation accuracy. In recent works, the long short-term memory (LSTM) module is applied to learn the temporal features of video and assists the propagation of spatial features [7]. The Netwarp structure uses optical flow to fuse the features of the previous frame with those of the current frame [8]. The spatial-temporal transformer gate recurrent unit (STGRU) module takes the neighbor frames of the current frame as inputs for training the optical flow based semantic segmentation model [9]. Reference [10] proposes the segmentation transformer in the coding stage to process continuous video sequences for global context information.

According to the similarities among the frames, the videos often contain redundant information that should be considered to be reduced. The classical deep feature flow (DFF) structure is combined with the fixed interval key frame selection strategy, where the features of the previous key frames are directly converted by the optical flow method in the feature extraction process of the current frame [11]. The optical flow method is much simpler than the feature extraction on the calculation issue. Because the fixed key frame is difficult to determine the time interval threshold, the adaptive key frame selection strategy is investigated [12]. Based on the DFF structure, a shallow neural network structure is added to the dynamic video segmentation network (DVSNet) to judge whether the current frame is a key frame [13].

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

Although the aforementioned works focus on the real-time performance of semantic segmentation method, most of them take the loss of accuracy as the cost, especially for the small targets. To tackle this problem, this research investigates the real-time video semantic segmentation under limited computing and storage conditions for soundable accuracy. We first propose a module combining spatial pyramid pooling module (SPP) [14] and attention mechanism (AM), which can extend the range of receptive field and effectively realize the classification of small targets. Then, we construct a lightweight image semantic segmentation model based on QNet [15], called QNet-attention, which is further combined with FlowNet2-s [16] to realize a video semantic segmentation framework. Finally, a comparative experiment with other state-of-the-art frameworks are conducted on the Cityscapes [17] dataset. The experimental results verify the excellent performance of the proposed video semantic segmentation framework.

In summary, this paper proposes a lightweight deep learning neural network based on the simplification of network structures, global and local feature fusion, and key frame selection. This research makes the following contributions:

1) Propose a module combining spatial pyramid pooling module (SPP) [14] and attention mechanism (AM), which can extend the range of receptive field and effectively realize the classification of small targets. We named this new module SPP-attention Module (SPP-A).
2) Construct a lightweight image semantic segmentation model based on QNet [15], called QNet-attention. The experimental results show that it realizes an excellent balance between accuracy and speed.
3) An adaptive threshold key frame scheduling strategy combined with optical flow method is proposed, which not only ensures the overall segmentation accuracy, but also improves the model reasoning speed.
4) Propose a video semantic segmentation framework of QNet-attention + FlowNet2-s [16], and carry out comparative experiments with other state-of-the-art frameworks on the Cityscapes [17] dataset. The experimental results verify the excellent performance of the video semantic segmentation framework proposed in this paper.

The rest of this paper is organized as follows. Section II introduces several related works. Section III introduces the proposed video semantic segmentation method. Experiments are depicted in Section IV. Finally, the paper is concluded in Section V.

II. RELATED WORKS

In this section, related works are introduced from two aspects: we first introduce several methods for video segmentation acceleration and then review the attention mechanism.

A. FRAME PROCESS STRATEGY

Videos often contain redundant temporal-spatial information, which can bring huge computational cost. Hence, the frame process strategy has been widely considered to reduce the redundancy.

1) INTER FRAME FEATURE PROPAGATION STRATEGY

Optical flow prediction usually estimates the position of each pixel in an image in the adjacent image from a pair of time-dependent image pairs [19]. The DFF algorithm [13] carries out for the inter frame feature propagation through the optical flow method, which reduces the number of feature extraction links and significantly improves the calculation speed. The low-latency video semantic segmentation uses convolutional neural network (CNN) to propagate the previous deep features to the current frame, and fuses it with the low-level features of the current frame [12]. The temporally distributed network (TDNet) circularly allocates several sub-networks to frames in chronological order, and performs a lightweight forward propagation on each frame [20]. Finally, all features are aggregated by reusing the sub-features extracted in the previous frames. Unlike DFF, we directly propagate the segmentation result of the key frame to the current frame instead of features.

2) KEY FRAME SCHEDULING STRATEGY

The scheduling of key frames is an essential step in feature propagation process. At present, there are mainly three types of methods, including clustering, optical flow, and quality, to extract key frames [28]. The clustering method maps the image information to the high-dimensional space composed of feature vectors and then classifies it by clustering [21]. The optical flow method obtains the motion information according to the optical flow between video frames, such as Lucas-Kanade optical flow method [26], to select the key frame. The quality method scores the image according to different measure standard [29]. These traditional methods often cannot meet the real-time requirement. To improve the real-time efficiency, the fixed interval key frame scheduling strategy has gradually become one of the research focus [30]. This method is simple and easy and greatly improves the efficiency of key frame scheduling. DVSNet measures the similarity of video images between frames through neural network. If the value exceeds a certain threshold, it indicates that the similarity of two frames is high, and the current frame is a non key frame. Otherwise, it is a key frame.

B. ATTENTION MECHANISM

The attention mechanism is used to determine where to focus and assists in making adaptive feature refinement. Recently, several attempts [22], [23], [24], [25] have been made to incorporate attention mechanisms into semantic segmentation tasks. Dual attention network (DANet) [22] append two types of attention modules on top of traditional dilated fully convolutional networks (FCN) [34] to adaptively

TABLE 1. Table of notations.

Notation	Meaning
k	Index of key frame
i	Index of current frame
I_k, I_i	Video frame
S_k, S_i	Segmentation results of video frames
W	Feature propagation function
$F_{k\to i}$	Interframe optical flow field
p, q	Pixels corresponding to two frames
$O_i(p), S_i(p)$	Semantic category label of pixel p
$C(u, v)$	Illustrative function that outputs 1 when $u = v$, otherwise 0
$score_i$	Confidence score of the ith frame and its nearest key frame
t	Confidence threshold for identifing new key frame

integrate local features with their global dependencies. Criss-cross attention network (CCNet) [23] uses a novel criss-cross attention module to capture contextual information from long-range dependencies in a more efficient and effective way. Spatial and channel squeeze & Excitation (scSE) [24] proposes three modules cSE, sSE and scSE, which can enhance meaningful features and suppress useless features. Hierarchical multi-scale attention (HMSA) [25] can improve the problem of category confusion and find the best prediction results from multiple scales.

Motivated by the successes of attention mechanisms, we proposed an attention mechanism module combined with spatial pyramid pooling module [14] for multi-feature fusion. It conducts global-level, channel-level, and spatial-level attention to refine the fused features, which enables the model to select the meaningful ones.

III. PROPOSED METHOD

To satisfy the requirements of real-time and accuracy, we propose a novel video semantic segmentation framework. In this section, we introduce our method in detail.

