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]
[1]University of Alberta [2]ByteDance [3]Johns Hopkins University
{wji3, jingjin1, lcheng5}@ualberta.ca, zzhou82@jh.edu, ayuille1@jhu.edu
https://jiwei0921.github.io/Multispectral-Video-Semantic-Segmentation

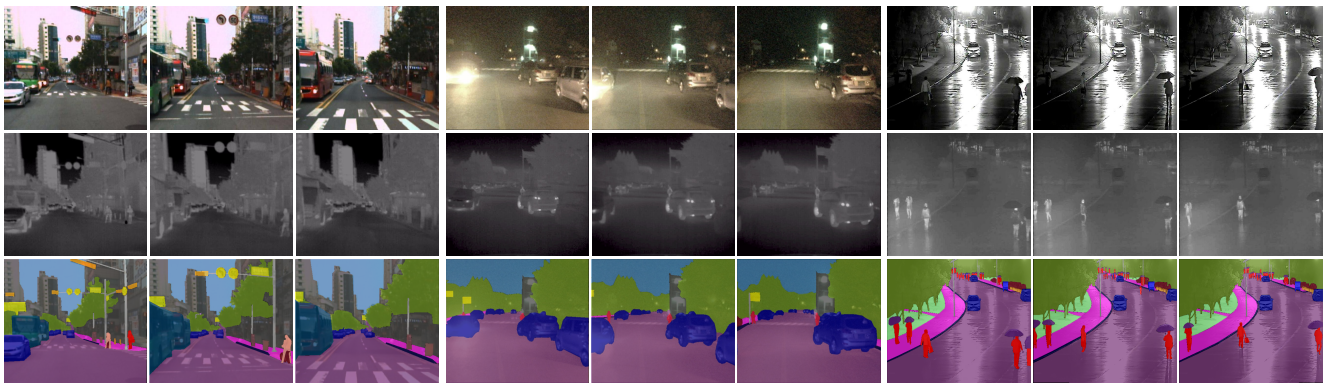

Figure 1. **Multispectral Video Semantic Segmentation**. Examples of three typical real-life video sequences under diverse conditions are given (*daytime* (left), *nighttime* and *overexposure* (middle), *rainy* and *low-light* (right)), where RGB images, thermal infrared images, and their pixel-level semantic annotations are shown through the first to the third rows, respectively.

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 RGBT 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

*Corresponding author.

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 RGBT 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-

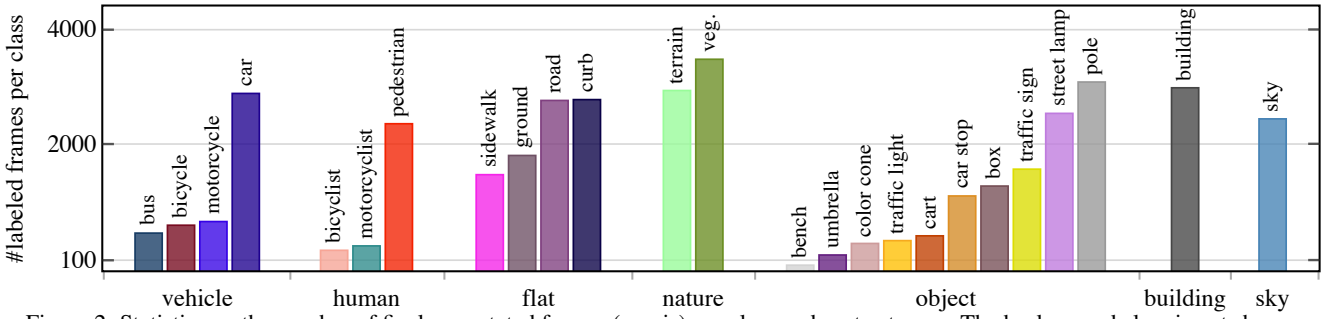

Figure 2. Statistics on the number of finely annotated frames (y-axis) per class and root category. The background class is not shown.

