Multispectral Video Semantic Segmentation: A Benchmark Dataset and Baseline

Wei Ji 1 Jingjing Li 1,* Cheng Bian 2 Zongwei Zhou 3
Jiaying Zhao 2 Alan Yuille 3 Li Cheng 1

1University of Alberta 2ByteDance 3Johns Hopkins University

{wji3, jingjin1, lcheng5}@ualberta.ca, zzhou82@jh.edu, ayuille1@jhu.edu

https://jiwei0921.github.io/Multispectral-Video-Semantic-Segmentation

Abstract

Robust and reliable semantic segmentation in complex scenes is crucial for many real-life applications such as autonomous safe driving and nighttime rescue. In most approaches, it is typical to make use of RGB images as input. They however work well only in preferred weather conditions; when facing adverse conditions such as rainy, overexposure, or low-light, they often fail to deliver satisfactory results. This has led to the recent investigation into multispectral semantic segmentation, where RGB and thermal infrared (RGBT) images are both utilized as input. This gives rise to significantly more robust segmentation of image objects in complex scenes and under adverse conditions. Nevertheless, the present focus in single RGB image input restricts existing methods from well addressing dynamic real-world scenes.

Motivated by the above observations, in this paper, we set out to address a relatively new task of semantic segmentation of multispectral video input, which we refer to as Multispectral Video Semantic Segmentation, or MVSS in short. An in-house MVSeg dataset is thus curated, consisting of 738 calibrated RGB and thermal videos, accompanied by 3,545 fine-grained pixel-level semantic annotations of 26 categories. Our dataset contains a wide range of challenging urban scenes in both daytime and nighttime. Moreover, we propose an effective MVSS baseline, dubbed MVNet, which is to our knowledge the first model to jointly learn semantic representations from multispectral and temporal contexts. Comprehensive experiments are conducted using various semantic segmentation models on the MVSeg dataset. Empirically, the engagement of multispectral video input is shown to lead to significant improvement in semantic segmentation; the effectiveness of our MVNet baseline has also been verified.

1. Introduction

As a fundamental computer vision problem, semantic segmentation concerns the assignment of category labels to each pixel in an image. It has received extensive research attention over the past decades [2, 5, 12, 36, 45, 64, 74, 80]. Existing semantic segmentation networks are predominantly designed to work with RGB images, which may fail in the presence of adverse conditions, such as rainy, low-light, or overexposure. On the other hand, we have evidenced a growing demand in using thermal images for semantic segmentation; a number of RGBT models have been subsequently developed, to engage both RGB and thermal images.
as input for semantic segmentation especially with complex scenes [21, 55, 76, 82, 83]. This may be attributed to the fact that thermal infrared imaging is relatively insensitive to illumination conditions, as it works by recording infrared radiations of an object above absolute zero temperature [19].

It is worth noting that the existing RGBT segmentation methods are based on single images. However, the lack of mechanism to account for the temporal contexts may limit their performance when working with video inputs containing dynamic scenes, which are omnipresent in our daily lives. This leads us to explore in this paper a relatively new task of Multispectral Video Semantic Segmentation, or in short MVSS, with a specific focus on RGB video inputs. Fig. 1 illustrates several exemplar multispectral video sequences and their ground-truth semantic annotations. As shown, the RGB frames and thermal frames provide rich and often complementary information for identifying moving foreground objects and static background scenes in low-light night or facing strong headlights. The new task opens up possibilities for applications that require a holistic view of video segmentation under challenging conditions, e.g., autonomous safe driving, nighttime patrol, and fire rescue. To our knowledge, this is the first work to address such multispectral video semantic segmentation problem.

