CrossNet: Cross-scene Background Subtraction Network via 3D Optical Flow

Dong Liang, Dong Zhang, Qiong Wang*, Zongqi Wei, Liyan Zhang

***Abstract*—This paper investigates an intriguing yet unsolved problem of cross-scene background subtraction for training only one deep model to process large-scale video streaming. We propose an end-to-end cross-scene background subtraction network via 3D optical flow, dubbed CrossNet. First, we design a new motion descriptor, hierarchical 3D optical flows (3D-HOP), to observe fine-grained motion. Then, we build a cross-modal dynamic feature filter (CmDFF) to enable the motion and appearance feature interaction. CrossNet exhibits better generalization since the proposed modules are encouraged to learn more discriminative semantic information between the foreground and the background. Furthermore, we design a loss function to balance the size diversity of foreground instances since small objects are usually missed due to training bias. Our whole background subtraction model is called Hierarchical Optical Flow Attention Model (HOFAM). Unlike most of the existing stochastic-process-based and CNN-based background subtraction models, HOFAM will avoid inaccurate online model updating, not heavily rely on scene-specific information, and well represent ambient motion in the open world. Experimental results on several well-known benchmarks demonstrate that it outperforms state-of-the-art by a large margin. The proposed framework can be flexibly integrated into arbitrary streaming media systems in a plug-and-play form. Codes are available at https://github.com/dongzhang89/HOFAM.**

***Index Terms*—CrossNet, Background subtraction, Cross-scene, 3D Optic flow, Streaming media**

I. Introduction

VIDEO background subtraction aims to recognize and segment all the pixel-level elements of moving foreground from a dynamic background, which has served as a fundamental role in a wide range of multimedia community, *e.g.*, video streaming summarization [1], video source/channel encoding and compressions such as MPEG series and H.264 [2], and large-scale streaming media synthesis [3].

In the past several years, this task has made significant progress. Yet, cross-scene background subtraction is still very challenging because most of the existing deep learning models [4], [5] and traditional background subtraction models [6]–[8] are scene-dependent trained, restricting the consistent performance in open-world cross scenes. In modern multimedia systems, edge computing devices capture video online and are required to real-time synopsize valuable information. However, edge computing devices usually rely on resource-constrained embedded platforms; thus, deploying an online background subtraction module without scene-specific supervised training is vital in the industry.

Following the most common way, the existing successful methods usually adopt strategies that are pre-training the background subtraction model in a vast dataset and then fine-tuning it in new scenes. Transfer learning [9]–[11], teacher-student [12], [13] and meta-learning strategy [14] could further adapt the model smoothly to new domains. However, the existing works that be able to transfer learning are originally designed for image classification tasks; The domain adaptation strategies for background subtraction require time-consuming pixel-wise labels in both source and target domains. An image with the size of 2048×1024 takes 1.5 hours to get a fine label and 7 minutes to generate a coarse one [15]. Therefore, the domain adaptation strategies learn from noisy data that may affect their inference performance in practice. In addition, some methods [16]–[18] also assume that the images in two domains have the same task prediction. Since the foreground semantic has large diversity, mapping the source and target domains is challenging. In our previous research work, we proposed a series of background subtraction methods to deal with dynamic scenes [7], [19]–[21], and we also proposed an interaction scheme [22], [23] between the background subtraction method and deep learning models to enhance the flexibility of both sides. However, none of the proposed approaches are free from extra labeling and supervised learning in a new scene to maintain reliable foreground segmentation performance.

This work focuses on the challenging task of video background subtraction in unseen scenes without any additional labeling and training. The proposed scheme aims to achieve scene adaptation for background subtraction via fine-grained motion feature representations and interactions. Since the optical flow is insensitive to gradual motion and is not robust to ambient lighting changes in the open world, we first design 3D hierarchical optical flows (3D-HOP) to convert instantaneous flows to fine-grained motion cues. We construct a cross-modal dynamic feature filter (CmDFF) to realize feature interactions between motion and appearance. Unlike using optical flow as a motion trigger, the proposed scheme tends to learn semantic-level discriminative information between the motion patterns of the foreground instances and the background. Therefore, better adaptability and robustness in cross-scene tasks can be obtained. In addition, since small foreground objects are usually missed in cross-scene background subtraction tasks due to

D. Liang, Z. Wei, L. Zhang are with MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Collaborative Innovation Center of Novel Software Technology and Industrialization, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

D. Zhang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

Q. Wang is with Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

∗ Corresponding author: Q. Wang (wangqiong@siat.ac.cn).

the dataset bias, we design a Class-in Scale Focal (CS-Focal) Loss to balance the size diversity of foreground instances. The proposed model can segment the video foreground by deploying them in an unseen scene without fine-tuning.

The main contributions of this paper are five folds:

- Dense optical flow is an instantaneous motion context that is less robust and inadequate to describe motions at the pixel level. We, therefore, design the 3D hierarchical optical flows (3D-HOP) to combine the long-term and short-term flow field estimation, converting instantaneous flows to fine-grained motion cues.
- We propose an end-to-end network, CrossNet, to realize cross-scene background subtraction in large-scale video streaming without any extra training. We embed a cross-modal dynamic feature filter (CmDFF) to realize the interaction between motion and appearance features. We also improve the pixel-level Focal Loss to a Class-in Scale Focal (CS-Focal) Loss fashion to balance the size diversity of the foreground instances.
- Our whole background subtraction framework, including the proposed 3D-HOP and CrossNet, called Hierarchical Optical Flow Attention Model (HOFAM), is compared with state-of-the-art methods via comprehensive experiments cross scenes. All results consistently endorse the superiority of the proposed approach.

This paper was partially presented in [24] while further theoretical investigation and extensions with thoroughly analyzed and discussed. Extensive evaluation has been undertaken to compare it against the state of the arts. The remainder of this paper is organized as follows. We discuss the related work in Section II. We describe the proposed method in detail in Section III. The experimental results are presented and discussed in Section IV, and the conclusions, limitations, and future work are presented in Section V.

II. Related Work

A. Unsupervised Background Subtraction

Early studies focused on statistical distributions to build the background model [25]–[27]. Spatio-temporal local descriptors [7], [20], [28]–[30] reveal the Spatio-temporal dependence that the background models can hold. The above statistical modeling methods usually have a low computational cost, which benefits resource-constrained multimedia systems. To eliminate the impact brought by illumination changes and dynamic background, imprecise progressive background updating solutions are commonly used [26]: 1) selective updating, in which a new sample is added to the model only if it is classified as a background sample, and 2) blind updating, in which every new sample is added to the model. Using selective updating, one must decide whether each pixel value is part of the background. Using the segmentation results as the updating criterion can be seen as a simple way to achieve this task, while invalid segmentation decisions may result in incorrect segmentation afterward. The blind updating mechanism is not subject to this deadlock scenario since it does not involve any updating decision; Blind updating mechanism allows intensity values not belonging to the background to be added to the model, which leads to more error accumulation as the foreground pixels may erroneously become part of the model. A high update rate leads to noisy segmentation due to the sensitivity to minor or temporary changes whereas a low update rate yields an outdated background model and results in false foreground segmentation.

B. Background Subtraction based on CNN

Brahamand [31] proposes the first approach using CNNs to undertake background subtraction. A scene-specific network is trained by corresponding image patches of frames. Branches with different sizes in Cascade CNN [5] are connected to detect multi-scale foreground objects, while temporal information in videos has not been considered. MFC3-D [32] leverages multi-scale 3-D convolution to detect the foreground. MSNet [33] uses Generative Adversarial Networks to generate the background. A probabilistic model [34] divides each video frame into patches, fed to a stacked denoising auto-encoder to extract significant features. All the methods are scene-specific. To our knowledge, DeepBS [35] is the first method to utilize a trained CNN for the background subtraction task across video scenes. For the training data, it randomly selects 5% samples with the corresponding ground truths of each subset from the CDNet2014 dataset. However, most CNN-based background subtraction models have not considered temporal information.

C. Semantic Segmentation

Semantic segmentation methods have enabled remarkable progress recently. Based on a pre-trained ResNet model, PSPNet [36] uses atrous convolution to perform feature extraction, in which the pyramid pooling module collects and integrates contextual information between different scales. DeepLabV3+ [37] applies depth-wise separable convolution to both the Atrous Spatial Pyramid Pooling and decoder modules. It utilizes the Xception model to integrate multi-scale information for the segmentation task. A boundary-aware feature propagation [38] shares local features within their regions and learns the boundary as an additional semantic class to create the network to be aware of object boundary layouts. CCL [39] is an aggregation scheme called gated sum, which aims to select different scale feature maps. It uses the context-contrasted local module in the network to generate multi-scale and multi-level context-aware local features. Most semantic segmentation approaches bring semantic annotations to independent frames, ignoring motion cues and temporal relevance. These are crucial to discriminating the foreground and dynamic background elements in background subtraction.