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 CamVid [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.

Table 1. High-level statistics of our MVSS dataset and existing MSS/VSS datasets. 'Seq.' means providing sequential video frames; 'TIR' means providing thermal infrared images. * Data annotations are obtained by human and models jointly.

Dataset	Seq.	TIR	#Videos(Frames)	#GTs	#Classes
Cityscapes [12]	✓		- (150k)	5,000	30
CamVid [4]	✓		5 (40k)	701	32
VSPW* [46]	✓		3,536 (252k)	252k	124
MFNet [21]		✓	-	1,569	9
PST900 [55]		✓	-	894	5
MVSeg	✓	✓	738 (53k)	3,545	26

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

Table 2. The pixel percentage per root category across existing multispectral (RGBT) semantic segmentation dataset, where '-' means no such classes.

Dataset	vehicle	human	flat	nature
MFNet [21]	5.05%	1.20%	0.59%	-
PST900 [55]	-	1.36%	-	-
Our Dataset	6.79%	0.91%	37.15%	33.22%
Dataset	object	building	sky	bkg.
MFNet [21]	1.02%	-	-	92.14%
PST900 [55]	1.66%	-	-	96.98%
Our Dataset	1.77%	11.76%	7.36%	1.04%

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.*

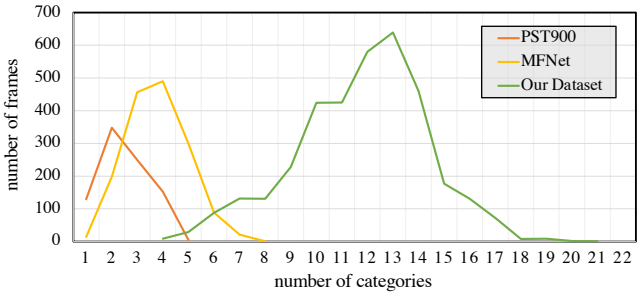

Figure 3. The number of categories per frame across existing semantic segmentation datasets with RGB and thermal pairs.

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. ❶ *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. ❷ *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-

