Supplementary Material: Unleashing Multispectral Video's Potential in Semantic Segmentation: A Semi-supervised Viewpoint and New UAV-View Benchmark

Wei Ji^{*1,2}**, Jingjing Li**^{*1}**, Wenbo Li**$^{\dagger 3}$**, Yilin Shen**^{3}**, Li Cheng**^{1}**, Hongxia Jin**^{3}

[1]University of Alberta, Edmonton, Canada
[2]Yale University, New Haven, USA
[3]Samsung AI Center, Mountain View, USA

In this supplementary material, we start by summarizing the notation definitions used in this paper in Sec. 1. Then, Sec. 2 gives the details of the CMF [1] applied in C3L and elaborates on the feature transformation process and corresponding dimension changes within DMR module. In Sec. 3, we illustrate various semi-supervised settings in related tasks and our Semi-MVSS. Sec. 4 presents more visualizations of the MVUAV dataset across various scenarios and discusses related aerial-view datasets and data alignment process. In Sec. 5, we provide more quantitative results, including benchmarking results on the new MVUAV dataset under the fully-supervised setting, segmentation results using different backbones, the exploration of SemiMV's extension capabilities, more detailed ablation analysis, using MVNet [2] as baseline, hyperparameter analysis, training scheme analysis, more analysis on $\mathcal{L}_{c3l}$ loss, and comparison with cross probability consistency. Finally, we discuss potential limitations with some feasible solutions for reference in Sec. 6.

1 Notation Table

In Table 1, a summary of the notation definitions is presented for better understanding.

2 Technical Details

2.1 Details of CMF in SemiMV

In the main paper, we have verified the necessity of cross-modal collaboration for the effective functioning of the SemiMV framework. It is implemented using the Cross-modal Fusion (CMF) module adapted from [1], which explicitly selects complementary information from two modalities and captures discriminative cross-modal features. Below is the details of the CMF moudle.

Specifically, $\mathbf{f}_i^R, \mathbf{f}_i^T \in \mathbb{R}^{D_i \times H_i \times W_i}$ represent the unimodal features extracted from the RGB and thermal encoders, where D_i, H_i, and W_i refer to the channel, height, and width number of the i-th layer, respectively. A gate function is employed to measure the usefulness of information propagated from each modality features, which is made up of a 3×3 convolutional layer followed by a Sigmoid activation function. Two normalized informative gate maps $G_i^R \in [0,1]^{H_i \times W_i}$ and $G_i^T \in [0,1]^{H_i \times W_i}$ can be obtained by:

$$\begin{cases} G_i^R = gate(\mathbf{f}_i^R) = \sigma(Conv(\mathbf{f}_i^R)), \\ G_i^T = gate(\mathbf{f}_i^T) = \sigma(Conv(\mathbf{f}_i^T)), \end{cases} \tag{1}$$

where σ is the Sigmoid activation function. Taking RGB as an example, the higher value of position (x, y) in this map indicates that RGB feature vector in $\mathbf{f}_i^R$ has useful information at the same position. Conversely, the lower value means that RGB feature has useless information or little information.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Summary of the notations and corresponding definitions used in this paper.

Notation	Definition
I_i^R, I_i^T	A pair of RGB and thermal frames at time step i.
t	The time subscript for query (current) frame.
M	The number of past frames used in memory.
$\{t-M, \ldots, t-1, t\}$	The set of time scripts of query (current) and memory (past) frames.
$\mathcal{V} = \{(I_i^R, I_i^T)\}_{i=1}^t$	A specific multispectral video clip unit.
$\mathcal{D}^L = \{(\mathcal{V}_n^L, y_n)\}_{n=1}^{n_L}$	The labeled set; n_L represents the number of labeled video clips; y_n denotes the ground-truth mask for the final frame of each clip, in a space of C classes.
$\mathcal{D}^U = \{\mathcal{V}_n^U\}_{n=1}^{n_U}$	The unlabeled set; n_U represents the number of unlabeled video clips.
$\mathcal{D}^V = \{(\mathcal{V}_n^V, y_n)\}_{n=1}^{n_V}$	The evaluation set; n_V represents the number of evaluation video clips.
P_i^R, P_i^T	The initial segmentation predictions.
Y_i^R, Y_i^T	The generated one-hot pseudo labels.
$\mathbf{f}_t^R, \mathbf{f}_t^T \in \mathbb{R}^{H\times W\times D}$	The initial query features, where $H\times W$ represents the spatial size, D is the channel dimension.
$\{\mathbf{f}_i^R, \mathbf{f}_i^T\}_{i\in[t-M,\cdots,t-1]}$	The initial memory features.
$\mathcal{R}_i \in \mathbb{R}^{H\times W\times 1}$	The pixel-wise reliability map in the DMR module.
$\mathbf{p}^*$	The prototypical memory feature bank on each modality, in which $\{\mathbf{p}^* \in \mathbb{R}^{MC\times D}\}_{*\in\{R,T\}}$.
$\mathbf{w}^*$	The intermediate attention matrix in the DMR module.
$\mathbf{F}_t^R, \mathbf{F}_t^T$	The memory-enhanced query features.
$\hat{P}_t^R, \hat{P}_t^T$	The updated predictions inferred from updated query features.
$\hat{Y}_t^R, \hat{Y}_t^T$	The updated one-hot pseudo labels.
P_t^{final}	The final segmentation prediction.