D. Pixel-level Motion Estimation

Sequences of ordered frames allow motion estimation as either instantaneous image velocities or discrete image displacements [40]–[42]. Most pixel-level motion estimation methods are based on optical flows. Lucas and Kanade [40] use spatial intensity gradients and Newton–Raphson iterations to formulate computable scene motion flows. Gunnar Farnebackused [43] uses exact polynomial transforms to calculate large displacement dense optical flow. FlowNet [44]

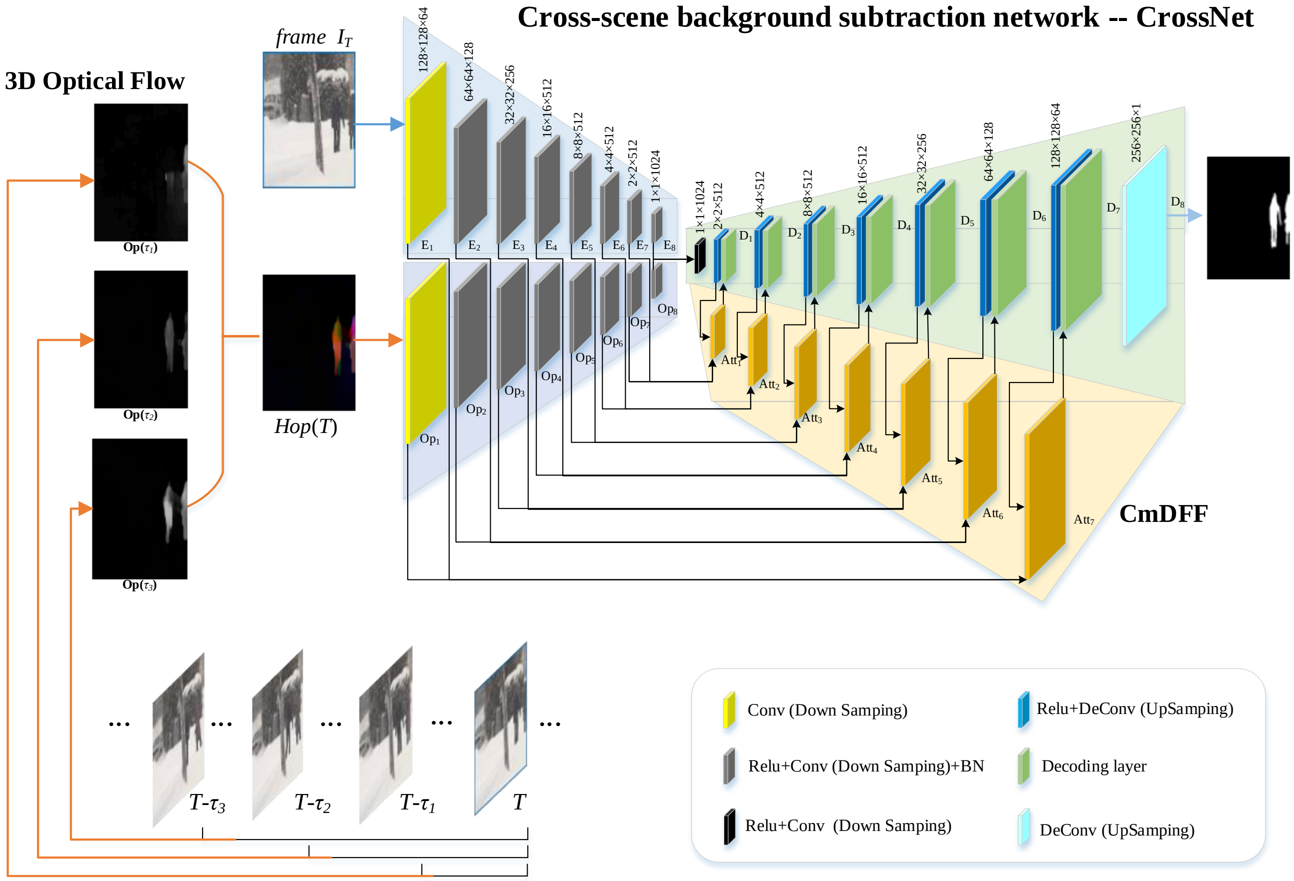

Fig. 1: The proposed background subtraction model includes an end-to-end cross-scene background subtraction network CrossNet and 3D optical flow. The input is the original frame I_T, its 3D optical Flow $Hop_(T)$, and the output is the binary foreground segmentation mask. CrossNet integrates multiple CmDFF modules in the up-sampling process to fuse the features of the encoder and the decoder, where E_i and Op_i are featured in the encoder for extracting appearance and optical flow features, respectively. D_i are the decoder features, and Att_i refer to the CmDFF modules in the corresponding layers. CrossNet uses a full convolution structure as the backbone network.

and FlowNet2.0 [41] update the training using warping structures to process optical flows accurately. ScopeFlow [45] handles optical flow prediction in dynamic environments. GLU-Net [46] uses global and local correlation layers to deal with dense optical flow problems. To deal with object occlusion, LiteFlowNet3 [47] introduces adaptive adjustment and local flow consistency. MANet [48] fuses multi-frame features and directly learns the motion cues of an extended period. Selflow [49] uses self-supervision to calculate optical flows. Flow-Guided Feature Aggregation [50] finds that optical flows can improve the performance of object segmentation.

III. The Approach

A. Motivations

In the above work, the existing background subtraction methods are unsupervised pixel-level binary classification. They build pixel-wise statistic model in a stochastic process [6], [8], [25], [26] or model the context among pixels [22], [23], [27]. The advanced semantic segmentation methods have enabled remarkable progress recently [36]–[38]. However, they usually bring high-cost pixel-wise annotations and scene-specific training while ignoring the motion cues and temporal relevance. Fundamental issues needed to be solved: 1) Traditional unsupervised background subtraction is often trapped by inaccurate online model updating; 2) The supervised methods, such as the off-the-peg deep models, heavily rely on scene-specific information, thereby limiting their cross-scene performance; 3) Optical flow inadequately represents ambient motion in the open world.

B. CrossNet

Our basic viewpoint is that background subtraction is related to fine-grained motion semantics. For example, motion regions can be foreground patterns, but areas of repetitive motions (*e.g.*, waving tree branches) belong to the background pattern. On the other hand, dense optical flow is an instantaneous motion context that is less robust and inadequate to describe motions at the pixel level. In contrast, the proposed hierarchical 3D optical flow HOP (c.f. Section III-C) can leverage both long-term and short-term motion to organize multi-scale optical flow to generate a fine-grained motion state space.

The proposed framework then embeds the 3D-HOP motion feature with the appearance feature via the proposed CmDFF (c.f. Section III-D). CmDFF is the critical module to modify a traditional CNN-based encoder and decoder deep network to build our CrossNet. CmDFF further highlights the scene-independent motion semantics to identify the suspected foreground while filtering out the scene-dependent background movements. The proposed scheme discriminates the scene-independent motion patterns in an unseen scene to avoid the domain shift. Therefore, a better cross-domain generalization can be obtained without any additional learning phase.

Loss function also matters. Besides the most commonly used L_1 and cross-entropy loss, one usually weights different loss functions in multi-task learning to adapt to specific tasks. Several loss functions have been formulated to deal with the imbalanced problem. For example, weighted cross-entropy loss [51], Focal loss [52], Dice loss [53], and Tversky loss [54]. In practice, however, preserving large foreground is much easier than small foreground under the same method. We aim to construct a new loss function (c.f. Section III-E) to preserve the foreground with a small size and to balance the size diversity of foreground instances.

We integrate two streaming (the original frame I_T and its 3D optical Flow $Hop(T)$) and then embed their features in CrossNet. The overall architecture of the proposed cross-scene background subtraction network is illustrated in Figure 1, which integrates multiple CmDFF modules in the up-sampling process to fuse the features of the encoder and the decoder, where E_i and Op_i are featured in the encoder for extracting appearance and optical flow features respectively in each layer. D_i are the decoder features, and Att_i refer to the CmDFF modules in the corresponding layers. CrossNet uses an encoder and decoder-based fully convolution structure as the backbone network. The image and its optical flows are fed to their respective encoders and then output the foreground segmentation mask 256×256×1. The encoder and decoder structures are entirely symmetrical, with 8 convolution layers and 16 convolution layers. The step sizes of the down-sampling (encoder) and up-sampling (decoder) convolution are 2. The width and length of the encoder's feature map are reduced to $1/2$ of the original one when it passes through a convolution layer. The length and width of the decoder's feature map are doubled after each up-sampling step. The model also uses convolution with step size 1 in the decoder layer when integrated with the CmDFF modules.

C. 3D Hierarchical Optical Flows

As an instantaneous motion cue, optical flow lacks sufficient stability when representing motion, which is mainly manifested in the following two aspects: 1) The motion vector is computed based on different gray levels of pixels; when the contrast is low, the motion vector would be invalid like the hole in Figure 4 $Op(\tau_1)$ and $Op(\tau_2)$; 2) The time interval between adjacent frames is very short, yet the foreground's movement speed is random, which leads to weak responses to the foreground with slow motion. Optical flow with a long interval has the object's long-term motion cues, but an object's outline is imprecise, while optical flow from a short interval has weak responses to the foreground with slow motion such as $Op(\tau_1)$ in Figure 1.

As illustrated in Figure 1, 3D-HOP aims to solve the problem that optical flows from adjacent image frames insufficiently describe motion cues. The current frame and neighboring frames with different lengths formulate three types of optical flows to complement each other. Suppose that the current frame is at time T, and the frames at time $T-\tau_1$, $T-\tau_2$ and $T-\tau_3$ using the intervals τ_1, τ_2 and τ_3, respectively. After calculating the optical flows at time T, denoted as $Op(\tau_1)$, $Op(\tau_2)$ and $Op(\tau_3)$, we assign the optical flows with different intervals in three channels to build hierarchical optical flows $Hop(T)$. We use selflow [49] to calculate all the optical flows since this method is a self-supervised method, which helps promote it into a general cross-scene training task without the need for complex optical flow annotation of video frames. We also prepared samples generated by the latest optical flow methods [55]–[57] for performance comparison. The performance comparison details of various optical flows will be detailed in the experiments.