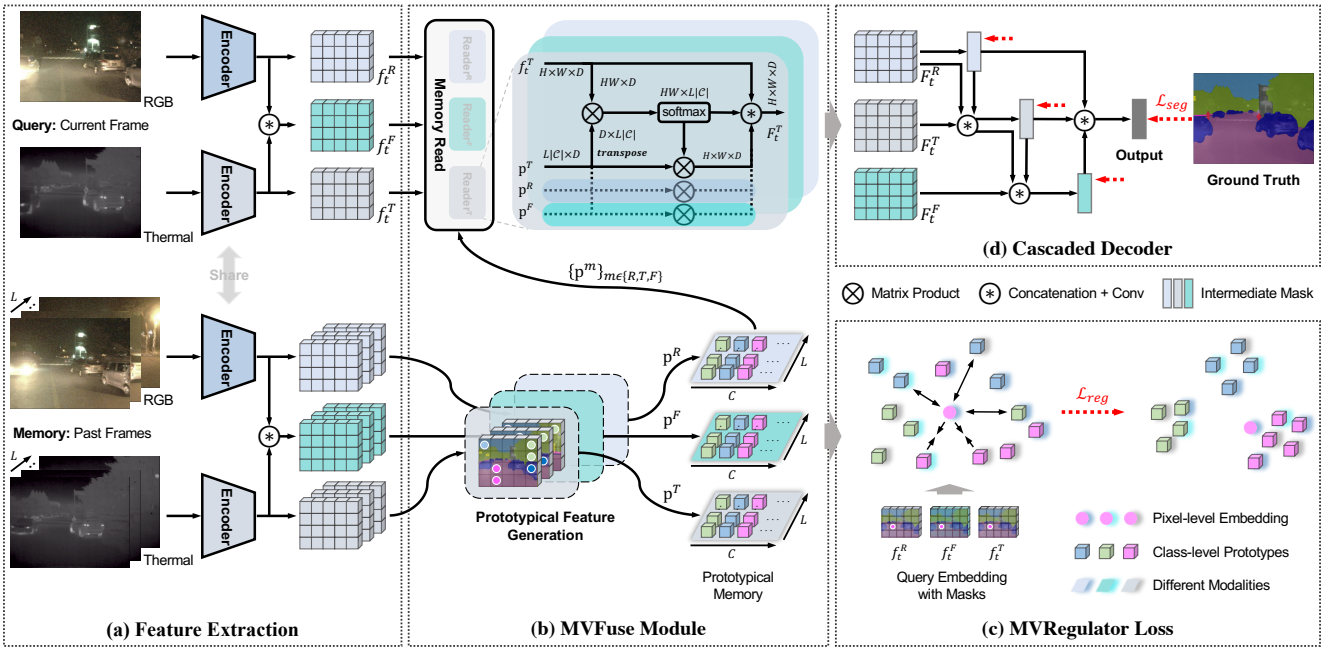

Figure 4. **Illustration of the proposed MVNet**. The input is a multispectral video clip, which contains one Query pair of RGB and thermal images, as well as L Memory pairs at past frames. The MVNet consists of four parts: (a) Feature Extraction to obtain the multispectral video features; (b) an MVFuse 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.

ity 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 address challenge ❶; (c) an MVRegulator loss to address challenge ❷; 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 $\{I_d^m\}_{d\in U, m\in\{R,T\}}$, 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 1×1 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 $\{\mathbf{f}_d^m \in \mathbb{R}^{H\times W\times D}\}_{d\in U, m\in\mathcal{M}}$, 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 $\mathbf{f}_d^m$, we derive $|\mathcal{C}|$ class-level prototypical features, by average pooling all the embeddings of pixels belonging to each category $c\in\mathcal{C}$. 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 $\{\mathbf{p}^m \in \mathbb{R}^{L|\mathcal{C}|\times D}\}_{m\in\mathcal{M}}$.

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 $\mathbf{f}_t^T$ as an example, we match it against all *keys* in prototypical memory. As shown in Fig. 4b, the inner product between the reshaped $\mathbf{f}_t^T$ and $\mathbf{p}^m$ are calculated as correlation maps, and transformed to weighting maps using a Softmax layer, expressed as:

$$\mathbf{w}^m = \text{Softmax}(\mathbf{f}_t^T \otimes \mathbf{p}^m), m \in \mathcal{M}. \quad (1)$$

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:

$$\mathbf{F}_t^T = \Phi(\text{Concat}[\{\mathbf{w}^m \mathbf{p}^m\}_{m\in\mathcal{M}} \cup \{\mathbf{f}_t^T\}]), \quad (2)$$

where Concat$[\cdot]$ denotes feature concatenation along channel dimension, and $\Phi(\cdot)$ is a 1×1 convolution operation to reduce the channel number to the original feature size.

Our MVFuse module finally outputs three informative features $\mathbf{F}_t^R$, $\mathbf{F}_t^T$, and $\mathbf{F}_t^F$ ($\mathbb{R}^{H\times W\times D}$) 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 $\mathbf{f}_t^m(i,j)$ at position (i,j) of modality m with its groundtruth semantic label $\bar{c}$, the positive set $\mathcal{P}$ 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 $\mathcal{C}/\bar{c}$. We include prototypical features of Query frame into the contrastive sets to consider within-frame contrasts. Formally, the MVRegulator loss is defined as:

$$\mathcal{L}_{reg}^m(i,j) = \frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}^+\in\mathcal{P}} -\log\frac{\exp(\mathbf{f}\cdot\mathbf{p}^+/\tau)}{\exp(\mathbf{f}\cdot\mathbf{p}^+/\tau)+\sum_{\mathbf{p}^-\in\mathcal{N}}\exp(\mathbf{f}\cdot\mathbf{p}^-/\tau)}, \quad (3)$$

$$\mathcal{L}_{reg} = \frac{1}{|\mathcal{M}|\times H\times W}\sum_{m\in\mathcal{M}}\sum_{(i,j)\in[H,W]}\mathcal{L}_{reg}^m(i,j). \quad (4)$$