A. OVERALL FRAMEWORK

This framework is divided into two branches: the optical flow branch and the segmentation branch. Notations used in this section are shown in Table 1.

1) *Step 1:* Current frame I_i and keyframe I_k are subjected to the optical flow calculation network simultaneously to obtain the optical flow field between the two frames. Then input the optical flow field to the decision network (DN). The decision network starts to analyze the similarity between the two input video frames and calculates the confidence of the predicted value between them. The confidence of the predicted value is further compared with the set confidence threshold t. Once the confidence is greater than the threshold, the current frame is then sent to the optical flow branch for further processing; otherwise, it is processed by the segmentation branch. As shown in Fig. 1, the current frame I_i processes through the optical flow branch (red flowchart in Fig. 1), the next frame I_{i+1} processes by dividing branches (blue flowchart). The high confidence of the predicted value implies that the current frame is similar to the key frame and good segmentation results can be obtained by the optical flow branch. Meanwhile, t determines the use frequency of the two branches and also affects the final segmentation speed and accuracy. For more discussion on the decision network, readers are referred to Section III-D.

2) *Step 2*: According to the similarity between the current frame and the key frame, the decision network sends each video frame into two subsequent different branches to obtain the segmentation result of the current frame. The segmentation branch directly sends the current frame to the semantic segmentation network for processing, the same as the general image semantic segmentation process. The optical flow branch takes the optical flow field of the current frame and the key frame in step 1 as the input and converts the previously processed key frame segmentation image to the segmentation result of the current frame through the propagation function W, which does not need to be processed through the segmentation network. Note that the optical flow branch cannot obtain the segmentation result only by relying on the optical flow computing network. The segmentation result and propagation function of the pre-nearest neighbor key frame must be used. Readers are referred to Section III-C for the specific discussion of the propagation process.

B. QNet-ATTENTION NETWORK

To improve the segmentation accuracy of small-scale targets, we combine the channel attention mechanism and spatial attention mechanism with the spatial pyramid pooling module, and design a lightweight semantic segmentation network based on QNet network, called QNet-attention. The structure diagram of the model is shown in Fig. 2, and the network structure is depicted in Table 2.

QNet-attention mainly uses the feature extraction unit based on Channel Split and Channel Shuffle [31]. Block 1 is composed of five basic units; block 2 is composed of nine basic units, where the first units are downsampling. The decoding stage mainly uses the spatial pyramid pooling-attention mechanism fusion module designed in this paper, as shown in Fig. 2. We use 1 × 1, 2 × 2, 4 × 4, 8 × 8 pooled kernel and step size to average the feature maps obtained at the coding stage, and obtain four feature maps of different sizes. To maintain the weight of global features, we use 1 × 1 convolution to reduce the channels of each feature map by half. Then upsample these low-dimensional feature maps by bilinear interpolation to recover their scale. Meanwhile, the four feature maps are added into two groups, the channel attention mechanism module follows the first group. The spatial attention mechanism module follows the second group. Finally, the two groups of feature maps are concatenated with the original input feature map and the feature maps at different levels are superimposed into the final global feature. This module uses receptive fields of different sizes to aggregate the information of different regions of the input

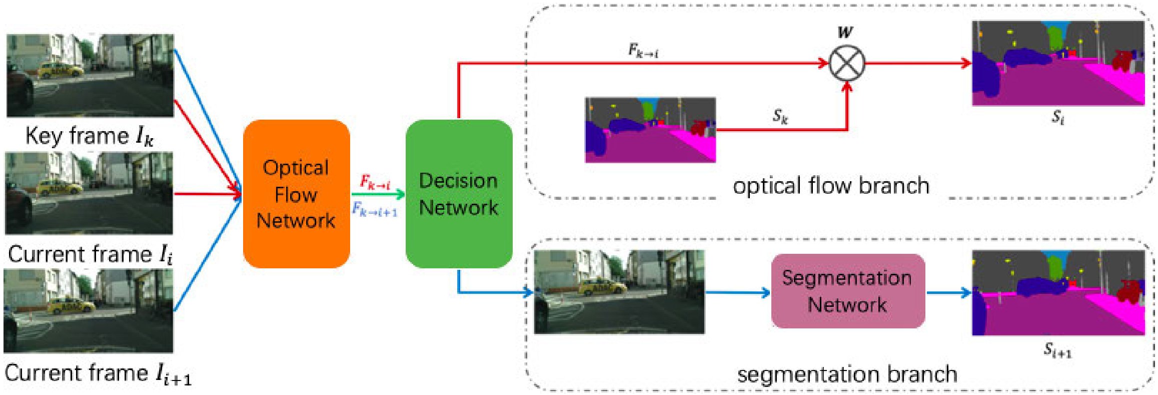

FIGURE 1. The framework of video semantic segmentation based on optical flow method. Different colors represent different similarity confidence.

TABLE 2. QNet-attention architecture (The input and output sizes are set to 1024 × 2048. C is the number of classes).

Stage	Type	Output size
Image	Initial Block	1024×2048×3
Initial		512×1024×16
Encoder	Block1.0-Basic Unit (DownSampling)	256×512×64
	4×Block1.x-Basic Unit	256×512×64
	Block2.0-Basic Unit (DownSampling)	128×256×128
	Block2.1-Basic Unit (dilated=1)	128×256×128
	Block2.2-Basic Unit (dilated=2)	128×256×128
	Block2.3-Basic Unit (dilated=5)	128×256×128
	Block2.4-Basic Unit (asymmetric=3)	128×256×128
	Block2.5-Basic Unit (asymmetric=5)	128×256×128
	Block2.6-Basic Unit (dilated=1)	128×256×128
	Block2.7-Basic Unit (dilated=9)	128×256×128
	Block2.8-Basic Unit (dilated=17)	128×256×128
Decoder	SPP-Attention Module	128×256×256
	Upsampling Unit	1024×2048×c

characteristic map to reduce the information loss between different regions. At the same time, the global information and local information of different scales are fully integrated to obtain the additional dimensional information. This improves the ability of the network to pay attention to small-scale targets and the overall reasoning ability of the decoding end. The operation of channel number reducing and final concatenating also fully ensures a low amount of calculation. The above operations improve the ability to understand the coded feature map, and fully utilize the spatial information through multi-scale feature map and attention mechanism, which effectively improve the recognition of small-scale targets, and consider both computational cost and efficiency.