Naturally, two uninformative gate maps $\tilde{G}_i^R \in [0,1]^{H_i\times W_i}$ and $\tilde{G}_i^T \in [0,1]^{H_i\times W_i}$ can be generated by an inverse operation. The formulation is

$$\begin{cases} \tilde{G}_i^R = 1 - G_i^R, \\ \tilde{G}_i^T = 1 - G_i^T. \end{cases} \tag{2}$$

Two enhanced unimodal features $\tilde{\mathbf{f}}_i^R$ and $\tilde{\mathbf{f}}_i^T$ are obtained by:

$$\begin{cases} \tilde{\mathbf{f}}_i^R = G_i^R \otimes \mathbf{f}_i^R, \\ \tilde{\mathbf{f}}_i^T = G_i^T \otimes \mathbf{f}_i^T, \end{cases} \tag{3}$$

where $\otimes$ denotes element-wise multiplication. Through this operation, information redundancy can be avoided by enhancing the useful information of these features and effectively suppressing the useless ones. To exploit the cross-modal complementary relationship, the same operation is conducted between the uninformative gate maps ($\tilde{G}_i^R$ and $\tilde{G}_i^T$) and enhanced unimodal features from another modality ($\tilde{\mathbf{f}}_i^T$ and $\tilde{\mathbf{f}}_i^R$) to get complementary cross-modal features $\tilde{\mathbf{f}}_i^{R\_T}$ and $\tilde{\mathbf{f}}_i^{T\_R}$:

$$\begin{cases} \tilde{\mathbf{f}}_i^{R\_T} = \tilde{G}_i^R \otimes \tilde{\mathbf{f}}_i^T, \\ \tilde{\mathbf{f}}_i^{T\_R} = \tilde{G}_i^T \otimes \tilde{\mathbf{f}}_i^R. \end{cases} \tag{4}$$

Note that the useful information from thermal branch is selected by uninformative gate map of RGB branch and propagated to those positions where there is little information in RGB features. With this operation, the RGB features are complemented by the thermal features. Besides, to preserve the original information of each modality, a residual connection is performed. Thereby, the RGB-dominated cross-modal intermediate feature $\tilde{\mathbf{f}}_i^{R\_compl}$ can be captured, which have the complementary information passing from thermal features. And in the thermal branch, the above two steps are adapted the same as in RGB branch, so the two cross-modal enhanced features can be obtained by:

$$\begin{cases} \tilde{\mathbf{f}}_i^{R\_compl} = \mathbf{f}_i^R + \tilde{\mathbf{f}}_i^R + \tilde{\mathbf{f}}_i^{R\_T}, \\ \tilde{\mathbf{f}}_i^{T\_compl} = \mathbf{f}_i^T + \tilde{\mathbf{f}}_i^T + \tilde{\mathbf{f}}_i^{T\_R}. \end{cases} \tag{5}$$

The two enhanced features $\tilde{\mathbf{f}}_i^{R\_compl}$ and $\tilde{\mathbf{f}}_i^{T\_compl}$, are each forwarded to subsequent layers of their respective encoder networks to further facilitate cross-modal feature extraction. Within our SemiMV framework, which incorporates the DeepLabv3+ architecture as segmentation network, the CMF module is inserted at the second and fifth encoding layers to enable an effective cross-modal collaboration between the RGB and thermal data streams.

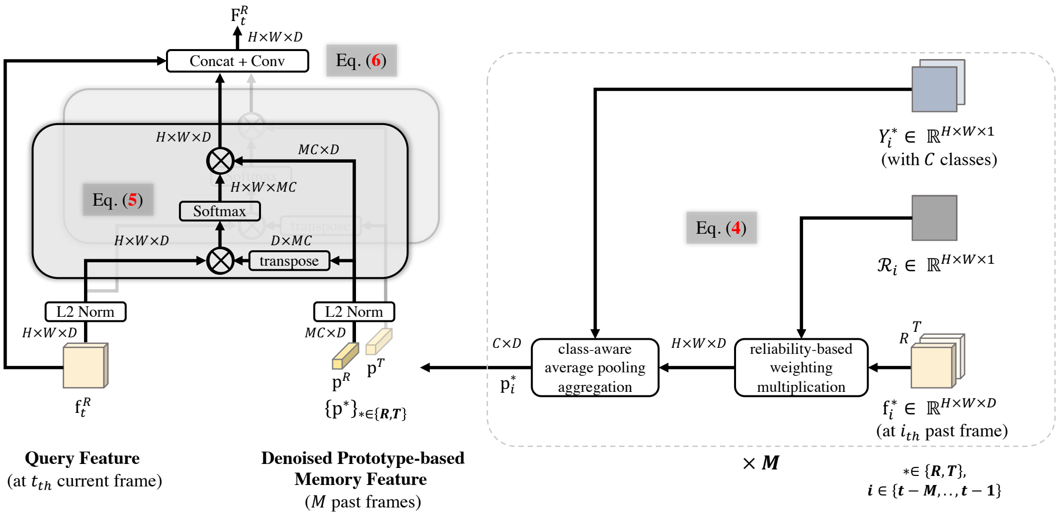


Figure 1: Detailed illustration of the internal transformation process within the DMR module.