D. Cross-modal Dynamic Feature Filters (CmDFF)

There are several encoder-decoder networks designed for image segmentation. The encoder-decoder networks mainly consider using different scales of features and gradually recovering sharp object boundaries in the decoder path. Most use bilinear upsampling directly, which lacks information sharing of feature maps at different levels and could harm spatial localization recovery. Unlike the above scheme, our task is to learn cross-modal (appearance and motion) features, which require higher representation and generalization capabilities for feature learning. Inspired by the work of global attention upsample [58], high-level features with abundant category information can be used to weight low-level information to select precise resolution details. Different from [58], the proposed CmDFF merges the decoder and encoder features through dense attention processes during the decoder phase. In detail, high-level features guide the CmDFF modules with global information to weight useful low-level features, which contribute to the prediction in the image. Meanwhile, the encoder's features are re-weighted by the decoder's layers at the pixel level and then concatenated with the latter. The proposed cross-modal dynamic feature filter module is shown in Figure 2. The decoding process is from the previous decoding layer D_{i-1} to the next layer D_i. In this process, the CmDFF module weights the decoder's features through the encoder. Inside the CmDFF module, the input includes E_i and Op_i, and the decoder's previous layer D_{i-1}. The output is a decoder layer D_i. To explain the operation mechanism of CmDFF clearly, we use B_w, $B_{\text{up\_sampling}}$ and B_{e_op} as the intermediate processes. When we have obtained two feature maps $E_i \in \mathbb{R}^{H\times W\times C}$ and $Op_i \in \mathbb{R}^{H\times W\times C}$ (H and W are the height and width of the input feature map, and C indicates the channel index of the feature map. To get D_i, we first

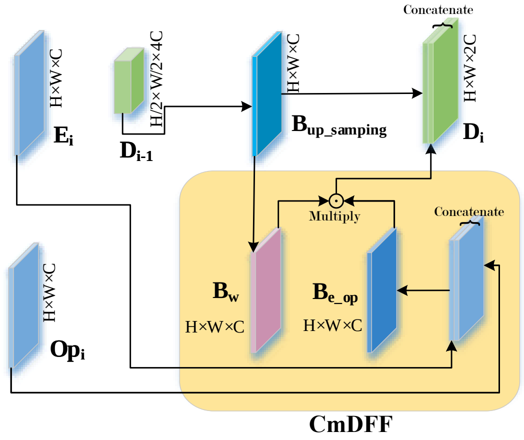

Fig. 2: Cross-modal dynamic feature filter (CmDFF). The decoding process is from the previous decoding layer D_{i-1} to the next layer D_i. In this process, the CmDFF module weights the decoder's features through the encoder. Inside the CmDFF module, the input includes E_i and Op_i, and the decoder's previous layer D_{i-1}. The output is a decoder layer D_i.

concatenate the two feature maps E_i and Op_i from two encoders. After concatenating, the channel becomes twice ($2C$) as much as the original one (C), and then $B_{e\_op} \in \mathbb{R}^{H\times W\times C}$ is obtained by convolution:

$$B_{e\_op} = \text{conv}_0(\text{Relu}(E_i \ominus Op_i)), \tag{1}$$

where conv_0 denotes a 3×3 convolution used to extract appearance features and reduce channels, $\ominus$ denotes the feature map concatenation along the channel dimension, and Relu is the active function.

We consider two ways of fusing appearance features E_i, and optical flow features Op_i. One of the processes is to fuse the two input features according to the corresponding pixel addition, just like the fusing from in [59]; the other way is to concatenate the two features in the channel dimension: the corresponding number of the channels is doubled, and then use kernel 3×3 convolution with step size 1 to reduce the channel dimension to the original number. Because the channels in our network express different feature information, the fusion method of directly adding corresponding channels cannot guarantee that the semantic information expressed by the features involved in the operation is consistent [60]. So, we concatenate the outputs of different channels and then use the subsequent network layer to learn the channel association.

In the decoding layer $D_{i-1} \in \mathbb{R}^{H/2\times W/2\times 4C}$, we have $B_{\text{up\_sampling}} \in \mathbb{R}^{H\times W\times C}$ by undertaking up-sampling convolution. Then, the weighted coefficient tensor $B_w \in \mathbb{R}^{H\times W\times C}$ (having been normalized to the range of 0 and 1) is obtained from convolution and activation operations:

$$B_w = \sigma(BN(\text{conv}_1(\text{Relu}(B_{\text{up\_sampling}})))), \tag{2}$$

where σ is the Sigmoid function, conv_1 is the convolution of kernel 3×3 and step 1 to learn the weighted coefficient and BN is batch normalization (BN). B_w is combined with the feature map B_{e_op} by multiplying pixel to obtain the weighted feature map. After batch normalization, we get the decode's features from $B_{\text{up\_sampling}}$. To prevent over-fitting and improve the robustness of the network, we also add the Dropout operation to the original decoder. Each node has a 50% probability of being suppressed in the training process, and we remove this dropout operation in network inference. The weighted feature map of the encoder and the features of the decoder are concatenated to obtain $D_i \in \mathbb{R}^{H\times W\times 2C}$ in the i^{th} decoding layer.

$$D_i = (B_w \odot B_{e\_op}) \ominus BN(\text{Dropout}(B_{\text{up\_sampling}})), \tag{3}$$

where $\odot$ denotes the Hadamard product operation. The cross-modal dynamic feature filter (CmDFF) realizes learning cross-modal (appearance and motion) features with higher representation and generalization capabilities for feature learning.

E. Class-In Scale Focal (CS-Focal) Loss

In background subtraction, there are two types of imbalance problems. One is the foreground/background imbalance (cross-class imbalance), where the background pixels dominate the whole image. Another case is that large objects dominate training, a kind of intra-class imbalance.

1) Focal Loss : Focal Loss [52] is proposed for solving the first type of unbalanced problem, which is based on the cross entropy function and expressed as:

$$\mathcal{L}_{\text{weighted-CE}}(p, y) = \begin{cases} -\alpha \log(p) & y = 1 \\ -(1-\alpha)\log(1-p) & y = 0, \end{cases} \tag{4}$$

where p represents the probability of model prediction, compared to the ground truth of the foreground label $y = 1$ and the background label $y = 0$. α is the parameter matrix of the foreground and background pixel samples.

2) CS-Focal Loss: For the case where the number of the background samples is larger than that of the positive ones, α in Eq (4) is set to be a large value so that the impact of the foreground samples on the model loss function is larger than that of the negative samples. On this basis, a hard-sample adjustment factor γ is added, and Focal Loss is finally obtained as follows:

$$\mathcal{L}_{focal} = \begin{cases} -\alpha(1-p)^\gamma \log(p) & y = 1 \\ -(1-\alpha)p^\gamma \log(1-p) & y = 0, \end{cases} \tag{5}$$

where γ regulates the contribution of hard and easy samples, for the hard sample case, it will get a lower p. The smaller p is, the larger $(1-p)^\gamma$ is; thus, a relatively large loss will be generated. Similarly, for a sample case, it may have a higher p. The larger p is, the smaller $(1-p)^\gamma$ is, and the focus of model training is on the hard samples. In the experiments, $focal\ loss$ is adopted as a baseline loss for segmentation.

We generate a visual interpretation of the loss function to explore better loss functions for background subtraction in the training process. In Figure 3, for example, $Fg = 0.6$ and $Bg = 0.4$ refer to the predicted p of the foreground and the background. Due to scale differences in categories, the focal loss may increase the training weight for the foreground,