C. FEATURE PROPAGATION STRATEGY

The optical flow method is an essential research content in video analysis. Variational methods, such as the classical Lucas-Kanade optical flow method, are some of the most widely used methods. These methods mainly solve the problem of small displacement of moving objects in video. In recent years, the method based on deep learning and semantic information has been widely used to address the issue of large removal and the robustness of the optical flow method. FlowNet first applies deep CNN to estimate motion directly and obtains good results, while the optical flow method is also used to assist visual tasks, which can speed up the processing speed of conventional video recognition tasks. In this paper, the optical flow method is used to represent the inter-frame correlation of video and propagate the inter-frame feature. Generally speaking, the similarity between consecutive video frames is significant, and the current frame is similar to the nearest neighbor key frame of the previous sequence, with only local differences. Feature propagation can be carried out between the two frames through optical flow. The introduction of the optical flow method for inter frame feature propagation can significantly reduce the use of image segmentation branches and greatly improve the overall reasoning speed of the model on a single frame image. Compared with image semantic segmentation frame by frame, the overall accuracy can also be slightly reduced. However, from the practical application scenarios, the trade-off between speed and accuracy is more important. To further improve the computational efficiency and reasoning speed, we will not propagate the feature in the feature extraction link, but directly propagate the segmentation graph of the key frame to the current frame through the optical flow method so as to obtain the segmentation result of the current frame.

After the current frame I_i is subjected, the optical flow field $F_{k\to i}$ between the two frames is calculated through the optical flow network together with the previous nearest key frame I_k. The position of the pixel p in the current frame I_i is projected back to the corresponding $p + \delta p$ in the key frame I_k through the optical flow field, where $\delta p = F_{k\to i}(p)$. Since δp is generally non-integer, feature conversion can be realized through bilinear interpolation, as shown

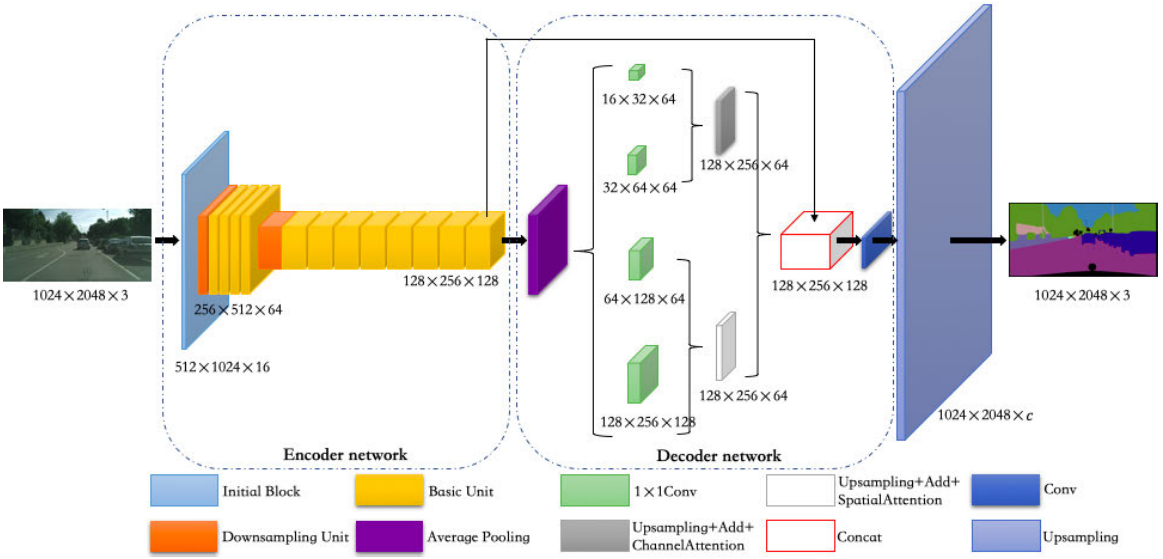

FIGURE 2. An Overview of our proposed QNet-attention.

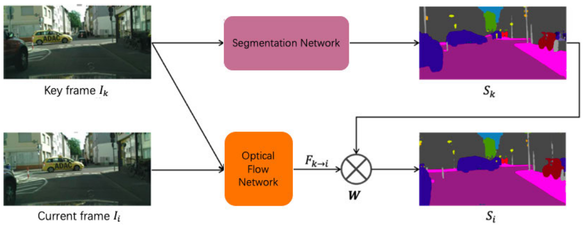

FIGURE 3. The structure of feature propagation strategy based on optical flow method.

in (1a) and (1b):

$$S_i(p) = \sum_q G(q, p + \delta p)S_k(q) \tag{1a}$$

$$G(q, p + \delta p) = g(q_x, p_x + \delta p_x) \cdot g(q_y, p_y + \delta p_y) \tag{1b}$$

where $g(a, b) = \max(0, 1-|a-b|)$. This propagation process can be abbreviated as (2):

$$S_i = W(S_k, F_{k\to i}) \tag{2}$$

In this paper, FlowNet2-s [16] is used as the optical flow computing network, which is significantly improved compared with FlowNet in data training and model structure, and the overall performance of the model is also the best at present. The semantic segmentation network adopts the lightweight image semantic segmentation network QNet-attention designed in this paper. The overall segmentation performance, especially the real-time performance, is significantly improved compared with other mainstream methods. The schematic diagram of feature propagation strategy based on optical flow method is shown in Fig. 3. The feature propagation strategy based on optical flow method proposed in this paper integrates semantic segmentation network and optical flow computing network to build a new model. Both networks can be pre-trained to reduce the computational cost. This strategy by combining the two networks can avoid semantic segmentation of each video frame, effectively reduce the amount of calculation and hereby the overall network will be fully accelerated. Meanwhile, the advanced optical flow computing network ensures the accuracy of feature propagation.

D. KEY FRAME SCHEDULING STRATEGY

Previous research works on key frame scheduling have been conducted based on fixed intervals or simple heuristic methods. Such methods usually cannot deal with complex changes in the video scene, such as sudden camera movement or large changes in the scene structure, which will seriously affect the overall performance of the model. At present, the existing key frame selection methods can be divided into three categories: clustering based, optical flow based and quality based. In this paper, we use the quality-based key frame selection method and the optical flow method to improve the overall quality and efficiency.

Fig. 4 is the overall structure diagram of the key frame scheduling strategy based on the decision network and the training strategy of the decision network. The decision network is a lightweight convolutional neural network composed of only a single convolution layer and three fully connected layers. Its input is the output of optical flow network. In the training phase, the current frame I_i and the nearest key frame I_k are used as inputs. The training purpose is to obtain the confidence score of the predicted value to represent the similarity of the two input images. Then, the optical flow fields $F_{k\to i}$ and Warp function W are calculated by the optical flow network FlowNet2-s, then the semantic segmentation result of the key frame is passed through W to calculate the semantic segmentation output O_i of the current frame.