2.2 Details of DMR in SemiMV

For better understanding, we illustrate the feature transformation process and corresponding dimension changes within the DMR module in Fig. 1, in relation to Eqs. 4, 5, and 6 of the main paper.

3 Task Definition

In Table 2, we illustrate the information used in the new semi-supervised MVSS task and related semantic segmentation tasks. As observed, the Semi-MVSS task utilizes a richer set of information, combining the benefits of both multispectral videos and semi-supervised learning. We also provide an intuitive visualization demo on our project website.

Table 2: Illustrations of information used in the semi-supervised MVSS (Semi-MVSS) task and related semantic segmentation tasks. * In our reimplemention of semi-supervised RSS models on the MVSS datasets, all unlabeled video frames are treated as *individual images* and engaged in the training process.

Task			Information		
	RGB	Thermal	Labeled	Unlabeled	Video
RSS	✓		✓		
Thermal-SS		✓	✓		
MSS	✓	✓	✓		
VSS	✓		✓		✓
MVSS	✓	✓	✓		✓
Semi-RSS	✓		✓	✓	✓*
Semi-VSS	✓		✓	✓	✓
Semi-MVSS	✓	✓	✓	✓	✓

4 The MUVAU Dataset

4.1 More Visualizations

Fig. 2 showcases a set of examples taken from the MVUAV dataset, offering an overarching view of the data it contains. Meanwhile, for better understanding, we also provide dynamic video visualizations on our project website.