Here $\mathcal{L}_{reg}^m(i,j)$ is the partial loss for query pixel $\mathbf{f}_t^m(i,j)$ (simplified as $\mathbf{f}$ in Eq. 3), τ denotes the temperature parameter, and all the embeddings are l_2-normalized.

With $\mathcal{L}_{reg}$, 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 $\mathbf{F}_t^R$, $\mathbf{F}_t^T$, and $\mathbf{F}_t^F$. 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:

$$\mathcal{L}_{seg} = \mathcal{L}_{wCE} + \sum\nolimits_{m\in\mathcal{M}}\mathcal{L}_{wCE}^m, \quad (5)$$

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

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda\mathcal{L}_{reg}, \quad (6)$$

where λ 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 320×480, 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], τ is set to 0.1. λ 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

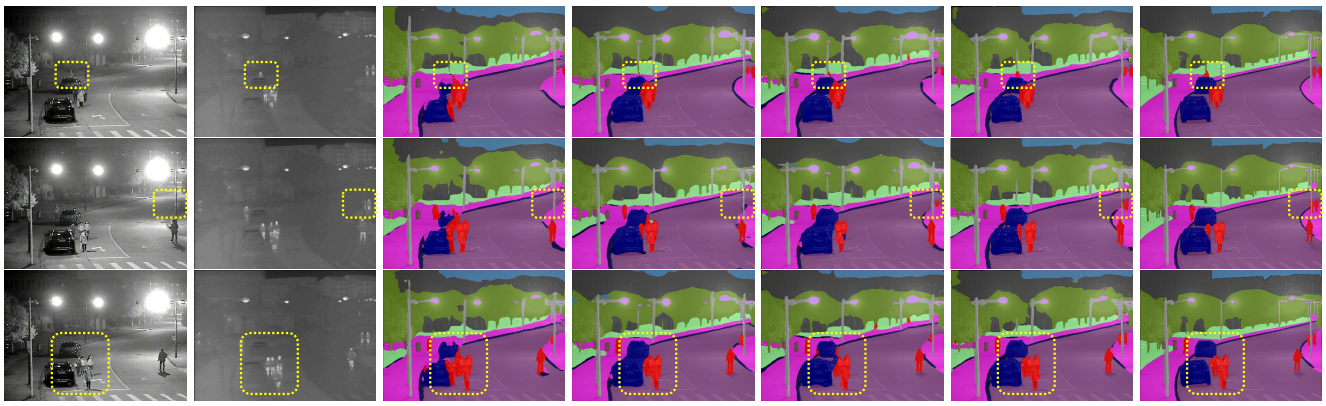

Figure 5. Qualitative results on the MVSeg dataset. We highlight the details with the yellow boxes. Best viewed in color and zoom in.

SS/MSS/VSS methods to provide a reference level. We reproduce these methods using their published codes with default setups. In our MVSS model, one important expectation compared to its image-level counterpart is whether the MVSS model improves per-frame segmentation accuracy by properly utilizing multispectral temporal features. To verify it, we apply our method to three popular image-based segmentation networks, including FCN [45], PSPNet [77], and DeepLabv3+ [8], to thoroughly validate the proposed algorithm. It is shown that our approach improves the performance of base networks by solid margins (*e.g.*, 51.38%→54.36% for PSPNet), suggesting that leveraging the multispectral temporal contexts is indeed beneficial for semantic segmentation, which has remained relatively untapped. Moreover, our MVNet shows a good generalization ability, which achieves consistently improved segmentation performance, independent of base networks.