In the deep learning era, benchmark datasets have become the critical infrastructure upon which the computer vision research community relies to advance the state-of-the-arts. Thanks to the publicly available benchmarks, such as MFNet [21], PST900 [55], Cityscapes [12], and CamVid [4], the related tasks of multispectral semantic segmentation (MSS) and video semantic segmentation (VSS) have evidenced notable progresses. Meanwhile, these existing datasets provide as input either single pairs of RGB and thermal images, or RGB only video sequences. There unfortunately lacks a suitable dataset to train and evaluate learning-based models for the proposed MVSS task. This leads us to curate a high-quality and large-scale MVSS dataset, referred to as MVSeg, that contains diverse situations. Specifically, our MVSeg dataset comprises 738 synchronized and calibrated RGB and thermal infrared video sequences, with a total of 52,735 RGB and thermal image pairs. Among them, 3,545 image pairs are densely annotated with fine-grained semantic segmentation labels, consisting of a rich set of 26 object categories in urban scenes. In particular, as showcased in Fig. 1, our MVSeg dataset involves many challenging scenes with adverse lighting conditions. It is expected to provide a sufficiently realistic benchmark in this field.

Furthermore, a dedicated baseline model is developed for this new task, which is called Multispectral Video Semantic segmentation NETwork or simply MVNet. Our MVNet possesses two key components in addressing the main challenges of MVSS task. Considering the high complexity of processing large-volume multispectral video data, a prototypical MVFuse module is devised to attend to rich contextual multispectral video features with a moderate memory footprint. A novel MVRegulator loss is further introduced, which regularizes the feature learning process to reduce cross-spectral modality difference and promote better exploitation of multispectral temporal data. Comprehensive experiments on various state-of-the-art semantic segmentation models are also carried out at the MVSeg dataset. Experimental results demonstrate the significance of multispectral video data for semantic segmentation, and verify the effectiveness of our MVNet model. We expect the MVSeg dataset and the MVNet baseline will facilitate future research activities toward the MVSS task.

2. Related Work

In this section, we review the most relevant literature in RGB semantic segmentation, multispectral semantic segmentation, and video semantic segmentation.

RGB Semantic Segmentation has achieved remarkable progress, driven by the availability of large-scale datasets (e.g., Cityscapes [12]), rapid evolution of convolutional networks (e.g., VGG [57] and ResNet [23]) and segmentation models (e.g., FCN [45]), and its wide applications (e.g., medical diagnosis [3, 10, 31]). In particular, FCN [45] is a pioneer work, which adopts fully convolutional networks to perform per-pixel representation learning. Since then, many other methods [1, 5–7, 17, 25, 42, 48, 53, 73, 75, 77, 85] have been proposed to increase the receptive field or representation ability of the network. Recently, vision transformer has been popular for semantic segmentation [11, 44, 58, 65, 68] by capturing global context [66], which give rise to impressive performance. Though tremendous progress has been made in image segmentation, the RGB-based models often perform less well when faced with adverse conditions, e.g., darkness or dim light.

Multispectral Semantic Segmentation are recently gaining grounds in addressing issues arising from traditional RGB models by incorporating multimodal visible and thermal images (RGBT). There is also another line of research focusing on multimodal RGBD-based segmentation [20, 32–35, 38, 39, 52, 54, 56, 67, 71, 79, 81], addressing limitations of RGB-based segmentation to some extent; interested readers can refer to surveys [16, 63] for more details. In terms of the MSS task, two challenging benchmark datasets have been released by MFNet [21] and PST900 [55], that are captured in adverse environments, such as nighttime road, underground tunnels, and caves. Based on them, many RGBT models [14, 21, 55, 61, 62, 76, 82–84] have been designed to leverage RGB and thermal imagery for semantic segmentation. Typically, two-stream encoders and one decoder are used to extract features from two modalities and decode semantic representations. Meanwhile, various fu-
sion strategies are designed to integrate multispectral information, e.g., concatenating the outputs of the encoders [21], fusing thermal feature into RGB encoder via element-wise summation [61], and more complex two-stage fusion [62], bridging-then-fusing strategy [76], attention fusion [69], and hierarchy shallow & deep fusion [84]. In addition, [82] improves edge quality by introducing edge prior into RGBT segmentation network. These methods are shown to gain robustness against poor lighting conditions thanks to the complementary RGB and thermal imagery.