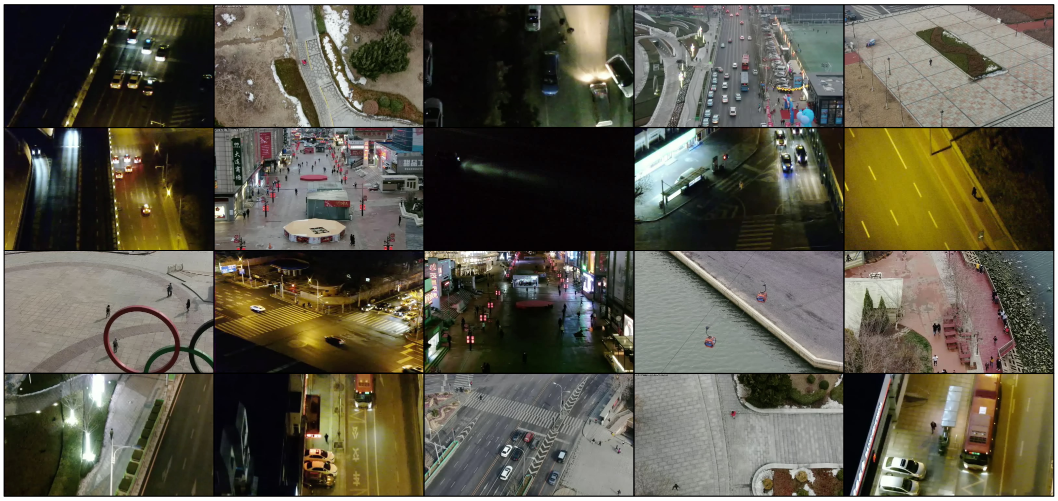


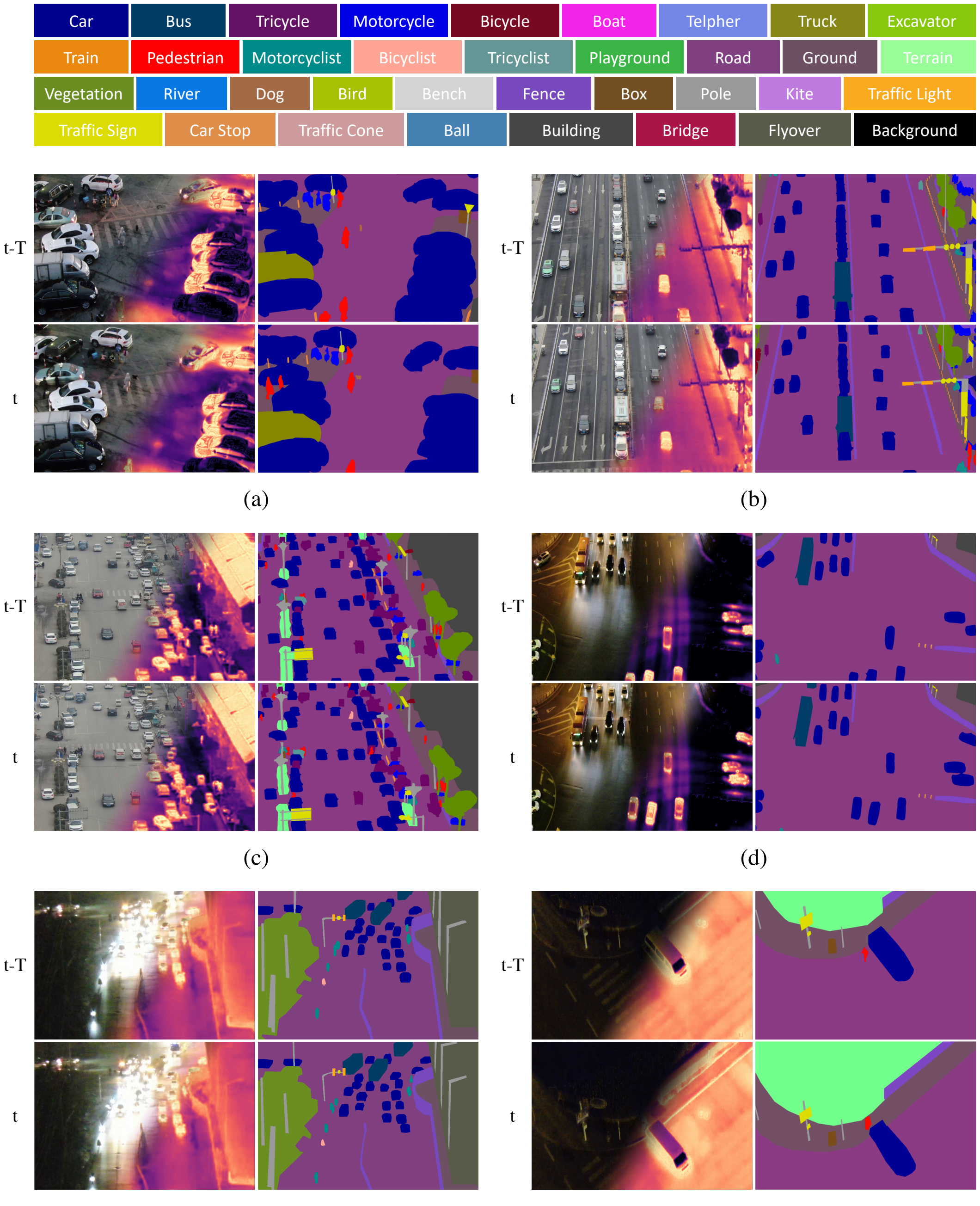
Figure 2: Some examples from our MVUAV dataset.

Fig. 3 displays several representative multispectral video sequences with dense semantic annotations from the MVUAV dataset under various conditions.

- (a) presents an urban street scene during twilight or early evening. The scene is alive with activity, featuring a parking area teeming with cars, trucks, and motorcycles, some parked, others in motion. Some individuals are observed engaging in various activities, either standing or walking through the area.
- (b) depicts a typical urban traffic scenario. A mix of vehicles and pedestrians coexist, navigating through the city's transport infrastructure. The vantage viewpoint allows for a clear observation of traffic patterns and road usage.
- (c) showcases a busy urban intersection, where traffic seems chaotic. Cars, trucks, and motorcycles are scattered across the intersection. Pedestrians are visible along the sidewalks, with some crossing the street or standing near the roadside. A row of buildings lines the right side of the image, and individuals are seen walking along the street. This image highlights the complexity and challenges of transportation in densely populated urban areas.
- (d) shows a city road at night with a modest flow of vehicles. It provides a contrast to the bustling daytime scenes.
- (e) portrays a challenging busy road at night, characterized by low visibility and the glare of strong headlight reflections. The difficulty in discerning specific details is notable. The inclusion of a thermal map offers significant assistance in identifying objects in such challenging conditions.
- (f) illustrates a street scene under extremely low illumination. A pedestrian is almost invisible in the RGB image. The complementary thermal map is crucial for enhancing scene understanding in such low-light conditions.

4.2 Dataset Analysis and Processing

Discussions with MVSeg and UAV-view Datasets. In this paper, we explore a new UAV-view perspective for multispectral video semantic segmentation, offering a distinct bird's-eye viewpoint that complements existing ground-level datasets like MVSeg [2], which advances the MVSS field. Meanwhile, thanks to the unique characteristics of UAVs, which provide a broader and more holistic view free from the constraints of ground-level capture, MVUAV encompasses extra challenging scenes such as rivers, boats, bridges, and playgrounds, as shown in Fig. 2. This characteristic is advantageous for applications that require comprehensive coverage in challenging conditions, such as aerial nighttime search and rescue, sea patrols, firefighting response support, traffic management, and UAV delivery services [3, 4, 5].



(a) (b) (c) (d) (e) (f)

Figure 3: Sample frames and annotations from six representative video sequences in our MVUAV dataset, showcasing a variety of conditions including daytime, low light, and darkness. From left to right: RGB/Thermal sequences with their pixel-level dense annotations.

Table 3: A comparison between existing aerial-view datasets and our proposed MVUAV dataset. 'Seg. Task' indicates whether the dataset supports segmentation tasks. '#GTs' and '#Cls' are the shorthand for the number of segmentation ground truths and semantic classes.