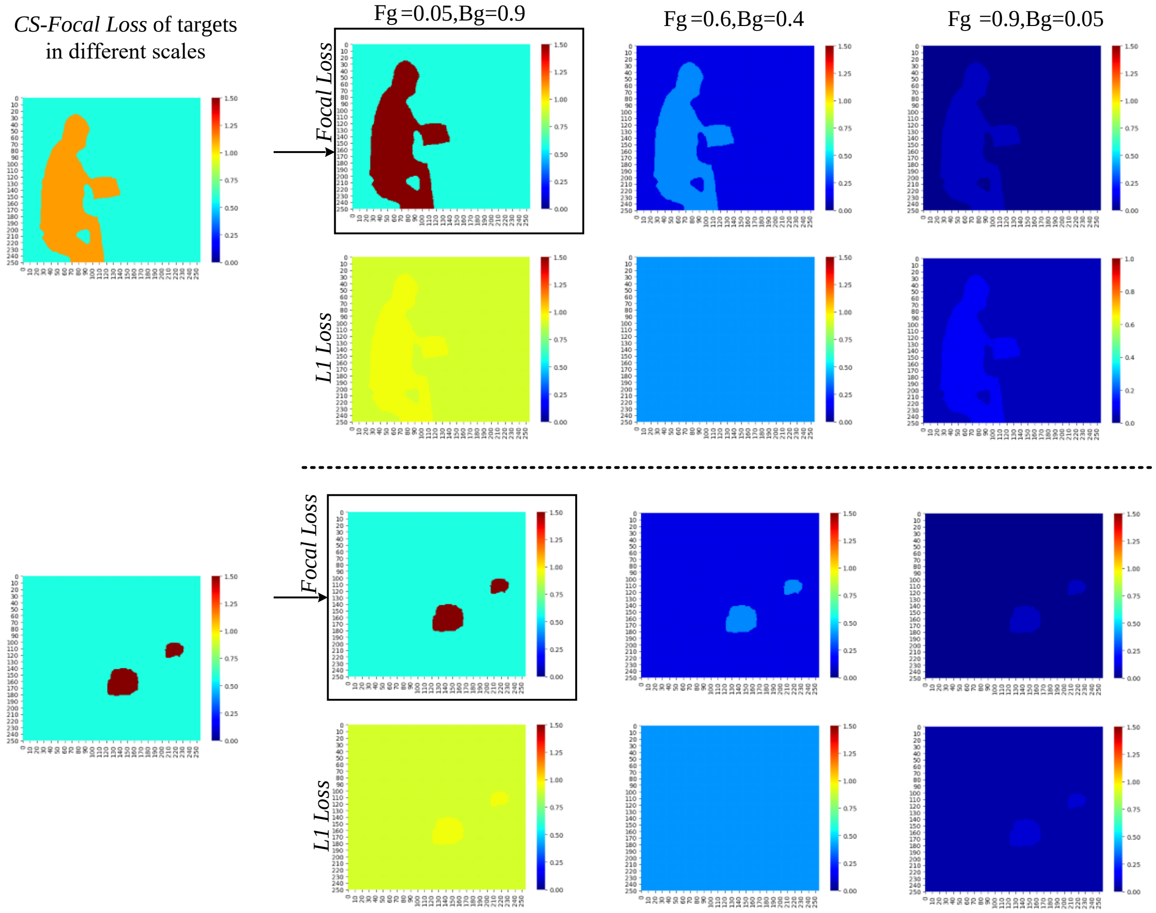

Fig. 3: Result comparisons of different loss in pixel-level. CS-Focal loss strengthens the loss of the small targets, thus effectively improving the possibility that small objects can be correctly segmented.

compared to $l1$ loss. However, foreground objects have different scales and objects with different scales must be balanced to improve the segmentation result of small objects. When $Fg = 0.9$ and $Bg = 0.05$ in Figure 3 left, the focal loss value of the human subjects is treated differently from that of the cars. We tend to improve the focal loss to balance the objects with different scales inside the foreground class based on focal loss.

First of all, we define the area ratio $S(fg)$ between the foreground and the background from one image frame and then define a balance coefficient inside class β, as follows:

$$\beta = t_3 \min(\frac{1}{S(fg)}, 50) \tag{6}$$

The reason we set the minimum value of $\frac{1}{S(fg)}$ and 50 is to prevent the potential scene from infinity, and 50 is the value set after sampling the small object area in the training images. t_3 is the normalized parameter. The proposed Class-in Scale Focal (CS-Focal) loss defined as,

$$\mathcal{L}_{CS-Focal} = \begin{cases} -\beta\alpha(1-p)^{\gamma}\log(p) & y=1 \\ -(1-\alpha)p^{\gamma}\log(1-p) & y=0. \end{cases} \tag{7}$$

From Figure 3, the visualization comparison of focal loss and CS-Focal loss is shown on the left. We can see that in the training process, CS-Focal loss strengthens the loss of the small targets, thus effectively improving the possibility that small objects can be correctly segmented. To train the model stably, we apply $l1$ loss as a standard regularization term. It is measured between the prediction p and ground truth y. The final loss function can be expressed as:

$$\mathcal{L} = t_1\mathcal{L}_{\text{CS-Focal}} + t_2\mathcal{L}_{l1}, \tag{8}$$

where t_1 and t_2 are two tunable hyper-parameters, which denote weights between the two terms in the final loss.

IV. EXPERIMENTS

In this section, we first give a brief introduction to the experimental settings, implementation details, and evaluation metrics (in Subsection IV-A). Then, we show the ablation study (in Subsection IV-B). After that, the experimental results and analyses on CDNet2014 [61], LIMU [62] and LASIESTA [63] are respectively given in Subsection IV-C and Subsection IV-D. Finally, we further show the precision-recall analysis in Subsection IV-E.

A. Training Settings and Implementation Details

Training CrossNet. Following the training setting of DeepBS [35], for the training of our model, 5% samples are randomly selected with their ground truths of each subset from

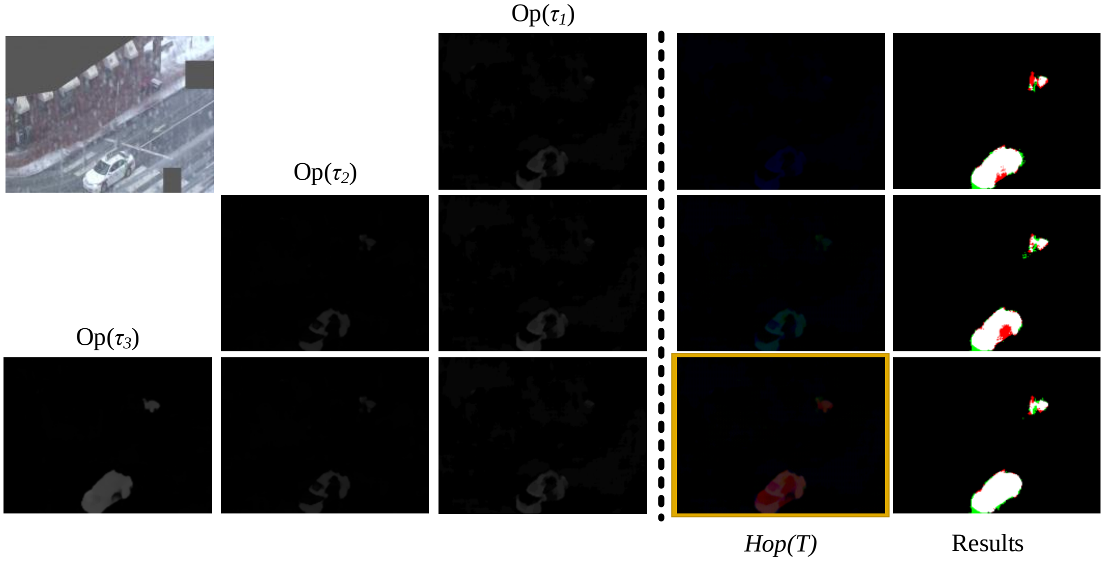

Fig. 4: Hierarchical optical flows and background subtraction results. The left is the image frame and optical flows (τ_1, τ_2, τ_3), and the right is the fused optical flows and the segmented result. Green: False Positive, Red: False Negative.

the dataset CDNet2014 to train CrossNet. The remaining 95% of samples in CDNet2014 are used as the test dataset without any overlap of the training set. The segmented foreground is also obtained without any post-processing.

Basic Optical Flow Calculation. For the proposed 3D-HOP, we utilize Selflow [49] to obtain basic optical flow. For training, Selflow extracts 10000 images of Sintel movie [64] for self-training. We infer optical flow directly in the background subtraction datasets without additional training, and the speed of inferring is 18 fps on two GTX2080Ti. To verify the impact of basic optical flow generation performance on 3D-HOP, we also prepared samples generated by the latest optical flow methods [55]–[57] for performance comparison. The performance comparison details of various optical flows will be detailed in the following part.

Hyper Parameters. All our hyperparameters settings used in the model are chosen experimentally. For the hierarchical optical flows, we set $\tau_1 = 1$, $\tau_2 = 5$, and $\tau_3 = 10$. In the proposed loss function, we set $t_1 = 0.8$, $t_2 = 0.2$, $t_3 = 0.25$, $\alpha = 0.75$, and $\gamma = 0$. The training batch size is 16, and we run 16000 epochs. Adam is used as the optimizer and its parameters beta1 = 0.95, and beta2 = 0.999. The learning rate is set to a small value 5×10^{-5}.

Comparison Methods. We add comparison methods in the experiments and divide them into three folds: 1) cross-scene deep models (single model); 2) specific-scene models (including deep models and background subtraction methods); 3) semantic segmentation models. For cross-scene deep models, STAM [59] and DeepBS [35] are trained in the same way as CrossNet. We also add experiments of the proposed model without CmDFF ($CrossNet_{noAtt}$) and Optical flow ($CrossNet_{noHOP}$). For semantic segmentation, PSPNet [36] and DeepLabV3+ [37] are trained with ADE20K [65] because there is no semantic annotation in CDNet2014. According to the protocol recommended in [66], we define some classes as foreground, including {person, car, cushion, box, book, boat, bus, truck, bottle, van, bag, and bicycle}. We compare PSPNet and DeepLabV3+ on cross-scene background subtraction because the semantic background subtraction presents potential performance advantages.