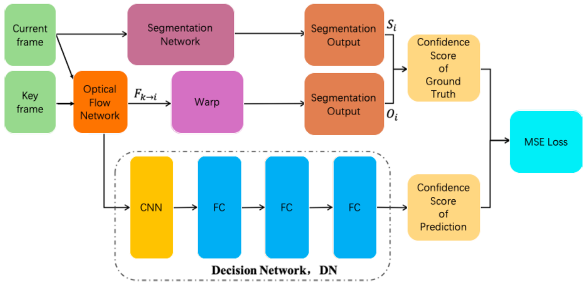

FIGURE 4. The structure and training methodology of Decision Network.

The other branch calculates the semantic segmentation output S_i directly through the image semantic segmentation network QNet-attention proposed. We define the confidence score of ground truth as the expected similarity between O_i and S_i as the following:

$$y_i = \frac{\sum_{p \in P} C(O_i(p), S_i(p))}{P} \tag{3}$$

where P is the total number of pixels in the current frame, p is the index of P. $O_i(p)$ and $S_i(p)$ represent the semantic category label of pixel p calculated by the two branches, respectively. $C(u, v)$ is an illustrative function that outputs 1 when u is equal to v, otherwise 0.

The output of decision network branch is the confidence score of prediction while the output of segmentation branch is the confidence score of ground truth. Based on them, a regression model can be trained with the mean squared error (MSE) loss function. The current frame is identified as a new key frame only when the similarity between the current frame and previous key frame lower than a preset confidence threshold t.

E. ADAPTIVE THRESHOLD STRATEGY

An optimal threshold can lead to a proper total key frame, and it is impossible to set a unified threshold. Therefore, an adaptive threshold strategy is proposed in this paper. Define the confidence score of the ith frame image and its nearest key frame as $score_i$, set the initial threshold to 95 then define the threshold:

$$t = \begin{cases} t + 0.2, & \frac{1}{k}\sum_{j=i-k}^{i} y_j > t \\ t - 2, & \frac{1}{k}\sum_{j=i-k}^{i} y_j \leq t \end{cases} \tag{4}$$

where k can be set manually. Through this method, the threshold can be adaptively and flexibly adjusted according to the video content, and help to improve the efficiency and accuracy of subsequent segmentation tasks.

IV. EXPERIMENTS

A. DATASET

To verify the real-time performance and accuracy of the proposed method, the benchmark datasets Cityscapes [17] are selected for training, testing and performance evaluation with other state-of-the-art semantic segmentation networks. Cityscapes is a large-scale dataset of 50 driving scenes images in different cities, with 19 categories of dense pixel annotation, eight of which have instance level segmentation. There are two sets of evaluation standards, fine and coarse. The former provides 5000 fine labeled images and the latter provides 5000 fine labeled images plus 20000 rough labeled images, with a maximum resolution of 1024 × 2048. In this paper, only fine labeled data are used, of which 2975 are trained, 500 are verified, and the remaining 1525 are tested.

B. EXPERIMENTAL SETTINGS

1) IMPLEMENTATION DETAILS

The video semantic segmentation framework based on optical flow method designed in this paper is mainly composed of segmentation branch, optical flow branch and discrimination network. Considering the training cost and efficiency, we test these three parts separately, and finally integrate them for testing. The segmentation branch uses the lightweight semantic segmentation network QNet-attention proposed in this paper and optical flow branch uses the optical flow computing network FlowNet2-s. As FlowNet2-s is a good pre-trained model, we directly use its pre-trained model. The semantic segmentation network QNet-attention and decision network DN are trained separately. All the training and testing issues are finished based on TensorFlow deep learning framework and carried out on a single NVIDIA Titan RTX GPU.

2) TRAINING DETAILS OF DECISION NETWORK

To improve the training efficiency and consider the continuity and stability of video, we first calculate the confidence score of 38675 frames from the 7th frame to the 19th frame of 2975 video clips in the training set of leftImg8bit_sequence video frame dataset for the decision network. In the calculation process, the 20th frame of each video clip is the key frame with semantic label as the reference frame of other frames in the video clip. Then we build a regression model for training, and take the MSE as loss function as the following:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

where N represents the total number of pictures in the training set, y_i represents the confidence score of the ground truth, and $\hat{y}_i$ represents the confidence score of the prediction. Adadelta optimizer is used to train the decision network. Considering the small scale of the decision network, we trained 100 epochs and the batch capacity is set to 32. The initial learning rate is set as 0.002 and decays at a rate of 0.99 after each epoch.

3) LOSS FUNCTION

Inter class sample imbalance is a common problem in semantic segmentation. The total number of small and medium-sized target pixels is much less than that of background pixels. Because the traditional semantic segmentation training process calculates the loss pixel-by-pixel from isolated pixels, it is difficult for the network to obtain the global

TABLE 3. **The definition of SPP-Attention module's layer *K*.**

k	pooling kernel
2	[1×1, 2×2]
3	[1×1, 2×2, 4×4]
4	[1×1, 2×2, 4×4, 8×8]
5	[1×1, 2×2, 4×4, 8×8, 16×16]
6	[1×1, 2×2, 4×4, 8×8, 16×16, 32×32]

TABLE 4. **The accuracy and real-time indicators results of each network on Cityscapes dataset.**

Method	Class IoU (%)	Category IoU (%)	FLOPs	Parameters	Frame (fps)	Model sizes (M)
QNet	42.3	68.1	20955312937	248778	18.0	1.9
QNet-attention	**49.2**	**70.1**	**19767648256**	**236298**	**18.2**	**1.8**

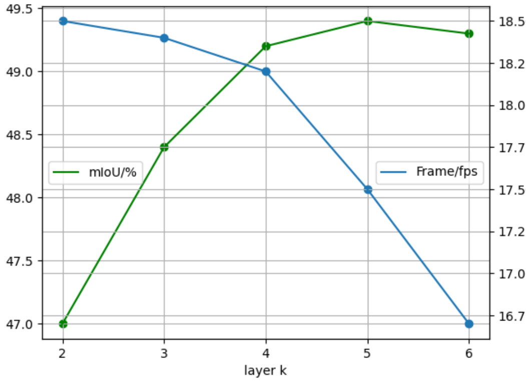

FIGURE 5. **Study of SPP-Attention module. mIoU and fps as a function of SPP-Attention module *k*. Empirically, the SPP-Attention module works best with $k = 4$.**

context information. Therefore, to strengthen the ability of network learning context semantic information and improve the segmentation accuracy of small-size targets, we add a full connection layer branch with sigmoid activation functions on the coding layer and use binary cross-entropy loss to predict the target categories in the image scene. The loss function of the whole network is the weighted sum of the cross-entropy loss of the final decoding layer and the loss of the class prediction branch, in which the weight of the class prediction branch loss is 0.4. Experiments show that this improves the segmentation accuracy of small-size targets.