Video Semantic Segmentation becomes attractive in practical applications that deal with videos rather than single images [26,40,41,51,59,78]; it aims to infer pixel-wise classes in each frame of a video. Due to the prohibitive cost of densely annotating video frames, most VSS datasets (e.g., Cityscapes [12] and CamVis [4]) compromise to annotate a sparse set of selected frames of the entire video. VSPW [46] adopts a segmentation model [72] to propagate sparse human annotations of key frames to adjacent unlabelled ones, then refine the annotations artificially. With these datasets, a series of VSS models have been proposed to explore temporal contexts of video frames for improving semantic segmentation. Among them, a group of works [18, 30, 47, 70] rely on optical flow [15] to warp features from neighboring frames for feature alignment and aggregation, which usually suffer from sub-optimal performance due to the error-prone optical flow estimation. Recently, attention-based approaches have been explored [40, 51, 59, 60], which attentively select relevant information from past reference frames to help segment the target one, producing promising results. In this paper, we extend this attention scheme to build a tailored MVSS model.

3. MVSS Benchmark Dataset

In this section, we focus on describing the construction of the MVSeg dataset, and analyzing the statistical results.

3.1. Dataset Construction

Data collection. Our goal is to collect a large-scale dataset with calibrated visible (RGB) and thermal infrared video sequences, covering a diverse set of challenging scenes, with high-quality dense annotations. We gathered RGB-thermal videos from multiple sources in related works, including OSU [13], INO [29], RGBT234 [37], and KAIST [28], and manually selected 738 high-quality video shots (5 seconds on average) to build our MVSeg dataset. Most of these videos are at the resolution of 480×640. This dataset covers many complex scenes during daytime, nighttime, normal weather conditions (e.g., sunny and cloudy), and adverse weather conditions (e.g., rainy, snowy and foggy). We illustrate several visual examples in Fig. 1, and more visualizations can be found in supplementary materials.

Classes and annotations. To identify object classes of interest, we carefully reviewed all paired videos of both RGB and thermal modes, and collected all object classes that appeared in the dataset. Then 26 object classes of interest were selected for annotation, which were grouped into 8 root categories, including vehicle, human, flat, nature, object, building, sky, and background (unlabeled pixels), as illustrated in Fig. 2. Guided by [12], criteria for selecting classes were based on their frequency, relevance to the applications, practical considerations for annotation efforts, and promoting compatibility with existing datasets, e.g., [12, 21].

Labeling the MVSeg dataset poses greater challenges compared to RGB-based segmentation datasets. Firstly, the MVSeg dataset contains many challenging scenes recorded under adverse conditions, which complicates the identification of less visible objects and the differentiation of their silhouettes. To assist annotators, we display RGB and thermal image pairs side by side, synchronizing their annotation traces to provide useful reference information. Secondly, we strive for consistent annotations between adjacent frames in a video by presenting a “global” view of annotated frames within each video. This allows inspectors
to more easily spot missing objects and inconsistent annotations. Despite these efforts, the annotation and quality control process for the MVSeg dataset still remains time-consuming, averaging over 50 minutes per video frame due to the intricate nature of dense pixel-level semantic labeling and the challenging scenes it encompasses.

Dataset splits. The dataset is split into training, validation, and test sets, which consist of 452/84/202 videos with 2,241/378/926 annotated image pairs, respectively. The entire test set is also divided into daytime and nighttime scenes (134/68 videos), to make a comprehensive evaluation.