Dataset	**Color**	**Infrared**	**Video**	**Seg. Task**	**#Vids(Frames)**	**#GTs**	**Resolution**	**#Cls**
VisDrone2018 [6]	✓	✗	✓	✗	263 (179k)	-	3840×2160 (max)	-
UAVDT [7]	✓	✗	✓	✗	100 (80k)	-	1080×540	-
URUR [8]	✓	✗	✗	✓	-	3,008	5120×5120	8
LoveDA [9]	✓	✗	✗	✓	-	5,987	1024×1024	7
MVUAV (**Ours**)	✓	✓	✓	✓	413 (54k)	2,183	1920 × 1080	36

Table 4: Benchmarking MVUAV dataset under the fully-supervised setting. *RSS*: RGB-based semantic segmentation methods; *VSS*: video semantic segmentation methods; *MSS*: multispectral semantic segmentation methods; *MVSS*: multispectral video semantic segmentation methods.

Method	**Category**	**Backbone**	**mIoU**
CCNet [27]	RSS	ResNet50	33.29
OCRNet [28]	RSS	ResNet50	33.76
STM [29]	VSS	ResNet50	33.91
LMANet [30]	VSS	ResNet50	34.05
CFFM [31]	VSS	MiT-B1	34.13
MFNet [10]	MSS	Mini-Inception	33.17
RTFNet [32]	MSS	ResNet152	34.09
EGFNet [14]	MSS	ResNet152	34.47
EAEFNet [11]	MSS	ResNet152	34.65
MVNet [2]	MVSS	ResNet50	35.21
SemiMV*	MVSS	ResNet50	36.43

In Table 3, we additionally review related aerial-view datasets, highlighting the differences compared to our own. As shown, VisDrone2018 [6] and UAVDT [7] are two large-scale datasets designed primarily for object detection and tracking tasks in UAV-view RGB videos and/or images, providing bounding box annotations for target objects. In contrast, our MVUAV dataset is focused on the semantic segmentation task in UAV-view RGB-thermal videos, offering dense pixel-wise semantic annotations. URUR [8] and LoveDA [9] are two high-resolution segmentation datasets collected by high-quality satellite or Spaceborne images. The key advantage of our MVUAV dataset compared to URUR and LoveDA is the inclusion of complementary multispectral (RGB-thermal) videos. This feature aids in detecting target objects at nighttime or in adverse lighting conditions, thereby enhancing low-light vision capabilities.

Dataset Alignment. The well-aligned RGB-T pairs are crucial for multimodal segmentation tasks [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. To ensure the quality of our MVUAV, careful attention was given to alignment during the collection and preparation stages of both the sourced VTUAV dataset [26] and our MVUAV dataset. In the VTUAV, [26] manually identified corresponding feature points on both RGB and thermal images and calculated an affine transformation matrix from these points. Using this matrix, one image was warped to align with the other, and the common overlapping regions were extracted and resized to a consistent resolution while maintaining the aspect ratio. This ensures that most frames are well-aligned. Additionally, we performed a visualization screening process by overlaying thermal heat maps onto paired RGB images. This made it easier for our inspectors to verify alignment and allowed us to filter out low-quality samples (*e.g.*, similar content, blurred, or misaligned images), thus further enhancing the overall quality of the MVUAV.

5 More Experimental Results

Fully-supervised Setting. In the main paper, we have presented comprehensive benchmarking results of diverse semantic segmentation models using different input modalities, specifically in the semi-supervised setting. Here, we further benchmark the new MVUAV dataset in the fully-supervised setting. We reproduce existing RSS, VSS, MSS, and MVSS models on the MVUAV dataset using their official codes. The results are systematically summarized in Table 4. Here, SemiMV* is trained with all labeled multispectral videos. We anticipate that the benchmarking results on both semi-supervised setting and fully-supervised setting will facilitate the utilization of the new MVUAV dataset.

Different Backbones. Table 5 presents an evaluation of the proposed SemiMV network using different backbones. These include two CNN-based backbones (FCN [33] and DeepLabv3+ [34]) and one transformer-based backbone (SegFormer [35]). It can be observed that our SemiMV networks significantly enhance the performance over the SupOnly (RGBT) baselines, *e.g.,* from 36.88% to 43.04% with DeepLabv3+. Notably, our SemiMV employing SegFormer as the segmentation network achieves an impressive mIoU score of 43.51%.

Table 5: Evaluation of SemiMV with different backbones on the MVSeg dataset under the 1/4 training partition setup. FCN and DeepLabv3+ adopt ResNet50 as feature extractor, and SegFormer uses MiT-B2 as feature extractor.

*	FCN [33]		DeepLabv3+ [34]		SegFormer [35]	
	SupOnly	Ours	SupOnly	Ours	SupOnly	Ours
mIoU	36.27	42.66	36.88	43.04	38.02	43.51

Table 6: Extensions of SemiMV framework by combining mean-teacher strategy [36] or data augmentation [37]. Experiments are carried out on the MVSeg dataset under the 1/4 training partition setup, with DeepLabv3+ as the backbone.