4) EVALUATION METRICS

This paper mainly considers the evaluation indexes of the accuracy and speed of semantic segmentation. The accuracy mainly includes pixel accuracy (PA), intersection over union (IoU), mean IoU (mIoU), floating point operations (FLOPs), etc. In terms of speed, frames per second (FPS), a total of parameters, and model size are used as the main evaluation index.

C. MODEL ANALYSIS

1) EFFECTIVENESS OF SPP-ATTENTION

To demonstrate the effectiveness of the proposed SPP-Attention module, we test different layers of SPP-Attention module $k = 2, 3, 4, 5, 6$ whose details are shown in Table 3. As shown in Fig. 5, we find $k = 4$ yields the best performance. Because when the SPP-Attention module's layer k increases to 4, the increase of mIoU is close to saturation, and the decrease of velocity is more obvious. Then we compare the accuracy of QNet and QNet-attention on the Cityscapes dataset, and the accuracy evaluation indexes are Class IoU and Category IoU. To verify the real-time performance of SPP-Attention module and consider the computing and storage requirements on mobile devices, we use a lower performance processor (GTX Titan GPU) as the test machine. Taking FLOPs, parameters, model size and FPS as the evaluation indexes. As shown in Table 4, it can be found that without using any pre-trained model, the Class IoU and Category IoU of QNet-Attention are both higher than QNet. At the same time, QNet-attention has lower computational power requirements and parameters than QNet. The overall model is small, which is conducive to storage. It shows that the SPP-Attention module proposed in this paper meets the real-time requirements and is suitable for practical application scenarios.

2) EFFECTIVENESS OF LOSS

To further improve the performance of the model, we introduce a loss function, which is different from the pixel-by-pixel cross-entropy loss function. We add a full connection layer branch with Sigmoid activation function, and use binary cross-entropy loss to predict the target categories in the image scene. The loss function of the whole network is the weighted sum of the cross-entropy loss per pixel and the branch loss. The weight of cross-entropy loss per pixel is always 1, and the weight of branch loss is a positive constant α less than 1. To verify the effectiveness of this loss function, we test the segmentation accuracy of the model under different α based on the Cityscapes dataset. The results show that when $\alpha = 0.4$, the performance is the best, and the segmentation accuracy of small-size targets (such as traffic sign, car, person, etc.) is improved to a certain extent. The test results are shown in Fig. 6 and Table 5, in which the bold part is the optimal value of the same group.

3) EVALUATION OF THRESHOLD ALGORITHM

To evaluate the effectiveness of the adaptive threshold algorithm proposed in this paper, we compare it with the results when the threshold t is 95. The experimental results in Table 7 show that although the adaptive threshold algorithm proposed in this paper still has some redundancy, it still follows the principle of ''better more than less'' in key frame extraction. In addition, most of the key frames extracted by the adaptive threshold algorithm proposed in this paper can well represent the video content. Compared with the methods with constant

TABLE 5. Test results of different loss function weight α IoU on Cityscapes dataset.

weight α	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sigh	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mIoU
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
0	90.3	64.5	87.5	25.6	19.5	31.0	17.5	56.3	86.7	57.7	88.5	59.7	8.7	83.6	67.3	0.0	0.0	0.0	48.6	47.0
0.1	92.2	64.7	88.7	25.7	19.9	31.8	17.9	56.4	87.3	58.3	89.3	60.8	9.2	84.9	69.8	0.0	0.0	0.0	51.3	47.8
0.2	93.9	65.8	90.0	26.7	20.2	32.7	18.1	57.6	88.4	59.7	91.0	61.3	10.0	85.8	70.0	0.0	0.0	0.0	52.2	48.6
0.3	94.6	65.9	90.1	26.8	20.3	32.8	18.2	57.5	**89.3**	59.8	**91.2**	61.4	10.1	85.9	70.1	0.0	0.0	0.0	58.9	49.1
0.4	94.7	**66.1**	**90.1**	**27.2**	**20.3**	**32.8**	**18.4**	**57.6**	86.6	**60.0**	90.5	**61.5**	**10.2**	**86.2**	**70.3**	**0.0**	**0.0**	**0.0**	**62.3**	**49.2**
0.5	**94.8**	66.0	89.9	27.1	20.2	32.7	18.4	57.6	88.0	59.9	90.5	61.4	9.8	86.1	70.3	0.0	0.0	0.0	60.2	49.1
0.6	94.7	66.1	89.8	26.9	20.1	32.6	18.3	57.6	87.9	59.8	90.6	61.3	9.9	86.0	69.9	0.0	0.0	0.0	59.5	49.0
1	94.6	65.8	89.9	26.7	20.2	32.8	18.4	57.6	87.2	59.9	90.6	61.3	10.0	85.9	69.8	0.0	0.0	0.0	58.4	48.9

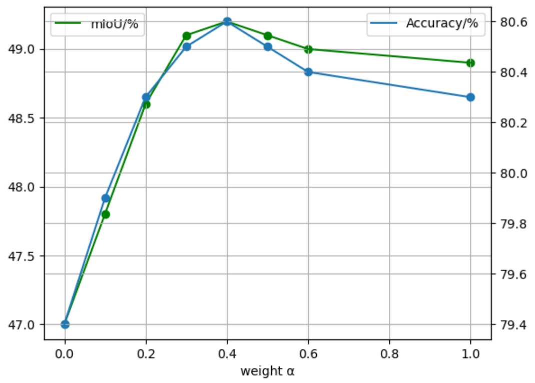

FIGURE 6. Study of loss function. mIoU and accuracy as a function of Loss function weight α. Empirically, the loss function works best with $\alpha = 0.4$.

TABLE 6. The accuracy and real-time indicators of each network on Cityscapes dataset.

Method	Class IoU	Category IoU	FLOPs	Parameters	Frame	Model sizes
	(%)	(%)			(fps)	(M)
PSPNet	**67.8**	**79.3**	5452610569665	86862246	2.3	662.7
ICNet	50.11	72.2	58484301607	6685715	8.8	51.0
ENet	40.1	64.3	126178826835	360618	14.8	2.8
QNet-attention	49.2	70.1	**19767648256**	**236298**	**18.2**	**1.8**

thresholds, our method shows fewer wrong key frames and less redundancy.

D. COMPARISON

1) COMPARISON WITH OTHER NETWORKS

The comparison is conducted with PSPNet [38], ICNet [43], and ENet [40]. Both PSPNet and ICNet use pre-trained models for transfer learning. Table 6 shows the Class IoU, Category IoU, FPS, Model size and other indicators of each network on the test picture. (The bold part in the table is the optimal value of the same group.) Compared with other networks, QNet-attention has lower computational power requirements and parameters without losing too much segmentation accuracy. At the same time, the whole model is small. The image processing speed achieved on the low-performance processor reaches 18.2fps, which is significantly higher than other networks. This result indicates that our lightweight network model meets the real-time requirements. The example results are shown in Fig. 7.