3.2. Statistical Analysis

Table 1 shows an overview of the statistical results of the proposed MVSeg dataset and related MSS/VSS datasets. Our MVSeg dataset contains 738 multispectral videos at a frame rate of 15 f/s, including 53K image pairs in total and 3,545 annotated image pairs of 26 categories. Similar to other VSS datasets (Cityscapes [12] and CamVid [4]), we annotate one frame for every 15 frames. We may notice that our MVSeg dataset and the MSS datasets (MFNet [21] and PST900 [55]) have fewer annotated GTs than VSS datasets. This is reasonable due to the scarcity of calibrated multispectral images/videos and the difficulty of annotating such data. Meanwhile, our MVSeg dataset has comparable or richer object categories compared to MSS & VSS datasets. Fig. 2 illustrates the detailed object sub- and root-categories in MVSeg, and plots the number of frames in each category. It shows that the distribution is unbalanced between each class, similar to any other semantic segmentation datasets. The common categories, e.g., car and pedestrian, appear in most of frames. Table 2 lists the pixel-wise annotate rate for each root-category in the multispectral-based datasets. It is shown that existing MSS datasets [21,55] only label a small fraction of pixels in a scene (7.86% and 3.02%, respectively). In comparison, our MVSeg dataset has a high pixel-wise annotation rate of 98.96%, which is more meaningful for understanding the entire scene. Finally, we display the distribution of categories in video frames in Fig. 3. It is shown that most frames in [21,55] only contain 4 or 2 categories, whereas that result is 13 categories in our MVSeg. More details are provided in the supplementary materials.

4. MVSS Baseline Design

4.1. Technical Motivation

To date, various network architectures have been developed for the tasks of MSS and VSS. In the former task, many advanced feature fusion techniques have been designed to fuse features extracted from multispectral images based on two-stream encoders. The latter task focuses more on exploiting temporal associations in video sequence, such as optical flow warping [30] or space-time attention [51]. The use of either multispectral or temporal information has demonstrated their individual advantages in improving segmentation accuracy & robustness. However, there is no research touching the joint learning of both multispectral and temporal contexts which are both essential for MVSS.

Drawing ideas from recent MSS/VSS models, a straightforward solution for a MVSS model is to, first extract features from different spectra data using two-stream encoders, then build an external memory to hold the rich temporal & multispectral features, and finally extend the conventional space-time attention to an advanced spectrum-space-time version, where pixels of query features attend to all pixels of memory features, including these of RGB and thermal modalities as well as these of past video frames. In this way, we can definitely exploit the rich source of multispectral video features, and learn a joint relationship from multispectral and temporal contexts for semantic segmentation.

However, there are two certain challenges associated with this straightforward solution. The first challenge is how to keep the computational and memory costs moderate when processing large amounts of multispectral video data. As revealed, conventional attention block that performs all-to-all matching of feature maps is memory-consuming and computationally expensive [51]; it is unsuitable and unaffordable for MVSS, as multispectral video streams usually come in sequentially and need to be processed on time. This requires us to devise more elegant strategies for efficient MVSS. The second challenge comes from the inherent modality differences between RGB and thermal modes. Due to imaging differences, RGB data usually provide rich visible appearance information, while thermal data present more invisible thermal radiations of objects. Such modal-
differences will cause the feature embeddings of RGB and thermal frames to be distributed in different embedding spaces, leading to suboptimal cross-spectral feature attending and affecting the full exploitation of cross-spectral complementarity. Therefore, we should properly address the modality difference issue to make better use of multispectral complementary information. In Sec. 4.2, we introduce a well-designed MVSS baseline, called MVNet, which addresses the two challenges for MVSS.

4.2. Proposed Method

Fig. 4 presents an overview of the proposed MVNet. Starting from the input multispectral video, its pipeline consists of four parts: (a) feature extraction; (b) an MVFuse module to furnish the query features with the rich semantic cues of memory frames; (c) an MVRegulator Loss to regularize the multispectral video embedding space; and (d) a Cascaded Decoder to generate the final segmentation mask.

Feature Extraction: The multispectral video input contains a Query pair of RGB and thermal images at current frame $t$, and $L$ Memory pairs at past frames. They are denoted as $\{I^m_d\}_{d \in U, m \in \{R, T\}}$, where $d$ represents the time subscript of a certain frame in the set of $U = \{t-L, \ldots, t-1, t\}$, and $m$ denotes the modality type in $\{R, T\}$. These image pairs are fed into two-stream encoders to extract RGB and thermal features, respectively. To enrich the features, we fuse the outputs of different spectra by concatenation and $1 \times 1$ convolution, resulting in a series of fused features. These RGB, thermal, and fused features, together constitute a rich source of multispectral temporal cues for MVSS. We represent these features as $\{f^m_d \in \mathbb{R}^{H \times W \times D}\}_{d \in U, m \in \mathcal{M}}$, where $H \times W$ represents the spatial size, $D$ is the channel dimension, and $\mathcal{M} = \{R, T, F\}$.