Extended models	**mIoU**
SemiMV	43.04
SemiMV + Mean-Teacher [36]	43.59
SemiMV + Augmentation (CutMix [37])	43.41

Table 7: Ablation analysis for the design of semi-supervised MVSS baseline. The architectures can be referred to Fig. 4.

Index	Setting	mIoU
(a)	RGB supervised only	35.79
(b)	CPS (RGB as input)	39.27
(c)	CPS (RGBT as 4-channel input)	39.81
(d)	C3L *w/o* cross collaboration	36.67
(e)	C3L with direct collaboration	40.28
(f)	C3L *w/o* cross supervision	39.12
(g)	C3L with both cross supervision and collaboration	40.73
(h)	C3L + DMR *w/o* denoised $\mathcal{R}_i$	41.85
(i)	C3L + DMR	42.39
(j)	SemiMV (*i.e.*, with Dual-C3L)	43.04

Extended Models. We also empirically study the extensions of our method by integrating it with the mean-teacher strategy [36] and data augmentation [37] in Table 6. In the mean-teacher extension, a SemiMV-teacher network is created, which is updated using an exponential moving average of the parameters from the SemiMV-student network. The output from the SemiMV-teacher is used to generate one-hot pseudo labels, which in turn supervise the final predictions of the SemiMV-student. In the strong data augmentation extension, we apply the powerful CutMix [37] technique to augment the training RGB-T video pairs. The results of these extensions are presented in Table 6. We can observe that consistent improvements in segmentation performance are achieved. This highlights the exceptional extension capabilities of our SemiMV framework.

Detailed Ablation Analysis. To facilitate a clear understanding of the various settings employed in our ablation study, we have illustrated these configurations in Fig. 4, with the corresponding results detailed in Table 7. First, the comparison among settings (a), (b), and (c) indicates that integrating semi-supervised learning and thermal infrared information substantially enhances performance. In (c), (e) and (g), we explore different strategies for RGB and thermal fusion. These include using RGB-Thermal as a 4-channel input (input fusion), direct collaboration (illustrated in the upper right corner of Fig. 4), and the CMF-based cross-modal collaboration within our C3L module. Notably, compared to the model without cross-modal collaboration (model (d)), the introduction of any cross-modal collaboration strategy consistently and significantly boosts segmentation performance, underscoring the crucial role of cross-modal collaboration in the effective functioning of the SemiMV

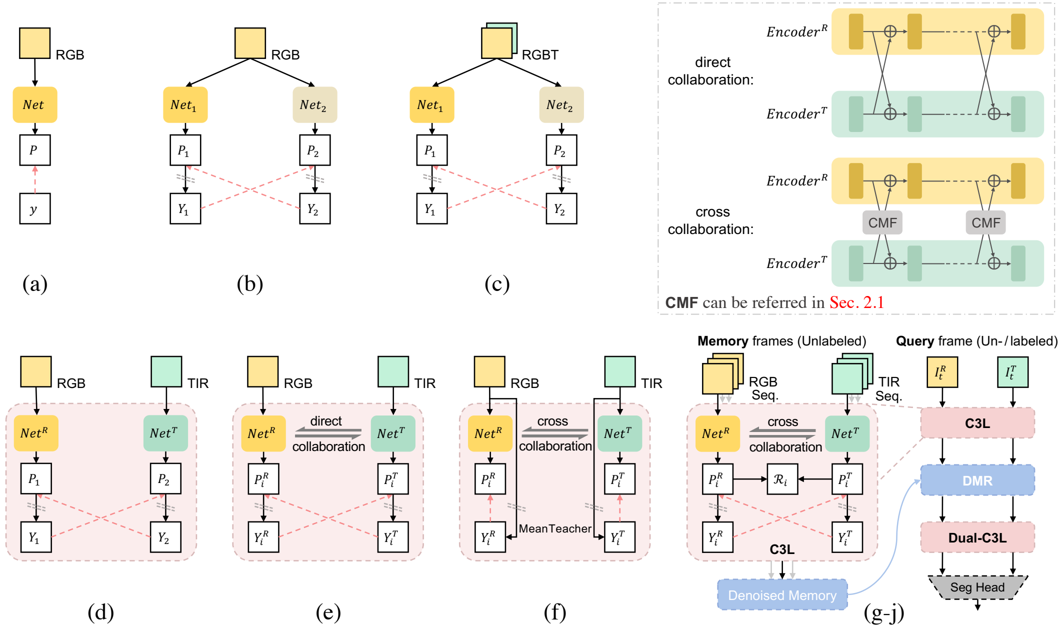


Figure 4: Diagrams of different ablation settings in Table 7: (a) RGB-based baseline using only supervised training. (b) RGB-based cross pseudo supervision (CPS [38]). (c) RGBT-based CPS using a 4-channel input. (d) C3L without cross collaboration. (e) C3L with direct collaboration. (f) C3L without cross supervision, employing individual mean-teacher pseudo supervision. (g-j) Our method incorporating C3L, DMR, and Dual-C3L. The top right legend illustrates direct and cross collaboration mechanisms.

framework. Among these, the CMF-based approach achieves the highest performance, benefiting from its robust cross-modal fusion capabilities. Additionally, the comparison between models (f) and (g) demonstrates that cross-supervision between two streams outperforms mean-teacher supervision of individual streams. This effectiveness is attributed to the C3L's capability to mutually correct potential errors, thereby enhancing overall accuracy. Lastly, we assess our denoising strategy, DMR and Dual-C3L in Table 7 (h)-(j). The results confirm that our proposed reliability estimation strategy effectively filters out unreliable features in the DMR fusion process. Furthermore, our Dual-C3L approach significantly enhances the combined strengths of our C3L and DMR strategies, effectively leveraging unlabeled multispectral videos for semi-supervised MVSS.