To further evaluate the methods, we test them on daytime and nighttime scenarios, with results reported in Table 4. Again, our approach brings impressive gains over three strong baselines on both daytime and nighttime scenarios. For example, our MVNet$_{DeepLabv3+}$ yields mIoU scores of 57.80% and 40.48% on daytime and nighttime scenes, respectively, which shows promising gains of 2.63% and 2.35% over its counterpart DeepLabv3+. This further demonstrates the advantages of our MVNet to segment target objects under diverse lighting conditions.

Fig. 5 visualizes the segmentation results of a challenging nighttime scene with dim light. Compared with the competing methods, the results from our MVSS model (*i.e.*, MVNet$_{DeepLabv3+}$) are more accurate. *We provide more experimental results in the supplementary materials.*

5.2. Ablation Analysis

To investigate the effect of our core designs, we conduct ablation studies on the test set of MVSeg, with results presented in Table 5-7. Throughout the ablation experiments, we use DeepLabv3+ [8] as the backbone encoder.

Table 3. Quantitative evaluation on the test set of MVSeg dataset. The notation † and ‡ mean the VSS and MSS models, respectively.

Method	Backbone	mIoU(%)
CCNet [27]	ResNet-50	51.70
OCRNet [74]	ResNet-50	52.38
STM† [49]	ResNet-50	52.51
LMANet† [51]	ResNet-50	52.73
MFNet‡ [21]	Mini-inception	51.63
RTFNet‡ [61]	ResNet-152	52.77
EGFNet‡ [82]	ResNet-152	53.44
FCN [45]	ResNet-50	50.67
MVNet$_{FCN}$	ResNet-50	53.90 (+3.23)
PSPNet [77]	ResNet-50	51.38
MVNet$_{PSPNet}$	ResNet-50	54.36 (+2.98)
DeepLabv3+ [8]	ResNet-50	51.59
MVNet$_{DeepLabv3+}$	ResNet-50	**54.52** (+2.93)

Table 4. Quantitative results on daytime and nighttime scenarios of MVSeg dataset, respectively, evaluated using mIoU (%) metric.

Method	*Daytime*	*Nighttime*
CCNet [27]	54.59	38.38
OCRNet [74]	55.42	38.79
STM† [49]	55.22	38.19
LMANet† [51]	56.52	38.54
MFNet‡ [21]	54.63	39.14
RTFNet‡ [61]	56.62	39.26
EGFNet‡ [82]	56.89	40.10
FCN [45]	53.02	37.40
MVNet$_{FCN}$	57.19 (+4.17)	40.05 (+2.65)
PSPNet [77]	54.62	37.29
MVNet$_{PSPNet}$	57.73 (+3.11)	39.53 (+2.24)
DeepLabv3+ [8]	55.17	38.13
MVNet$_{DeepLabv3+}$	**57.80** (+2.63)	**40.48** (+2.35)

Multispectral Information. We first investigate the benefits of multispectral information in Table 5(a)&(b). As shown, the model trained with RGB images along 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

Table 5. Quantitative results of ablation study. 'TIR' means thermal infrared image. #Params refers to model parameters. #Mem means GPU memory usage during training. The inference time (ms) per frame is calculated under the same input scale.

*	Model Setups	#Param (M)	#Mem (G)	Times (ms)	mIoU (%)
(a)	RGB	41.6	4.6	8.1	51.59
(b)	RGB+TIR (direct fusion)	85.5	7.1	15.5	52.53
(c)	RGB+TIR (cascade fusion)	87.5	7.6	15.9	52.87
(d)	(c)+MVFuse$_{stm}$	96.1	45.7	32.6	53.74
(e)	(c)+MVFuse$_{lma}$	95.6	25.3	25.3	53.95
(f)	(c)+MVFuse$_{proto}$	88.4	18.7	18.4	54.03
(g)	(f)+MVRegulator$_{uni}$	88.4	18.8	18.4	54.26
(h)	(f)+MVRegulator (Ours)	88.4	18.8	18.4	54.52

concatenation). This reveals the benefits of leveraging multispectral information to improve semantic segmentation.