Evaluation Metrics. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are used in the evaluation.

$$Recall = TP/(TP + FN) \tag{9}$$

$$Precision = TP/(TP + FP) \tag{10}$$

$$F - Measure = 2{\times}P.{\times}R./(P. + R.) \tag{11}$$

are employed as stranded metrics for quantitative evaluation. Recall, Precision, and F-measure for image segmentation are pixel-level evaluations that accumulate all the positive and negative pixels in all the testing image frames but ignore the foreground scale, which is unfair to small regions' evaluation. Taking no account of object scale, to evaluate the results of background subtraction for small regions fairly, we use $Mean\ Dice$ based on the Dice coefficient as follows:

$$Mean\ Dice = \frac{2}{N}\sum_{i=1}^{N}\frac{(TP+FN)_i \cap (TP+FP)_i}{(TP+FN)_i \cup (TP+FP)_i} \tag{12}$$

where N is the number of the input frames that contain the foreground, $(TP + FN)_i$ is the ground truth label in frame i^{th}, $(TP + FP)_i$ is the prediction result of frame i. $Mean\ Dice$ is calculated separately for evaluating the integrity of individual foreground objects rather than using global pixel-level counting.

B. Ablation Study

In the ablations, we verify the roles of HOP, CmDFF, and CS-Focal loss function with related combinations in Table I.

TABLE I: Ablation study on CDNet2014 [61] dataset.

	$Op[\tau_1]$	$Op[\tau_2]$	$Op[\tau_3]$	CmDFF	Loss$_{CS-Focal}$	Loss$_{focal}$	Loss$_{l1}$	**F-measure**	**Mean Dice**
1	✓	✓	✓	✓	✓		✓	**.9776**	**.9466**
2	✓	✓		✓	✓		✓	.9704	.9416
3	✓			✓	✓		✓	.9642	.9368
4	✓					✓		.9433	.8747
5	✓					✓	✓	.9716	.9346
6	✓	✓	✓	✓		✓	✓	.9730	.9408
7	✓	✓	✓	✓	✓			.9735	.9423
8	✓	✓	✓	✓		✓		.9706	.9385
9	✓	✓	✓	✓			✓	.9661	.9334
10				✓	✓		✓	.9030	.8705
11	✓	✓	✓		✓		✓	.8791	.8502

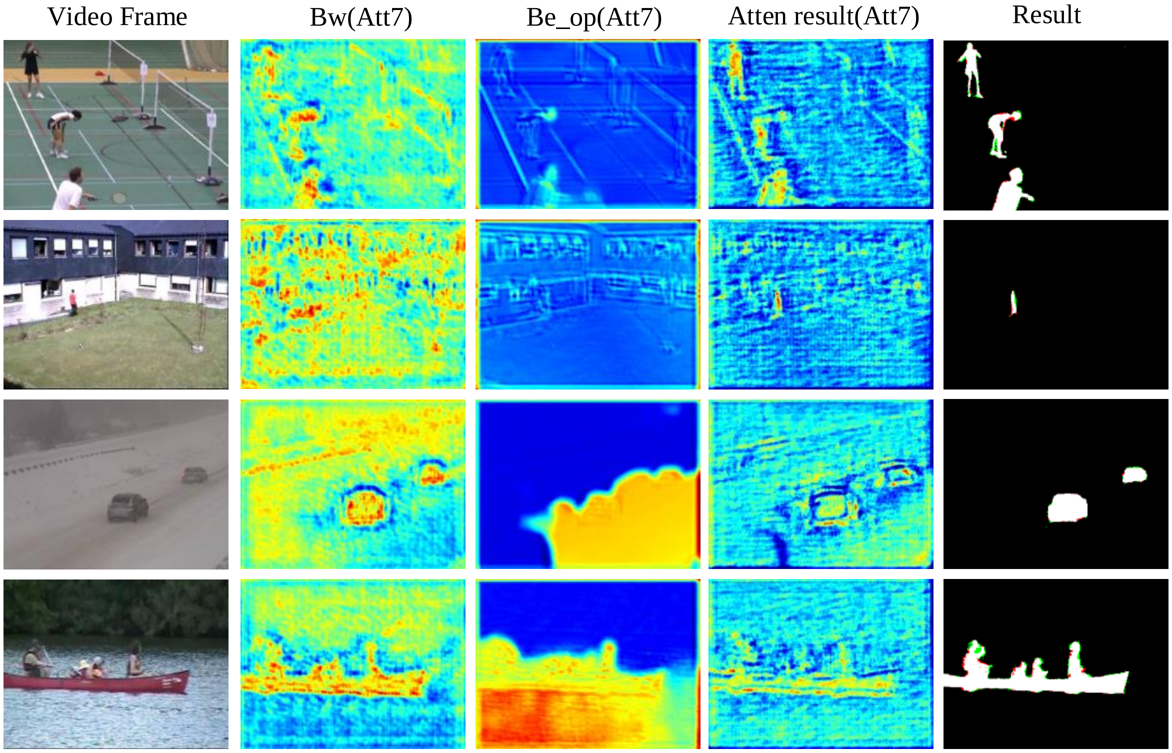

Fig. 5: Visualization of the CmDFF results. Each column has five images, including the image frame, processing results (B_w, B_{e_op}, $Attenresult$) of the CmDFF (Att7) module, and the segmentation result. Green: False Positive, Red: False Negative.

Effectiveness of 3D-HOP. Compared to the model using optical flows calculated with adjacent frames, the model using the proposed 3D-HOP gets a noticeable F-measure and Mean Dice improvement. As shown in Figure 4, hierarchical optical flows (orange border) provide sufficient motion cues to guide the background subtraction. In the scene shown in Figure 4, optical flows $Op(\tau_1)$, $Op(\tau_2)$ appear to occupy a large area inside the car. Meanwhile, $Hop(T)$ presents a relatively complete foreground area in the segmented result.

The effectiveness of CmDFF. For CmDFF, compared to the model without using CmDFF, adding it brings obvious F-measure and Mean Dice gains.

In Figure 5, we visualize the processing results of the 7^{th} (Att7) in the decoder. Because the CmDFF involves multi-channel and multi-layer processes, it is difficult to visualize the process of results directly through two-dimensional images. So we average the results of one layer in the channel dimension to reveal this trend roughly. The results of CmDFF highlight the foreground object's area, comparing the output (Att7) and the final result. B_w and B_{e_op} are the intermediate steps to get the Atten result. In B_w (Att7) and B_{e_op} (Att7), B_w and B_{e_op} present distributions of the original feature from the decoder and the encoder with object appearance and optical flows.

Effectiveness of CS-Focal loss. Compared to $focal + l1\ loss$, $CS - Focal + l1\ loss$ has noticeable improvement and achieves the best Mean Dice scores and F-measure. particularly, the gain in Mean Dice is significant. In Figure 6, the proposed loss has better performance with small objects.

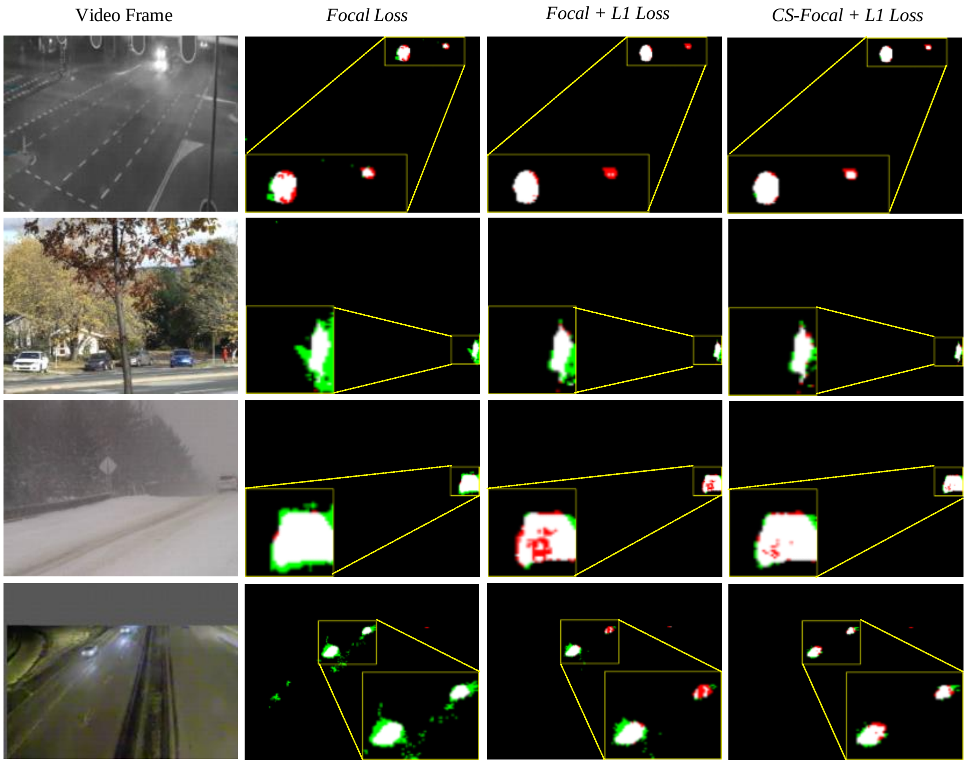

Fig. 6: Comparisons of small objects with different losses. Each row has four images including image frame, segmentation results of focal loss, focal+l1 loss, and CS-Focal+L1 loss, from left to right. Green: False Positive, Red: False Negative.

TABLE II: The average performance on CDNet2014 [61] dataset.

