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[3], Yilin Shen[3], Li Cheng[1], Hongxia Jin[3]

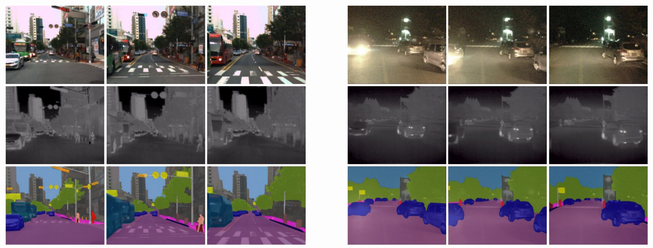
[1]University of Alberta [2]Yale University [3]Samsung AI Center-Mountain View

NeurIPS 2024
December 10-15, 2024, | Vancouver, Canada

Background

Multispectral Video Semantic Segmentation (MVSS)

Multispectral Video Semantic Segmentation (MVSS) focuses on RGBT video inputs for semantic segmentation tasks. With the rapid advancements in RGB and thermal imaging—known as multispectral imaging—MVSS has gained significant attention. Notably, it offers promising opportunities to enhance segmentation performance under challenging visual conditions, such as low light or overexposure. The new task opens possibilities for applications that require a holistic view of video segmentation under challenging conditions, e.g., autonomous safe driving, nighttime patrol, and fire rescue.

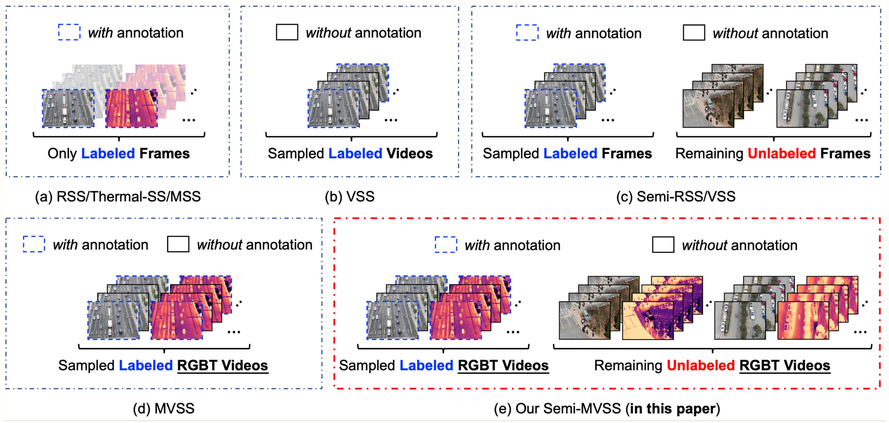


Daytime Nighttime and overexposure

New Task

Semi-supervised MVSS Task (Semi-MVSS)

This figure below illustrates the information used in the semi-supervised MVSS task and related semantic segmentation tasks.

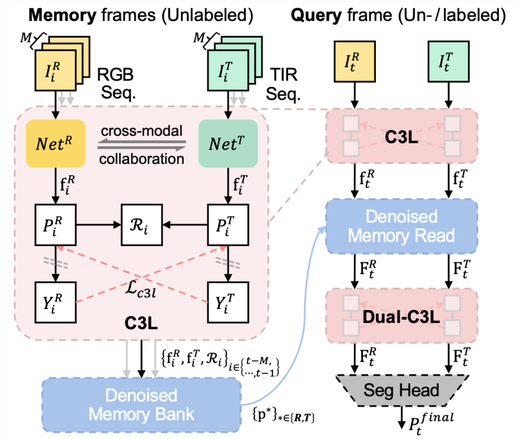


Features:

- A limited amount of sparsely labeled RGB-Thermal videos, with blue dotted box. (MVSS)
- Massive unlabeled ones with black box, are used to improve segmentation. (**Semi-MVSS**)

Proposed Method for Semi-MVSS

SemiMV - Cross-collaborative Consistency Learning

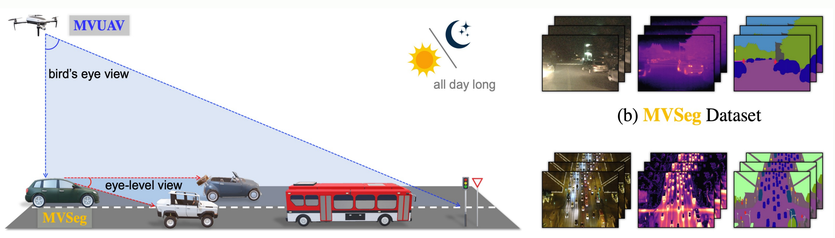


This figure illustrates the overview of proposed method. The C3L loss aims to learn from unlabeled RGB-Thermal pairs. The DMR is responsible for integrating temporal information from the denoised memory bank to update query features. A dual-C3L loss is further applied to regularize updated query features. Finally, a segmentation head predicts the final mask.