MVFuse: An MVFuse module is then developed in Fig. 4b to furnish the Query features by engaging the rich yet cumbersome features of Memory frames. This is realized by two key designs: a memory-efficient prototypical memory and a computationally-efficient memory read block.

To preserve as many representative “pixels” as possible with minimal memory consumption, we build a prototypical memory that stores only a small number of the most representative categorical features of memory frames. Specifically, for each memory feature $f^m_d$, we derive $|C|$ class-level prototypical features, by average pooling all the embeddings of pixels belonging to each category $c \in C$. The estimated semantic masks are employed here to provide the required pixel category information of memory frames. Therefore, the memory features are summarized into a condensed set of prototypical features. We group the prototypical features of each modality as $\{p^m \in \mathbb{R}^{C \times D}\}_{m \in \mathcal{M}}$.

Afterwards, we devise an efficient Memory Read block, which enables a fast and efficient access of relevant semantic cues from prototypical memory to refine query features. This is achieved via an all-to-prototype attention. Taken the query feature $f^R_t$ as an example, we match it against all keys in prototypical memory. As shown in Fig. 4b, the inner product between the reshaped $f^R_t$ and $p^m$ are calculated as correlation maps, and transformed to weighting maps using a Softmax layer, expressed as: 
Here we process the attending of each modality separately, due to their different characteristics. The learned weighting maps are then used to selectively retrieve relevant information from memory, and update the query feature by:

\[ F_i^t = \Phi(\text{Concat}(\{ w^m \in \mathcal{M} \cup \{ t_i^t \}),) \text{m} \in \mathcal{M} \]\n
where Concat[.] denotes feature concatenation along channel dimension, and \( \Phi(\cdot) \) is a \( 1 \times 1 \) convolution operation to reduce the channel number to the original feature size.

Our MVFuse module finally outputs three informative features \( F_i^R, F_i^T, \) and \( F_i^F \) that have equipped with rich temporal and multispectral contexts, by modeling both cross-spectral and cross-frame relationships. In practice, we find that this strategy is not only more efficient (reducing the complexity from \( \mathcal{O}(L(HW)^2) \) to \( \mathcal{O}(L(HW) \times |\mathcal{C}|) \), where \( |\mathcal{C}| \ll HW \), but also more effective (increasing mIoU by 0.3%) than conventional attention that densely models pixel-to-pixel relationships. This may be partly due to the way of dense pixel matching may introduce some unnecessary or wrong correlations between regions with similar semantic but different classes, whereas our prototypical memory can degrade the side effects of ambiguous pixels and preserve the most typical representations.

**MVRegulator:** Inspired by the contrastive loss in unsupervised representation learning [9, 22], we further design a tailored MVRegulator loss for MVSS. Intuitively, features from different spectra or video frames but with the same object class should be closer to each other than any other features with different object classes in the same video.

Specifically, for a query pixel \( f_i^m(i, j) \) at position \( (i, j) \) of modality \( m \) with its groundtruth semantic label \( \bar{c} \), the positive set \( P \) includes prototypical features also belonging to the class \( \bar{c} \), and its negative set \( \mathcal{N} \) consists of prototypical features belonging to the other classes \( \mathcal{C} \setminus \bar{c} \). We include prototypical features of Query frame into the contrastive sets to consider within-frame contrasts. Formally, the MVRegulator loss is defined as:

\[
L_{reg}^m(i, j) = \frac{1}{|P|} \sum_{p \in P} - \log \frac{\exp(f \cdot p^m(\tau))}{\exp(f \cdot p^m(\tau)) + \sum_{p \in \mathcal{N}} \exp(f \cdot p^m(\tau))},
\]

\[
L_{reg} = \frac{1}{|\mathcal{M}| \times HW} \sum_{m \in \mathcal{M}} \sum_{(i, j) \in [HW]} L_{reg}^m(i, j).
\]

Here \( L_{reg}^m(i, j) \) is the partial loss for query pixel \( f_i^m(i, j) \) (simplified as \( f \) in Eq. 3), \( \tau \) denotes the temperature parameter, and all the embeddings are \( l_2 \)-normalized.