2) COMPARISON WITH OTHER FRAMEWORKS

We compare the accuracy and speed of video semantic segmentation frameworks PSPNet + FlowNet2-s, ICNet + FlowNet2-s, and QNet-attention + FlowNet2-s corresponding to each network. Class IoU, Category IoU, FLOPs, FPS, Model size and other indicators are adopted. The experimental results are shown in Table 8. Compared with other frameworks, the QNet-attention + FlowNet2-s video semantic segmentation framework proposed in this paper has no advantages in accuracy indicators such as IoU. But its performance is more advanced in the indicators of FLOPs, Parameters and Model size. The total number of video frames it processed per second reaches 23.5fps, which is significantly higher than other frameworks. This shows that our proposed framework has more advantages under the condition of limited computing power and storage conditions, and is more suitable for mobile devices in actual scenarios.

Fig. 8 compares the segmentation results of various video semantic segmentation frameworks. It can be found that the video semantic segmentation framework based on optical flow method is effective. Each object has clear segmentation and clear edge, which can accurately reflect the semantic information of the scene. Among them, ICNet + FlowNet2-s has better overall effect, fewer error points and clear object contour. The accuracy of QNet-attention + FlowNet2-s proposed in this paper is guaranteed to a certain extent. At the same time, small objects such as electric poles can be segmented, and each segmented object can correspond to the actual picture. Fig. 9 compares the segmentation results of the image semantic segmentation network QNet-attention and the video semantic segmentation framework QNet-attention + FlowNet2-s proposed in this paper. It can be found that the pure image semantic segmentation has better effect than the video semantic segmentation with optical flow method, the segmented object edge is of less noise. To speed up the segmentation process, using the same segmentation method under the same conditions, video semantic segmentation will lose a certain accuracy.

E. SCENARIO TEST

We also carried out experiments in real scenes. We used DJI Matrice 210 RTK V2 UAV equipped with ZENMUSE X5 camera to fly at low altitude at Beichen Road, the main campus of Dalian University of technology to obtain video data

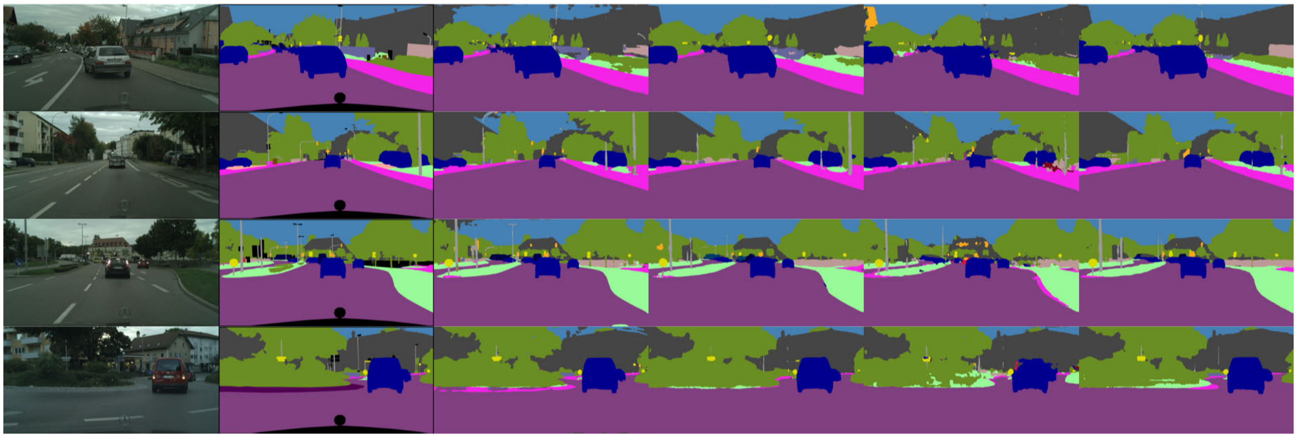

FIGURE 7. Example results of the different networks on Cityscapes dataset (From left to right, it represents input, label, PSPNet, ICNet, ENet, and our QNet-attention).

TABLE 7. Comparison of threshold algorithms on Cityscapes dataset.

Threshold Algorithm	Scene	Total Frames	Real Key Frames	Key Frames Extracted by the Algorithm
$t = 95$	stuttgart	5880	196	2385
Adaptive Threshold				**1401**
$t = 95$	lindau	1770	59	846
Adaptive Threshold				**443**
$t = 95$	munster	5220	174	1996
Adaptive Threshold				**1283**

TABLE 8. Each video semantic segmentation framework accuracy and real-time indicators test results on Cityscapes dataset.

Method	Class IoU(%)	Category IoU(%)	FLOPs	Parameters	Frame(fps)	Model sizes(M)
PSPNet + FlowNet2-s	**50.1**	72.2	5508974748531	91400731	15.2	669.9
ICNet + FlowNet2-s	46.2	**81.3**	114848480473	11224200	16.7	63.7
ENet + FlowNet2-s	33.7	62.8	218374651606	4899103	20.4	39.7
QNet + FlowNet2-s	36.5	67.4	77319491803	4787263	23.4	39.2
QNet-attention + FlowNet2-s (Ours)	43.8	68.8	**64603747328**	**4774783**	**23.5**	**39.1**

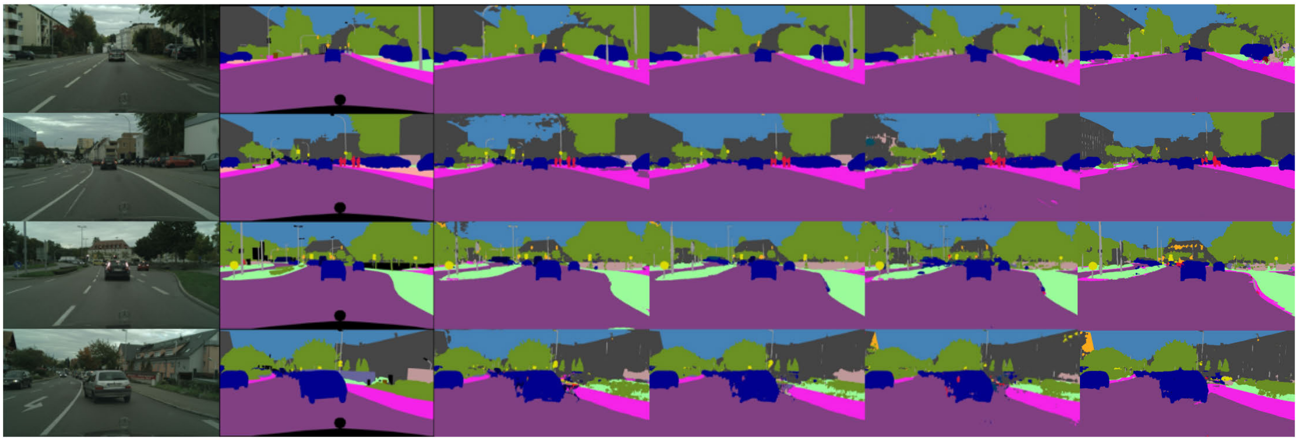

FIGURE 8. Example results of the different video semantic segmentation frameworks on Cityscapes dataset (From left to right, it represents input, label, PSPNet + FlowNet2-s, ICNet + FlowNet2-s, QNet + FlowNet2-s, and our QNet-attention + FlowNet2-s).