Method	**Mean Dice ↑**	**Recall ↑**	**Precision ↑**	**F-measure ↑**	**Model Types**
CrossNet-HOFAM	**.9466**	**.9661**	**.9893**	**.9776**	Cross-scene deep models
CrossNet$_{CmDFF}$	.8502	.8369	.9268	.8795	
CrossNet$_{noHop}$	.8705	.9297	.8789	.9036	
DeepBS [35]	.7041	.7545	.8332	.7548	
STAM [59]	.9452	.9458	.9851	.9651	
BSUV-Net 2.0 [67]	.7598	.8619	.8295	.8556	
Cascade CNN [5]	.8947	.9506	.8997	.9209	Specific-scene deep models
FgSegNet [4]	.5738	.6073	.6235	.6094	
FgSegNetV2 [68]	7544	7161	7632	.7389	
Motion U-Net [69]	.9046	.9188	.9557	.9369	
GMM [6]	.5361	.6846	.6025	.5707	Specific-scene background subtraction
CPB [7]	.6157	.7049	.6223	.6325	
SuBSENSE [8]	.6843	.8124	.7509	.7408	
RT-SBS [70]	.7341	.8507	.8064	.8280	

We utilize the color green and red to mark the false positives and false negatives in the results.

C. Results on CDNet2014

Since CrossNet has been trained on the CDNet2014 dataset, this experiment's purpose is not to test the capability of cross-scene background subtraction but to evaluate the proposed single model compared with other scene-specific training models. For the cross-scene models STAM [59], BSUV-Net 2.0 [67], and DeepBS [35] are trained in the same way as CrossNet. We also add experiments of the proposed model without CmDFF ($CrossNet_{noAtt}$) and Optical flow ($CrossNet_{noHOP}$). For specific-scene models, four deep models FgSegNet [4], FgSegNetV2 [68], Motion U-Net [69], and CascadeCNN [5], three background subtraction models GMM [6], CPB [7], RT-SBS [70] and SuBSENSE [8] are trained with a scene-specific style on CDnet2014 following their default experiment settings. For semantic segmentation models, PSPNet [36] and DeepLabV3+ [37] are trained with ADE20K [65] dataset using their default model settings.

From Table II, the Recall of CrossNet is 0.9661, the Recall of Cascade CNN with 0.9506 ranks second, and STAM ranks third with 0.9458. CrossNet improves Recall by 1.55%. For F-measure, CrossNet, as a single model, gains the best performance of F-measure with 0.9776, which is 4% better than the best specific-scene deep model Motion U-Net with F-measure 0.9369. These results indicate CrossNet could maintain high performance even with limited training data.

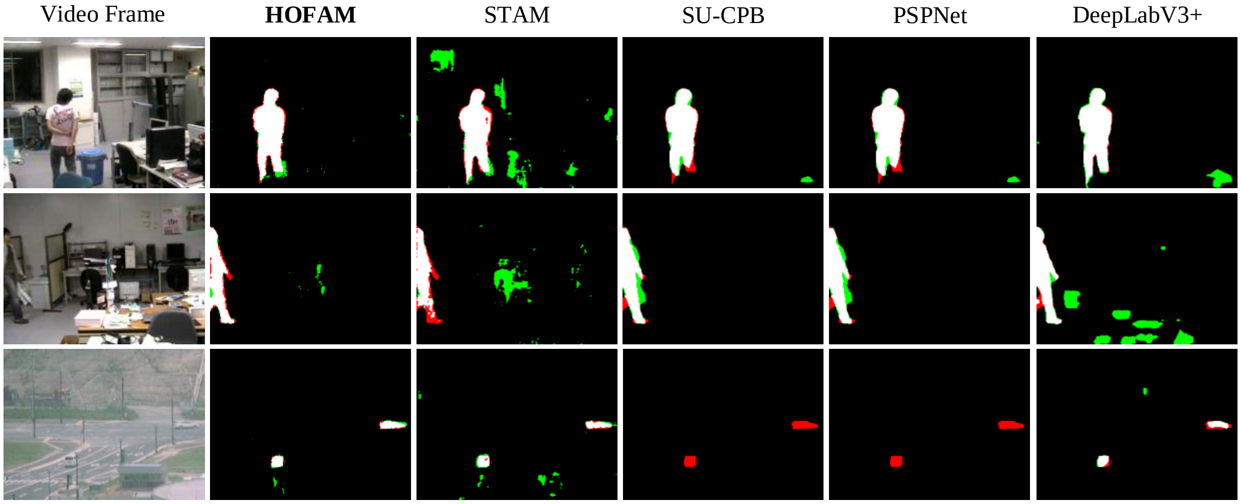

Fig. 7: Comparison on cross-scene dataset LIMU. Each column: image frame, segmentation results of the proposed CrossNet-HOFAM, SU-CPB, STAM, PSPNet, and DeepLabV3+, from left to right. Green: False Positive, Red: False Negative.

TABLE III: F-measure of different methods on LIMU [62] dataset.

Method	CameraParameter	Intersection	LightSwitch	Overall	Model Types
CrossNet-HOFAM	.7979	**.7851**	**.8493**	**.7981**	Cross-scene training on CDnet2014
CrossNet$_{noCmDFF}$	.6998	.7364	.7965	.7291	
CrossNet$_{noHOP}$	.7055	.7294	.6981	.7130	
DeepBS [35]	.6705	.5545	.6332	.6073	
STAM [59]	.7742	.6749	.7163	.7344	
Cascade CNN [5]	.1025	.0453	.0277	.0585	
FgSegNet [4]	.2668	.1428	.0414	.1503	
GMM [6]	.6372	.6423	.6743	.6519	Specific-scene background subtraction
CPB [7]	.6545	.6778	.6633	.6652	
SU-CPB [23]	.7484	.7672	.8211	.7789	
SuBSENSE [8]	.6744	.6530	.6934	.6753	
PSPNet [37]	**.8656**	.1303	.6510	.7506	Semantic training on ADE20k
DeepLabV3+ [36]	.7739	.6766	.3330	.6986	

D. *Cross-scene Test*

We compared the existing state-of-the-art cross-scene deep models, specific-scene deep models, background subtraction methods, and semantic segmentation models with CrossNet without additional training.

LIMU and LASIESTA Dataset. LIMU includes scenes with a variety of dynamic backgrounds. LASIESTA collects many typical real indoor and outdoor sequences organized into different categories, each covering a specific challenge.

Cross-scene Training Setting. We apply CrossNet, STAM, DeepBS, CascadeCNN, and FgSegNet trained on CDNet2014 as a single model to test without any additional training phase on the two cross-scene datasets. The background model GMM, CPB, SU-CPB, and SuBSENSE are trained in specific scenes on the two datasets with their default experiment settings. The semantic segmentation models PSPNet and DeepLabV3+ are trained on the semantic segmentation dataset ADE20K.

Results on LIMU.

On LIMU, from Table III, CrossNet-HOFAM performs better on two subsets than the other models. On the subset of CameraParameter, CrossNet ranks second with an F-Measure of 0.7979, compared with PSPNet with the highest F-Measure of 0.8656. Overall, CrossNet gains the best performance of F-measure 0.7981 while SU-CPB ranks second with 0.7789, and PSPNet ranks third with 0.7506. We illustrate the results of the proposed CrossNet-HOFAM, STAM, SU-CPB, PSPNet, and DeepLabV3+ in Figure 7.

Results on LASIESTA.

On LASIESTA, from Table IV as on LIMU, CrossNet is compared with the approaches presented above. Two indoor and two outdoor subsets, *i.e.*, outdoor Moving camera (O_MC), outdoor Cloudy conditions (O_CL), indoor Occlusions (I_OC), and indoor Moving camera (I_MC), are shown. Overall, CrossNet gains the best performance of F-measure 0.8072, while STAM ranks second with 0.6807. On the outdoor subsets, CrossNet gets a much higher F-measure than PSPNet and DeepLabV3+. Significantly, some scenes in LASIESTA may change slowly, which is why the methods using background subtraction with specific-scene perform less satisfactorily in some test scenes. This also reflects the effectiveness and robustness of the proposed method. Figure 8 demonstrates the above observation.

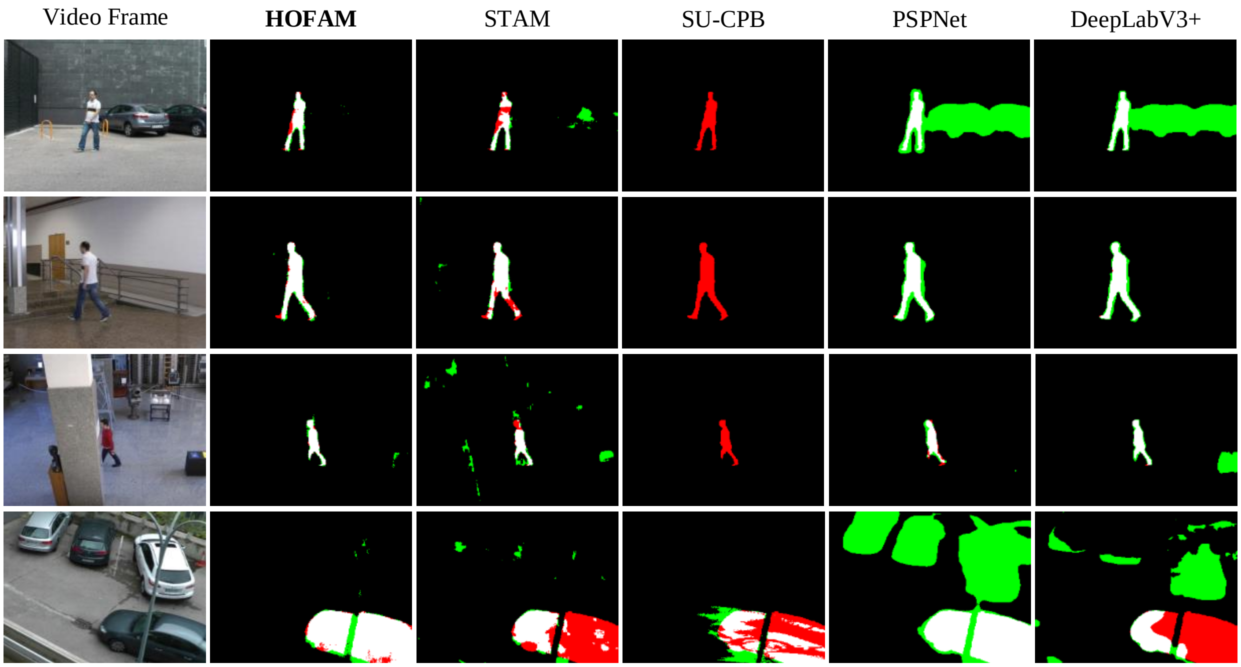

Fig. 8: Result comparisons of different models on cross-scene dataset LASIESTA. Each column: image frame, segmentation results of the proposed CrossNet-HOFAM, SU-CPB, STAM, PSPNet and DeepLabV3+, from left to right. Green: False Positive, Red: False Negative.