Table 8: Ablation analysis on using MVNet as backbone.

SemiMV using basic backbone	SemiMV using MVNet as backbone
43.04	44.10

Table 9: Ablation study on the impact of memory size (M).

Memory Size (M)	*w/o* temporal $M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
mIoU (%)	40.73	41.95	42.61	43.04	43.10	43.05

MVNet as Baseline. In the MVSS field, MVNet [2] serves as a strong baseline. Here, we explore using MVNet as our backbone in the Semi-MV framework. As shown in Table 8, our SemiMV using MVNet as backbone further improves the segmentation performance, benefiting from semi-supervised learning to effectively utilize unlabeled RGB-thermal videos and our denoising strategy to further improve MVRegulator loss in MVNet [2].

Hyperparameter Analysis. We further investigate the impact of M and λ (the number of past frames stored in memory, and the loss weight). As shown in Table 9 and Table 10, adding memory frames consistently improves mIoU scores, with a noticeable increase from 40.73% to 43.04% when $M = 3$. Raising M further beyond 3 gives marginal returns. Thus, we set $M = 3$ for a better trade-off

Table 10: Ablation study on the impact of the trade-off loss weight (λ).

Parameter (λ)	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$
mIoU (%)	41.89	42.66	43.04	42.95

Table 11: Ablation analysis of training strategy.

SemiMV with warm-up training (Ours)	SemiMV *w/o* warm-up training
43.04	41.53

Table 12: Robustness analysis of our SemiMV baseline. In this paper, we report the middle performance of the three experiment results as final results.

Training Case 1	Training Case 2	Training Case 3
43.04	43.03	43.07

Table 13: Ablation analysis of different loss combinations. The results are obtained on MVSeg dataset under the 1/4 data partition protocol. $\mathcal{L}_{Sup}$ represents the supervision losses involved in SemiMV for the labeled frames. $\mathcal{L}^{l}_{C3L}$, $\mathcal{L}^{u-iv}_{C3L}$ and $\mathcal{L}^{u-v}_{C3L}$ represent the SemiMV using our C3L based on cross pseudo supervision for labeled frames, unlabeled intra-video past frames within sparsely labeled videos and entirely unlabeled videos, respectively. The subscript '$pC3L$' means the C3L with the cross probability consistency loss on the labeled/unlabeled set. The overall performance on C3L with the cross pseudo supervision on both the labeled and unlabeled data is the best.

Analysis of losses							mIoU
$\mathcal{L}_{Sup}$	$\mathcal{L}^{l}_{C3L}$	$\mathcal{L}^{u-iv}_{C3L}$	$\mathcal{L}^{u-v}_{C3L}$	$\mathcal{L}^{l}_{pC3L}$	$\mathcal{L}^{u-iv}_{pC3L}$	$\mathcal{L}^{u-v}_{pC3L}$	
✓							40.17
✓	✓						40.49
✓	✓	✓					41.16
✓		✓	✓				42.84
✓	✓	✓	✓				**43.04**
✓				✓	✓	✓	41.78

between accuracy and memory cost. For λ, we found that $\lambda = 1$ balances supervised and pseudo losses effectively.

Training Analysis. In Table 11, we study the impact of the backbone warm-up training scheme (*i.e.*, supervised loss in Eq. 1). The results indicate a performance decrease of 1.51% when the backbone warm-up stage is omitted. This degradation occurs because the warm-up process enables the model to generate more meaningful supervision guidance for unlabeled video frames during the early stages of training, which in turn enhances the feature representation quality for training the SemiMV. Besides, we conduct three running experiments on the SemiMV, as presented in Table 12. These results further verify the reasonability and robustness of our SemiMV framework.

More Analysis on $\mathcal{L}_{C3L}$ Loss. We also conduct more analysis on the $\mathcal{L}_{C3L}$ loss in Table 13. We modify the videos/frames engaged in the training process of C3L, including using only labeled frames ($\mathcal{L}^{l}_{C3L}$), using sparsely labeled videos (labeled frames and their unlabeled past frames, $\mathcal{L}^{l}_{C3L}$ + $\mathcal{L}^{u-iv}_{C3L}$) and using all videos (sparsely labeled videos and extra unlabeled videos, $\mathcal{L}^{l}_{C3L}$ + $\mathcal{L}^{u-iv}_{C3L}$ + $\mathcal{L}^{u-v}_{C3L}$). As detailed in Table 13, the gradual incorporation of more unlabeled data consistently enhances performance, validating the effectiveness of our approach and demonstrating the potential of semi-supervised learning for the MVSS task. Our design of engaging both sparsely labeled videos and extra unlabeled videos achieves the best performance.