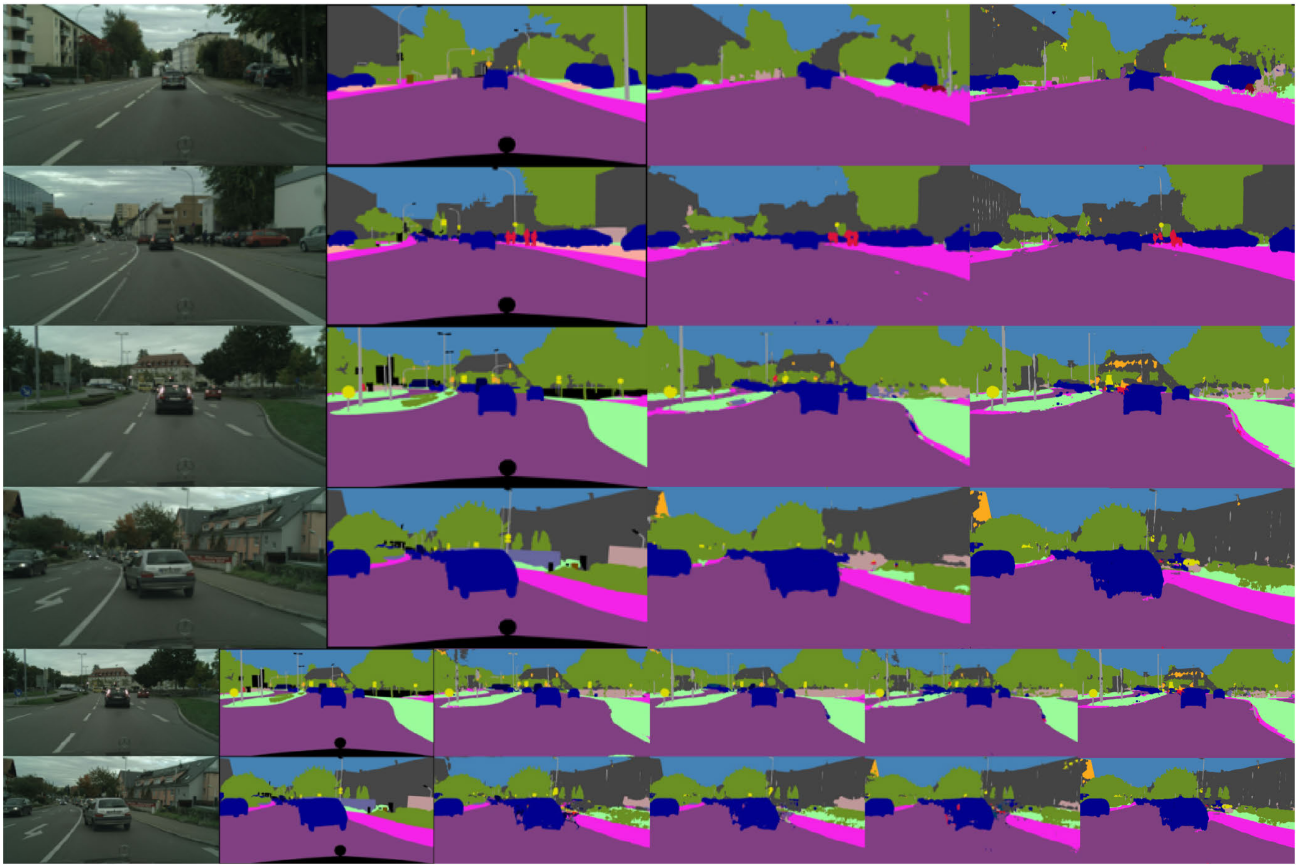

FIGURE 9. Example results of image semantic segmentation and video semantic segmentation results on Cityscapes dataset (From left to right, it represents input, label and QNet-attention, QNet-attention + FlowNet2-s).

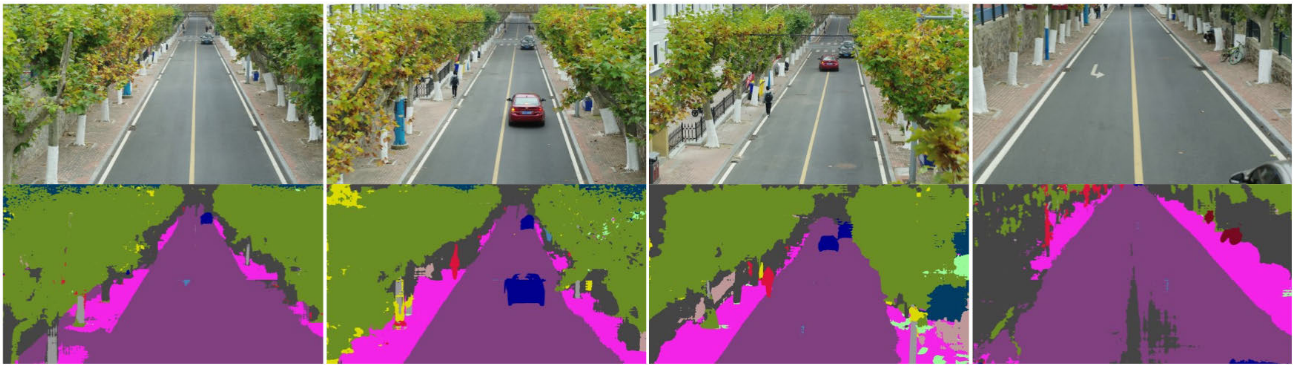

FIGURE 10. The actual scene segmentation from the drone's low-altitude perspective (Based on QNet-attention+FlowNet2-s).

similar to Cityscapes street view dataset. Then we used QNet-attention + FlowNet2-s proposed in this paper to directly perform semantic segmentation. The overall effect is shown in Fig. 10. It can be found that the overall segmentation effect is good. The segmentation of main objects such as roads, pedestrians, cars and trees is clear, but some images have more noise and some false segmentation, especially when the flight speed of UAV is unstable or the scene is too complex. To ensure the segmentation effect of the practical scene, the generalization ability of the segmentation model should be further improved.