TABLE IV: F-measure of different methods on LASIESTA [63] dataset.

Method	O_MC	O_CL	I_OC	I_MC	Overall	Model Types
CrossNet-HOFAM	**.6919**	**.8602**	.8456	.7895	**.8072**	Cross-scene training on CDnet2014
CrossNet$_{noCmDFF}$	.5518	.6364	.7067	.5683	.6148	
CrossNet$_{noHOP}$	.5656	.6637	.6883	.6030	.6312	
DeepBS [35]	.7020	.7673	.6758	.5911	.6774	
STAM [59]	.6365	.7624	.7362	.6735	.6807	
Cascade CNN [5]	.1028	.1414	.1155	.1799	.1288	
FgSegNet [4]	.1539	.1687	.4923	.4306	.2447	
GMM [6]	.3125	.8027	.7746	.2513	.4527	Specific-scene background subtraction
CPB [7]	.2910	.8407	.8095	.0641	.4304	
SU-CPB [23]	.2803	.8430	.7823	.0722	.4412	
SuBSENSE [8]	.3029	.8327	.7412	.1164	.4425	
PSPNet [37]	.1652	.3533	**.9281**	.7086	.3723	Semantic training on ADE20k
DeepLabV3+ [36]	.1675	.2319	.8294	**.8276**	.3395	

E. Precision-Recall Analysis of HOFAM

The Precision-Recall (PR) comparisons of the proposed background subtraction framework HOFAM with other models are shown in Figure 9. We compare the results in method trained dataset (CDNet2014 [61]) and two cross-scene datasets (LIMU [62] and LASIESTA [63]). We can observe that the proposed HOFAM can achieve a better result and performs more stably than the other models in different datasets. Meanwhile, compared with those performances on CDNet2014, although all the supervised methods (HOFAM, STAM, DeepBS, CascadeCNN, and FgSegNet trained on CDnet2014) have inevitable performance degradation on cross-scene datasets LIMU and LASIESTA, the proposed HOFAM has more stable performance, which indicates that our model has better cross-scene generalization.

F. Test Speed and Model Size

The test speed of CrossNet is 5.33 fps for the image size of 256 by 256 on two GTX2080TI with 32 GB RAM, i9 CPU, and Ubuntu 16.04 LTS operating system. The entire network uses Tensorflow 1.13 version. The model size checkpoint of CrossNet is 1.6 GB. Note that because the structure of different background subtraction models is very different, including supervised and unsupervised methods, cross-scenario, and specific scenario methods, it is unfair to compare the scale of parameters between different models. In general, the model size of the proposed method is equivalent to that of STAM method. In addition, if the optical flow extraction network is considered, the model size will vary according to the difference of the primary optical flow generation model.

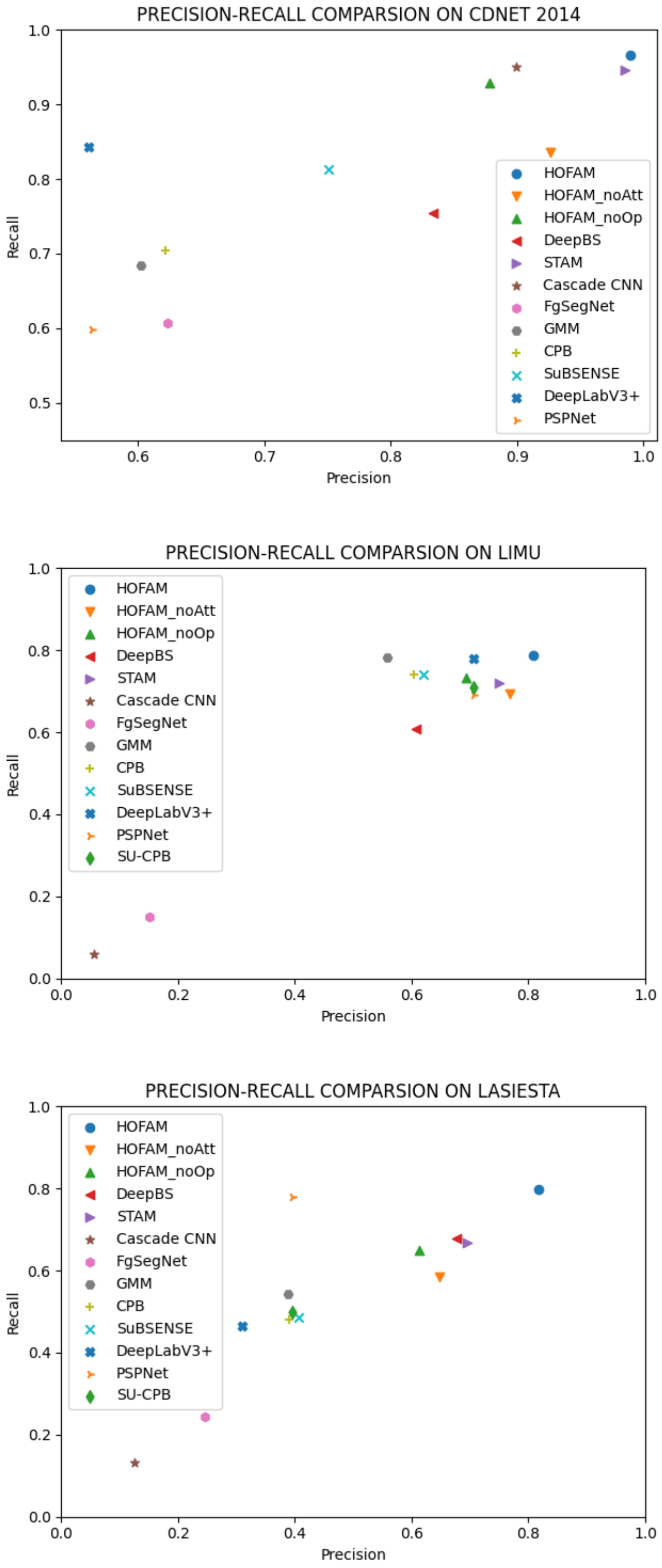

Fig. 9: Precision-Recall comparison of different models on CDNet2014 [61], LIMU [62] and LASIESTA [63].

G. Impact of Basic Optical Flow Generation Models

Without additional training, we infer basic optical flow directly in the background subtraction datasets. To just verify the impact of basic optical flow generation performance on 3D-HOP, we also prepared samples generated by the latest optical flow methods, including KPA [55], Gmflow [56], equilibrium [57], and the classical LK [40] for performance comparison. The performance comparison of various optical flows is detailed in Table V. We can observe that two of CrossNet integrated three SOTA optical flow methods exceed the performance of the current CrossNet integrated SelfFlow. This also confirms the importance of basic optical flow generation quality. We still prefer to use the self-supervised optical flow generation method Selflow [49], mainly considering its potential versatility and flexibility to build a full unsupervised background subtraction framework in our future work.

TABLE V: The average performance of F-measure on CDNet2014 using different basic optical flow generation models.

KPA [55]	Gm [56]	equilibrium [57]	Sel [49]	LK [40]
.9839	.9738	.9803	.9776	.9295

V. Conclusion and Future Work

We have proposed a method to realize cross-scene background subtraction. Compared to the existing state-of-the-art cross-scene deep CNNs, specific-scene deep CNNs, traditional background subtraction methods, and modern semantic segmentation models on CDNet2014, LIMU and LASIESTA benchmarks, CrossNet has shown promising generalization to discriminate the scene-independent motion patterns without any additional training for a new scene.

The limitation of this work is that we fixed the interval settings of 3D-HOP for efficient training. Since the model has to be retrained with optical flows at different intervals, we intend to develop a more flexible way to use an adaptive interval of 3D-HOP. In future work, realizing a complete self-supervised learning manner could extend the flexibility further. In addition, applying the proposed module to other computer vision tasks is also a potential direction.

References

[1] W. Hu, X. Li, X. Zhang, X. Shi, S. Maybank, and Z. Zhang, "Incremental tensor subspace learning and its applications to foreground segmentation and tracking." in *International Journal of Computer Vision (IJCV)*, 2011.