With \( L_{reg} \), the model is able to not only reduce modality differences between different spectra, but also promote intra-class compactness & inter-class separability. We would note that the MVRegulator loss is performed only during training, so it does not affect the inference time.

**Cascaded Decoder:** The final stage of the MVNet involves a cascaded decoder to predict segmentation mask based on \( F_i^R, F_i^T, \) and \( F_i^F \). Instead of direct prediction, we propose to cascadely integrate these features, and impose multiple supervisions on each level and the final result. This strategy is able to further promote multi-modal feature interaction and help filter unnecessary information redundancy. The segmentation loss in the decoder is then computed by the sum of these supervisions as:

\[
L_{seg} = L_wCE + \sum_{m \in \mathcal{M}} L_{wCE}^m,
\]

where we adopt the weighted cross-entropy loss \( L_{wCE} \) suggested by [50, 76, 82] for training. The overall training objective of the MVNet is thus defined as:

\[
L_{total} = L_{seg} + \lambda L_{reg},
\]

where \( \lambda \) is a weighting parameter for balancing the losses.

### 4.3. Implementation Details

Our code is implemented on the Pytorch platform and trained using two Nvidia RTX A6000 GPUs.

**Training:** We adopt DeepLabv3+ [8] as the encoders unless otherwise specified. For the thermal stream, we generate 3-channel thermal images by repeating the 1-channel thermal images. Each image is uniformly resized to \( 320 \times 480 \), and we perform random horizontal flipping and cropping to avoid potential over-fitting. We use Adam optimizer with an initial learning rate of 2e-4, which is adaptively scheduled based on training loss [51]. We set the batch size to 2. Following [9], \( \tau \) is set to 0.1. \( \lambda \) is set to 0.001 empirically. We select 3 memory frames (i.e., \( L = 3 \)). More details can be found in our source code and supplementary materials.

**Testing:** During testing, the system moves forward frame-by-frame, and the computed features at previous steps are added to the memory for the next frame. Therefore, the access of past video frames for inferring current frame only incurs a lightweight overhead. We report the mean Intersection over Union (mIoU) for evaluation.

### 5. Experiments

In this section, we conduct experiments to benchmark the new MVSS task and evaluate the proposed MVNet.

#### 5.1. Benchmark Results

We first benchmark MVSS by performing comprehensive experiments on various segmentation methods, including *image-based SS models* (CCNet [27], OCRNet [74], FCN [45], PSPNet [77], and DeepLabv3+ [8]), *MSS models* (MFNet [21], RTFNet [61], and EGFNet [82]), *VSS models* (STM [49] and LMANet [51]), and our proposed MVSS model - MVNet, using the MVSeg dataset.