For implementation, the proposed framework can be applied to the outdoor scene of low altitude and low speed flight of UAV or the outdoor street scene of automobile. In terms of hardware, it needs to be equipped with a camera and have a certain computing power. And it is better to use the model in a bright day environment. In a simple environment with little scene change, the model reasoning speed is faster and more efficient than in a complex environment with fast scene change.

V. CONCLUSION

In this paper, we present an SPP-Attention module and analyze its effectiveness. Accompany this, we also propose a lightweight image semantic segmentation model QNet-attention. Based on QNet-attention, a video semantic segmentation framework is proposed. The proposed framework consists of QNet-attention segmentation branch and FlowNet2-s optical flow branch, in which the segmentation branch performs semantic segmentation of key frames, and the optical flow branch performs inter frame feature propagation and key frame scheduling. We have proposed an adaptive threshold key frame scheduling strategy based on decision network for the key frame scheduling problem. In the experimental part, we conducted comparative experiments with other frameworks on the Cityscapes dataset under the same conditions, and the experimental results verified the excellent performance of the proposed framework. In addition, we have tested the proposed video semantic segmentation framework in the UAV cruise scene, and the segmentation effect is good, which can meet a certain degree of accuracy and real-time.

ACKNOWLEDGMENT

(Jiawen Cai and Yarong Liu contributed equally to this work.)

REFERENCES

[1] T. Ohgushi, K. Horiguchi, and M. Yamanaka, *Road Obstacle Detection Method Based on an Autoencoder With Semantic Segmentation*. Cham, Switzerland: Springer, 2020.

[2] X. Tian, L. Wang, and Q. Ding, "Overview of image semantic segmentation methods based on deep learning," *J. Softw.*, vol. 30, no. 2, pp. 440–468, 2019.

[3] D. He and C. Xie, "Semantic image segmentation algorithm in a deep learning computer network," *Multimedia Syst.*, vol. 28, no. 6, pp. 2065–2077, Dec. 2022.

[4] H. Hao and J. Wu, "Overview of image semantic segmentation technology based on deep learning," *Comput. Eng. Appl.*, vol. 55, no. 19, pp. 12–21, 2019.

[5] L. Han and C. Meng, "Overview of video semantic segmentation based on deep learning," *Comput. Syst. Appl.*, vol. 28, no. 12, pp. 1–8, 2019.

[6] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, Sep. 2018.

[7] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[8] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video CNNs through representation warping," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4453–4462.

[9] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6819–6828.

[10] 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," 2021, *arXiv:2012.15840*.

[11] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2349–2358.

[12] Y. Li, J. Shi, and D. Lin, "Low-latency video semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5997–6005.

[13] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, "Dynamic video segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6556–6565.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[15] A. Xia, D. Li, J. Cai, H. Gu, and P. Qin, "QNet: A quick deep neural network for real-time semantic segmentation," in *Proc. IEEE 5th Int. Conf. Signal Image Process. (ICSIP)*, Oct. 2020, pp. 102–107.

[16] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

[19] X. Li, "Overview of visual optical flow vector field estimation algorithms," *J. Beijing Univ. Technol.*, vol. 39, no. 11, pp. 1638–1643, 2013.

[20] P. Hu, F. C. Heilbron, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," 2020, *arXiv:2004.01800*.

[21] Y. Zhu and D. Zhou, "A key frame extraction method based on video clustering," *Comput. Eng.*, vol. 30, no. 4, p. 3, 2004.

[22] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[23] 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.

[24] A. Roy, N. Nav, and C. Wachinger, *Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks*. Cham, Switzerland: Springer, 2018.

[25] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," 2020, *arXiv:2005.10821*.

[26] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell. (IJCAI)*, 1981, pp. 674–679.

[27] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[28] X. Qi, C. Liu, and S. Schuckers, "Boosting face in video recognition via CNN based key frame extraction," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 132–139.

[29] J. Cao, "Research on video key frame extraction," Chongqing Univ., Tech. Rep., 2008.

[30] S. Ghatak, "Key-frame extraction using threshold technique," *Int. J. Eng. Appl. Sci. Technol.*, vol. 1, no. 8, pp. 51–56, 2016.

[31] 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.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," 2016, *arXiv:1605.08695*.

[33] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[35] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[36] 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.

[37] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[38] 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. 2881–2890.

[39] 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.

[40] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[41] 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.

[42] 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*.

[43] H. Zhao, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Euro. Conf. Comput. Vis.*, 2018, pp. 405–420.

[44] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Euro. Conf. Comput. Vis.*, 2018, pp. 552–568.

[45] 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.

[46] H. Si, Z. Zhang, F. Lv, G. Yu, and F. Lu, "Real-time semantic segmentation via multiply spatial fusion network," 2019, *arXiv:1911.07217*.

[47] T. Akiba, S. Suzuki, and K. Fukuda, "Extremely large minibatch SGD: Training ResNet-50 on ImageNet in 15 minutes," 2017, *arXiv:1711.04325*.

[48] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[49] 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.

[50] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[51] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, 2009.

[52] M. Wang, B. Liu, and H. Foroosh, "Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial 'bottleneck' structure," 2016, *arXiv:1608.04337*.

[53] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Euro. Conf. Comput. Vis.*, 2018, pp. 116–131.

[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[55] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, 2005.

JIAWEN CAI was born in Quanzhou, Fujian, China, in 1997. He received the B.S. degree in control engineering from the Dalian University of Technology, Liaoning, China, where he is currently pursuing the M.S. degree under the guidance of Prof. Pan Qin. His research interests include semantic segmentation and point cloud data analysis.

YARONG LIU received the M.S. degree from the Dalian University of Technology, Dalian, China, in 2021, where he is currently pursuing the Ph.D. degree with the Faculty of Electronic Information and Electrical Engineering.

His research interests include deep learning, bioinformatics, and statistical modeling.

PAN QIN received the B.S. and M.S. degrees in the study and research of aircraft power supply system from Northwest Polytechnic University, Shaanxi, China, and the Ph.D. degree in the research of identification and predictive control algorithms for multi rate systems from Kyushu National University, Japan. He worked as an Academic Researcher for nearly six years with the School of Mathematics and Science and the Institute of Mathematics for Industry (IMI). He is currently an Associate Professor with the Dalian University of Technology. His current research interests include statistics and data mining.