[2] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," in *IEEE Transactions on Image Processing (TIP)*, 2004.

[3] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *CVPR*, 2020.

[4] L. A. Lim and H. Y. Keles., "Foreground segmentation using convolutional neural networks for multiscale feature encoding," in *Pattern Recognition (PR)*, 2018.

[5] Y. Wang, Z. Luo, and P. Jodoin, "Interactive deep learning method for segmenting moving objects." in *Pattern Recognition (PR)*, 2017.

[6] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking." in *CVPR*, 1999.

[7] W. Zhou, K. Shun'ichi, H. Manabu, S. Yutaka, and D. Liang, "Foreground detection based on co-occurrence background model with hypothesis on degradation modification in dynamic scenes," in *Signal Processing (SP)*, 2019.

[8] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," in *IEEE Transactions on Image Processing (TIP)*, 2014.

[9] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," in *Neurocomputing*, 2018.

[10] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *CVPR*, 2019.

[11] Y. C. Chen, Y. Y. Lin, M. H. Yang, and J. B. Huang, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *CVPR*, 2020.

[12] Z. Li and D. Hoiem, "Learning without forgetting," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[13] L. Wang and K. J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[14] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn : Towards general solver for instance-level low-shot learning," in *CVPR*, 2020.

[15] 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 *CVPR*, 2016.

[16] A. Zlateski, R. Jaroensri, P. Sharma, and F. Durand, "On the importance of label quality for semantic segmentation," in *CVPR*, 2018.

[17] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," in *ICCV*, 2019.

[18] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *CVPR*, 2018.

[19] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, X. Zhao, and Y. Satoh, "Co-occurrence-based adaptive background model for robust object detection," in *AVSS*, 2013.

[20] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, and X. Zhao, "Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes." in *Pattern Recognition (PR)*, 2015.

[21] D. Liang, S. Kaneko, H. Sun, and B. Kang, "Adaptive local spatial modeling for online change detection under abrupt dynamic background." in *ICIP*, 2018.

[22] D. Liang and X. Liu, "Coarse-to-fine foreground segmentation based on co-occurrence pixel-block and spatio-temporal attention model," in *ICPR*, 2021.

[23] D. Liang, B. Kang, X. Liu, P. Gao, X. Tan, and S. Kaneko, "Cross-scene foreground segmentation with supervised and unsupervised model communication," in *Pattern Recognition (PR)*, 2021.

[24] D. Liang, Z. Wei, H. Sun, and H. Zhou, "Robust cross-scene foreground segmentation in surveillance video," in *ICME*, 2021.

[25] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body." in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1997.

[26] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," in *Proceedings of the IEEE*, 2002.

[27] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," in *IEEE Transactions on Image Processing (TIP)*, 2011.

[28] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes." in *CVPR*, 2010.

[29] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection." in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI))*, 2005.

[30] T. Huynh-The, O. Banos, S. Lee, B. H. Kang, E. S. Kim, and T. Le-Tien, "Nic: A robust background extraction algorithm for foreground detection in dynamic scenes," in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2016.

[31] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *SMC*, 2016.

[32] Y. Wang, L. Zhu, and Z. Yu, "Foreground detection for infrared videos with multiscale 3-d fully convolutional network." in *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 2018.

[33] P. W. Patil and S. Murala, "Msfgnet: A novel compact end-to-end deep network for moving object detection." in *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[34] J. García-González, J. de Lazcano-Lobato, R. Luque-Baena, M. Molina-Cabello, and E. López-Rubio, "Foreground detection by probabilistic modeling of the features discovered by stacked denoising autoencoders in noisy video sequences." in *Pattern Recognition (PR)*, 2019.

[35] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," in *Pattern Recognition (PR)*, 2018.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network." in *CVPR*, 2017.

[37] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation." in *ECCV*, 2018.

[38] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *ICCV*, 2019.

[39] H. Ding, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *CVPR*, 2018.

[40] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, 1981.

[41] 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 *CVPR*, 2017.

[42] J. Pan and D. Liang, "Holistic crowd interaction modelling for anomaly detection," in *Biometric Recognition*, J. Zhou, Y. Wang, Z. Sun, Y. Xu, L. Shen, J. Feng, S. Shan, Y. Qiao, Z. Guo, and S. Yu, Eds. Cham: Springer International Publishing, 2017, pp. 642–649.

[43] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion." in *SCIA*, 2003.

[44] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, and T. Brox, "Flownet: Learning optical flow with convolutional networks." in *ICCV*, 2015.

[45] A. Bar-Haim and L. Wolf, "Scopeflow: Dynamic scene scoping for optical flow." in *CVPR*, 2020.

[46] P. Truong, M. Danelljan, and R. Timofte, "Glu-net: Global-local universal network for dense flow and correspondences." in *CVPR*, 2020.

[47] T. W. Hui and C. C. Loy, "Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation." in *arXiv*, 2020.

[48] S. Wang, Y. Zhou, J. Yan, and Z. Deng, "Fully motion-aware network for video object detection." in *ECCV*, 2018.

[49] P. Liu, M. Lyu, I. King, and J. Xu, "Selflow: Self-supervised learning of optical flow." in *CVPR*, 2019.

[50] X. Zhu, Y. Wang, J. Dai, and L. Yuan, "Flow-guided feature aggregation for video object detection," in *ICCV*, 2017.

[51] S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, and S. Vitaladevuni, "Multi-task learning and weighted cross-entropy for dnn-based keyword spotting." in *Interspeech*, 2016.

[52] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection." in *ICCV*, 2017.

[53] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation." in *3DV*, 2016.

[54] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks." in *MLMI*, 2017.

[55] A. Luo, F. Yang, X. Li, and S. Liu, "Learning optical flow with kernel patch attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8906–8915.

[56] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8121–8130.

[57] S. Bai, Z. Geng, Y. Savani, and J. Z. Kolter, "Deep equilibrium optical flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 620–630.

[58] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation." in *BMVC*, 2018.

[59] D. Liang, J. Pan, H. Sun, and H. Zhou, "Spatio-temporal attention model for foreground detection in cross-scene surveillance videos," in *Sensors*, 2019.

[60] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[61] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset." in *CVPR*, 2012.

[62] U. Kyushu, "Limu," in *http://limu.ait.kyushu-u.ac.jp/dataset/en/*, 2008.

[63] C. Cuevas, E. M. Yáñez, and N. García., "Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta," in *CVIU*, 2016.

[64] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012.

[65] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," in *International Journal of Computer Vision (IJCV)*, 2019.

[66] M. Braham, S. Pierard, and M. Van Droogenbroeck, "Semantic background subtraction," in *ICIP*, 2017.

[67] M. O. Tezcan, P. Ishwar, and J. Konrad, "Bsuv-net 2.0: Spatio-temporal data augmentations for video-agnosticsupervised background subtraction," *CoRR*, vol. abs/2101.09585, 2021.

[68] F. Gao, Y. Li, and S. Lu, "Extracting moving objects more accurately: A cda contour optimizer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4840–4849, 2021.

[69] G. Rahmon, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Motion u-net: Multi-cue encoder-decoder network for motion segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 8125–8132.

[70] A. Cioppa, M. V. Droogenbroeck, and M. Braham, "Real-time semantic background subtraction," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3214–3218.

ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program of China (2020YFB1313900), NSFC General Project 62272229 and 62072452, the Natural Science Foundation of Jiangsu Province under grant BK20222012, the Regional Joint Fund of Guangdong under Grant 2021B1515120011, and Shenzhen Science and Technology Program JCYJ20200109115627045. Dong Liang and Dong Zhang contributed equally to this work.

Dong Liang received a B.S. in Telecommunication Engineering and an M.S. in Circuits and Systems from Lanzhou University, China, in 2008 and 2011, respectively. In 2015, he received Ph.D. at the Graduate School of IST, Hokkaido University, Japan. He is an associate professor at the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include robust pattern recognition and large-scale video streaming processing. He has published several research papers in IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, and AAAI. He was awarded the Excellence Research Award from Hokkaido University in 2013, the Best Student Paper Award in International Symposium on Optomechatronic Technology (ISOT) 2013, and the Outstanding Contribution Award from the China Conference on Biometric Recognition 2021.

Dong Zhang received a BS degree and an MS in Computer Science and Technology at Nanjing Forestry University, China. He is now a Ph.D. student at the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include machine learning and computer vision, especially in object detection, semantic segmentation, video object segmentation, and cross-scene segmentation. He has published several journal and conference papers in these areas, including IEEE Transactions on Cybernetics, IEEE Transactions on Geoscience and Remote Sensing, AAAI, IJCAI, ECCV, and NeurIPS.

Qiong Wang is a Researcher with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. Her research interests include VR applications in medicine, visualization, medical imaging, human-computer interaction.

Zongqi Wei received a BS degree in 2018 in Computer Science and Technology at Nanjing Post and Telecommunication University, China. He is now a master student at the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include computer vision, especially object detection and video object segmentation. He is currently a research intern with ByteDance research center Beijing.

Liyan Zhang received a Ph.D. degree in computer science from the University of California, Irvine, Irvine, CA, USA, in 2014. She is currently a Professor at the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include multimedia analysis, computer vision, and deep learning. Dr. Zhang received the Best Paper Award from the International Conference on Multimedia Retrieval (ICMR) 2013 and the Best Student Paper Award from the International Conference on Multimedia Modeling (MMM) 2016.