Table 3 presents the segmentation results on the test set of MVSeg. Since there is no prior work directly applicable to the new MVSS task, we first present the closely-related
We first investigate the benefits of multispectral information in Table 5(a)&(b). As shown, the model trained with RGB images alone achieves an mIoU score of 51.59%; adding the thermal infrared (TIR) branch brings a substantial performance gain of 0.94% even using a simple direct fusion strategy (i.e., direct fusion).
concatenation). This reveals the benefits of leveraging multispectral information to improve semantic segmentation. **Cascade Decoder.** We then validate the efficacy of our cascaded decoder by using it to replace the direct fusion strategy. As shown in Table 5(c), the cascaded decoder leads to an mIoU gain of 0.34%, thanks to the advantages of our cascaded decoder to better filter & fuse complementary information from RGB and thermal modes. **MVFuse Module.** We deeply investigate the design of our MVFuse module in Table 5(d)-(f). Based on “model (c)”, we examine three MVFuse variants, i.e., MVFuse$t_{stm}$, MVFuse$t_{lma}$, and our proposed MVFuse$prot_{o}$, which differ in the design of memory and attention, while remaining all other settings the same. Technically, MVFuse$t_{stm}$ performs an all-to-all matching attention between query and memory frames with a large pixel-wise memory; MVFuse$t_{lma}$ reads only the spatial neighborhood regions of each position in query frame from pixel-wise memory. The results suggest that, i) leveraging multispectral video data is indeed useful, since all MVFuse variants yield increased mIoU scores compared to the single-frame baseline (c), ranging from 0.87% to 1.16%; and ii) our MVFuse$prot_{o}$ module is more favored, since it performs better, has smaller model size, faster inference, and requires less GPU memory, compared to MVFuse$t_{stm}$ and MVFuse$t_{lma}$. We attribute this to the superior performance of memory-efficient prototypical memory to preserve as many representative “pixels” as possible in the video, and our computationally-efficient memory read block to engage the rich multispectral temporal knowledge. **MVRegulator Loss.** We evaluate the MVRegulator loss in Table 5(g)&(h). As shown, integrating our MVRegulator loss improves mIoU score by 0.49% (i.e., 54.03%→54.52%), without introducing any extra model parameters or affecting inference time, which demonstrates its effectiveness to generate a more structured feature embedding space. We also derive an MVRegulator$un_{v}$ variant, which removes the cross-spectral contrast in Eq. 3. As seen, the mIoU score degrades, further showcasing the necessity of addressing the modality differences issue in MVSS. **Memory Frames Selection.** This part examines the impact of memory size $M$ and sample rate $S$ for memory frame selection. As shown in Table 6, adding memory frames consistently improves mIoU scores compared to the single-frame baseline (i.e., $M = 0$). When using more memory frames (i.e., $M = 3$), we see a clear performance increase (i.e., 52.87%→54.52%). Raising $M$ further beyond 3 gives marginal returns in performance. As a result, we set $M = 3$ for a better trade-off between accuracy and memory cost. Then we fix memory size $M = 3$, and experiment with different sample rate $S$. As shown in Table 7, best result is achieved when using a moderate sample rate $S = 3$. We set $M$ and $S$ to 3 in MVNet, which can efficiently make use of past video frames without holding on too old information.

### 5.3. Discussion and Outlook

Here we discuss three challenges and potential directions for future research. i) **Accuracy:** Motivated from the well-studied semantic segmentation of RGB images, the accuracy of MVSS model can be further advanced by exploring, e.g., multi-scale learning and boundary-aware modeling. ii) **Efficiency:** Although the engagement of multispectral videos brings significant improvement, it introduces additional model parameters. More lightweight schemes can be explored to improve efficiency, such as exploiting knowledge distillation [24] to transfer thermal knowledge to RGB stream. iii) **Evaluation metrics.** Due to the complex nighttime scenes, the popular TC metric [43] that evaluates temporal consistency based on optical flow warping may not correctly reflect the performance of MVSS models. How to design suitable metrics for MVSS is still an open issue.

### 6. Conclusion

In this paper, we have presented a preliminary investigation on the new task of semantic segmentation of multispectral video inputs. Specifically, we have provided a new challenging and finely annotated MVSeg dataset, developed a simple but efficient baseline framework (i.e., MVNet), conducted comprehensive benchmark experiments, and highlighted several potential challenges and future directions. The above contributions provide an opportunity for the community to design new algorithms for robust MVSS. In the future, we plan to expand the MVSeg dataset and provide other forms of annotations, e.g., instance annotations.

**Acknowledgements.** This research was funded by the CFI-JELF, Mitacs, NSERC Discovery (RGPIN-2019-04575) grants, NSFC (62001464), Guangzhou Key Research and Development Project (202206080008). This work was partially done at ByteDance.
References


[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1, 2, 3, 4


[37] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. Pattern Recognition, 96:106977, 2019. 3


Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016. 2