Comparison with Cross Probability Consistency. We compare our method with cross probability consistency (using L$_2$ loss, $\mathcal{L}_{pC3L}$) in the last two rows of Table 13. We can see that our original cross pseudo supervision design outperforms cross probability consistency by 1.26% mIoU. This empirical result is align with findings reported in CPS [38]. We conjecture this is due to that one-hot pseudo labels provide more confident supervision signals compared to the probabilistic predictions, which encourages the model to output confident predictions on unlabeled data.

Table 14: Analysis of the small-object problem in the MVUAV, under the 1/4 training partition setup.

SemiMV	SemiMV using multi-scale ensemble technique
26.52	27.64

6 Discussion and Outlook

Here we discuss three challenges and potential directions for future work. (1) Due to the high cost of labeling, the existing MVSS datasets are still relatively small in size. While the introduced semi-supervised MVSS task can mitigate the issue of label scarcity, its performance still falls short of the demands of real-world applications. Investigating photorealistic rendering technologies [39] could offer another promising avenue for progress. We could explore creating simulated data or augmenting existing annotations using these techniques, potentially lowering labeling costs substantially. (2) Meanwhile, our introduction of a SemiMV baseline opens new avenues in semi-supervised MVSS. However, this line of research is still in its initial stage, and the accuracy has large room for improvement. We may extend the SemiMV framework to incorporate ideas from the established field of semi-supervised learning, *e.g.*, mean-teacher strategy [36] and data augmentation/perturbation [37], to enhance model's performance. Besides, we can consider extra guidance signals (*e.g.*, contour [40, 41]) to aid in identifying unreliable areas in pseudo labels, further refining the process. (3) In addition, this work introduces a new UAV-view MVSS dataset, MVUAV, which can complement the eye-level viewpoint of existing MVSeg dataset and serve as an additional resource for a more thorough evaluation of MVSS models. However, this dataset presents specific challenges, particularly with smaller targets and scale variation in UAV views. To address these issues, we adapted a widely-used multi-scale ensemble technique [42] from aerial imagery analytics to our SemiMV framework, incorporating an additional resolution of 360×540 for ensemble operations. As shown in Table 14, addressing the unique challenges of the UAV view can significantly enhance overall performance. Furthermore, when higher resolution and finer analysis are required, off-the-shelf super-resolution tools can be employed to enhance image quality. We encourage further research to continue exploring these challenges.

References

[1] Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022.

[2] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *CVPR*, pages 1094–1104, 2023.

[3] Seunghyeon Lee, Youngkeun Song, and Sung-Ho Kil. Feasibility analyses of real-time detection of wildlife using uav-derived thermal and rgb images. *Remote Sensing*, 13(11):2169, 2021.

[4] Gaetano Messina and Giuseppe Modica. Applications of uav thermal imagery in precision agriculture: State of the art and future research outlook. *Remote Sensing*, 12(9):1491, 2020.

[5] Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Thomas Lagkas, and Ioannis Moscholios. A compilation of uav applications for precision agriculture. *Computer Networks*, 172:107148, 2020.

[6] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.

[7] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *ECCV*, pages 370–386, 2018.

[8] Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *CVPR*, pages 23621–23630, 2023.

[9] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *NeurIPS*, 2021.

[10] 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.

[11] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, 2023.

[12] Wei Ji. Deep learning-based segmentation for complex scene understanding. *University of Alberta*, 2024.

[13] 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.

[14] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb thermal scene parsing. In *AAAI*, pages 3571–3579, 2022.

[15] 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.

[16] Jingjing Li, Wei Ji, Size Wang, Wenbo Li, and Li Cheng. Dvsod: Rgb-d video salient object detection. In *NeurIPS*, pages 8774–8787, 2023.

[17] Wei Ji, Jingjing Li, Cheng Bian, Zhicheng Zhang, and Li Cheng. Semanticrt: A large-scale dataset and method for robust semantic segmentation in multispectral images. In *ACM MM*, pages 3307–3316, 2023.

[18] 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.

[19] 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*, 131(4):855–876, 2023.

[20] 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.

[21] 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.

[22] Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 25:5142–5154, 2022.

[23] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. Lfnet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020.

[24] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. *NeurIPS*, pages 898–908, 2019.

[25] 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.

[26] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *CVPR*, pages 8886–8895, 2022.

[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.

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

[29] 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.

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

[31] 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.

[32] 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.

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

[34] 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.

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

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, pages 1195–1204, 2017.

[37] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019.

[38] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021.

[39] Ansys Inc. Thermal analysis simulation software. `https://www.ansys.com/applications/thermal-analysis-simulation-software`.

[40] Shuo Li, Yue He, Weiming Zhang, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, and Jingdong Wang. Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. In *ICCV*, pages 16348–16358, 2023.

[41] 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.

[42] Adam Van Etten. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv preprint arXiv:1805.09512*, 2018.