Cascaded 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$_{stm}$, MVFuse$_{lma}$, and our proposed MVFuse$_{proto}$, which differ in the design of memory and attention, while remaining all other settings the same. Technically, MVFuse$_{stm}$ performs an all-to-all matching attention between query and memory frames with a large pixel-wise memory; MVFuse$_{lma}$ 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$_{proto}$ module is more favored, since it performs better, has smaller model size, faster inference, and requires less GPU memory, compared to MVFuse$_{stm}$ and MVFuse$_{lma}$. We attribute this to the superiority of our 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$_{uni}$ 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.

Table 6. Ablation on the impact of memory size using mIoU(%).

Memory Size	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
Ours, $S = 3$	52.87	53.96	54.21	54.52	54.57	54.52

Table 7. Ablation on the impact of sample rate using mIoU(%).

Sample Rate	$S = 1$	$S = 2$	$S = 3$	$S = 4$	$S = 5$
Ours, $M = 3$	54.25	54.47	54.52	54.41	54.39

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

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[2] Qi Bi, Shaodi You, Wei Ji, and Theo Gevers. Learning rotation equivalent scene representation from instance-level semantics: A novel top-down perspective. *CVIU*, 229:103635, 2023. 1

[3] Qi Bi, Shuang Yu, Wei Ji, Cheng Bian, Lijun Gong, Hanruo Liu, Kai Ma, and Yefeng Zheng. Local-global dual perception based deep multiple instance learning for retinal disease classification. In *MICCAI*, pages 55–64, 2021. 2

[4] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 2, 3, 4

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 1, 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 6, 7

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 6

[10] Wenting Chen, Shuang Yu, Kai Ma, Wei Ji, Cheng Bian, Chunyan Chu, Linlin Shen, and Yefeng Zheng. Tw-gan: Topology and width aware gan for retinal artery/vein classification. *Medical Image Analysis*, 77:102340, 2022. 2

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 2

[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

[13] James W Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2-3):162–182, 2007. 3

[14] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *IROS*, pages 4467–4473, 2021. 2

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 3

[16] Fahimeh Fooladgar and Shohreh Kasaei. A survey on indoor rgb-d semantic segmentation: from hand-crafted features to deep convolutional neural networks. *Multimedia Tools and Applications*, 79:4499–4524, 2020. 2

[17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 2

[18] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *ICCV*, pages 4453–4462, 2017. 3

[19] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25(1):245–262, 2014. 2

[20] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014. 2

[21] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115, 2017. 2, 3, 4, 6, 7

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 6

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[24] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 8

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2

[26] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *CVPR*, pages 8818–8827, 2020. 3

[27] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–612, 2019. 6, 7

[28] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015. 3

[29] INO. Video analytics dataset. https://www.ino.ca/en/technologies/video-analytics-dataset/, 2012. 3

[30] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *CVPR*, pages 8866–8875, 2019. 3, 4

[31] Wei Ji, Wenting Chen, Shuang Yu, Kai Ma, Li Cheng, Linlin Shen, and Yefeng Zheng. Uncertainty quantification for medical image segmentation using dynamic label factor allocation among multiple raters. In *MICCAI on QUBIQ Workshop*, 2020. 2

[32] Wei Ji, Jingjing Li, Qi Bi, Chuan Guo, Jie Liu, and Li Cheng. Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. *ICLR*, 2022. 2

[33] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated rgb-d salient object detection. In *CVPR*, pages 9471–9481, 2021. 2

[34] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pages 52–69. Springer, 2020. 2

[35] Wei Ji, Ge Yan, Jingjing Li, Yongri Piao, Shunyu Yao, Miao Zhang, Li Cheng, and Huchuan Lu. Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:2321–2336, 2022. 2

[36] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, 2021. 1

[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

[38] Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, et al. Joint semantic mining for weakly supervised rgb-d salient object detection. *NeurIPS*, 34:11945–11959, 2021. 2

[39] Jingjing Li, Wei Ji, Miao Zhang, Yongri Piao, Huchuan Lu, and Li Cheng. Delving into calibrated depth for accurate rgb-d salient object detection. *International Journal of Computer Vision*, pages 1–22, 2022. 2

[40] Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Video semantic segmentation via sparse temporal transformer. In *ACMM*, pages 59–68, 2021. 3

[41] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *CVPR*, pages 19914–19924, 2022. 3

[42] Jianbo Liu, Junjun He, Jiawei Zhang, Jimmy S Ren, and Hongsheng Li. Efficientfcn: Holistically-guided decoding for semantic segmentation. In *ECCV*, pages 1–17. Springer, 2020. 2

[43] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *ECCV*, pages 352–368. Springer, 2020. 8

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 2, 6, 7

[46] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, pages 4133–4143, 2021. 3

[47] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *CVPR*, pages 6819–6828, 2018. 3

[48] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015. 2

[49] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 6, 7

[50] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 6

[51] Matthieu Paul, Martin Danelljan, Luc Van Gool, and Radu Timofte. Local memory attention for fast video semantic segmentation. In *IROS*, pages 1102–1109, 2021. 3, 4, 6, 7

[52] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 2

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[54] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *ICRA*, pages 13525–13531. IEEE, 2021. 2

[55] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *ICRA*, pages 9441–9447, 2020. 2, 3, 4

[56] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV*, 7576:746–760, 2012. 2

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 2

[58] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021. 2

[59] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *CVPR*, pages 3126–3137, 2022. 3

[60] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. In *ECCV*, 2022. 3

[61] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 2, 3, 6, 7

[62] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 2020. 2, 3

[63] Changshuo Wang, Chen Wang, Weijun Li, and Haining Wang. A brief survey on rgb-d semantic segmentation using deep learning. *Displays*, 70:102080, 2021. 2

[64] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020. 1

[65] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. 2

[66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[67] Zhengtuo Wang, Yuetong Xu, Jiongyan Yu, Guanhua Xu, Jianzhong Fu, and Tianyi Gu. Instance segmentation of point cloud captured by rgb-d sensor based on deep learning. *International Journal of Computer Integrated Manufacturing*, 34(9):950–963, 2021. 2

[68] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, volume 34, pages 12077–12090, 2021. 2

[69] Jiangtao Xu, Kaige Lu, and Han Wang. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, 146:179–184, 2021. 3

[70] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. Dynamic video segmentation network. In *CVPR*, pages 6556–6565, 2018. 3

[71] Zhengtian Xu, Shu Liu, Jianping Shi, and Cewu Lu. Outdoor rgbd instance segmentation with residual regretting learning. *IEEE Transactions on Image Processing*, 29:5301–5309, 2020. 2

[72] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, pages 332–348, 2020. 3

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

[74] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, pages 173–190, 2020. 1, 6, 7

[75] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018. 2

[76] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *CVPR*, pages 2633–2642, 2021. 2, 3, 6

[77] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 2, 6, 7

[78] Xiaoqi Zhao, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In *ACM MM*, pages 2645–2653, 2021. 3

[79] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Joint learning of salient object detection, depth estimation and contour extraction. *IEEE Transactions on Image Processing*, 31:7350–7362, 2022. 2

[80] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI*, pages 120–130. Springer, 2021. 1

[81] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pages 646–662. Springer, 2020. 2

[82] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb thermal scene parsing. In *AAAI*, 2022. 2, 3, 6, 7

[83] Wujie Zhou, Xinyang Lin, Jingsheng Lei, Lu Yu, and Jeng-Neng Hwang. Mffenet: Multiscale feature fusion and enhancement network for rgb-thermal urban road scene parsing. *IEEE Transactions on Multimedia*, 2021. 2

[84] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. 2, 3

[85] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, pages 593–602, 2019. 2