New MVUAV Dataset

Multispectral Video Semantic Segmentation in UAV Videos

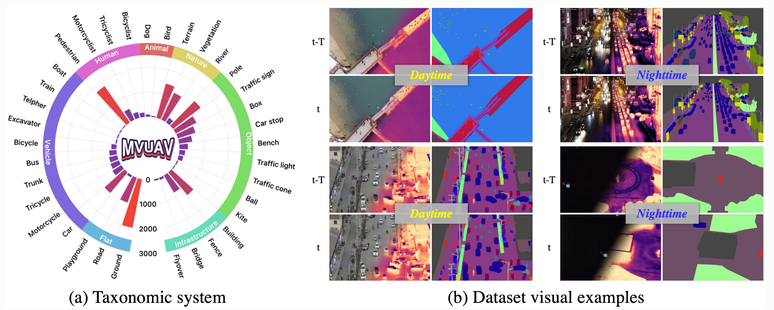
We introduce MVUAV, a new MVSS dataset containing a wide range of RGB-T videos captured by Unmanned Aerial Vehicles (UAVs) from an oblique bird's-eye viewpoint. This viewpoint offers a complementary perspective to the eye-level viewpoint adopted by existing MVSeg dataset.



(a) Data Acquisition (b) MVSeg Dataset (c) MVUAV Dataset

Features:

- (a) Viewpoint diversity of the existing MVSeg dataset and the new MVUAV dataset.
- (b) & (c) Representative samples from the MVSeg & MVUAV datasets, where RGB videos, thermal videos, and the corresponding semantic annotations are visualized.



(a) Taxonomic system (b) Dataset visual examples

Table 1: Statistics of various semantic segmentation datasets in diverse modalities. 'Surv.', '#Cls' and 'Anno.' are the shorthand for surveillance, the number of classes and annotation density, respectively.

Dataset	Year	Color	Infrared	Video	UAV	Capture	#Vids(Frames)	#GTs	Resolution	#Cls	%Anno.
Cityscapes [1]	2016	✓	✗	✓	✗	Car	- (150k)	5,000	2048×1024	30	97.10%
CamVid [82]	2009	✓	✗	✓	✗	Car	5 (40k)	701	960×720	32	96.20%
UAVid [83]	2020	✓	✗	✓	✓	Drone	42 (38k)	420	3840×2160	8	82.69%
SODA [84]	2020	✗	✓	✗	✗	Pedestrian	-	2,168	640×480	21	79.73%
SCUT-Seg [85]	2021	✗	✓	✗	✗	Car	-	2,010	720×576	10	56.50%
MFNet [46]	2017	✓	✓	✗	✗	Car	-	1,569	640×480	9	7.86%
PST900 [47]	2020	✓	✓	✗	✗	Robot	-	894	1280×720	5	3.02%
SemanticRT [45]	2023	✓	✓	✗	✗	Surv.	-	11,371	1280×1024	13	21.27%
FMB [86]	2023	✓	✓	✗	✗	Car	-	1,500	800×600	15	98.16%
CART [87]	2024	✓	✓	✗	✓	Drone	-	2,282	960×600	11	99.98%
MVSeg [21]	2023	✓	✓	✓	✗	Car/Surv.	738 (53k)	3,545	480×640	26	98.96%
MVUAV	-	✓	✓	✓	✓	Drone	413 (54k)	2,183	1920×1080	36	99.18%

Experiments

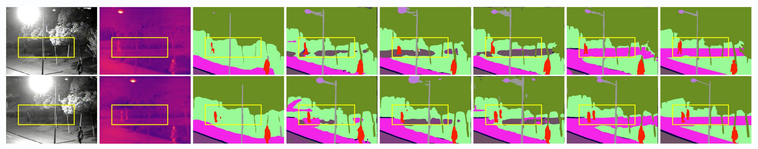


Table 2: Quantitative evaluation on the MVSeg dataset. SupOnly stands for the model trained on the labeled data.

Method	1/16 (26,140)	1/8 (54,282)	1/4 (111,561)	1/2 (228,1119)
SupOnly (RGB)	21.95	27.09	35.79	42.37
MT [22]	23.39	29.45	38.75	44.51
CCT [23]	23.81	29.66	39.04	44.89
CPS [25]	23.88	30.05	39.27	45.34
UniMatch [24]	24.73	30.47	39.39	45.42
Accel [77] (Video)	23.16	28.41	37.31	43.75
IFR [80]	24.79	30.97	40.69	46.21
SupOnly (RGBT)	23.26	28.45	36.88	43.75
MVNet [21]	24.70	30.32	39.89	46.08
SemiMV (Ours)	25.48	34.12	43.04	49.73

Table 3: Quantitative evaluation on the MVUAV dataset.

Method	1/16 (23,91)	1/8 (40,184)	1/4 (70,365)	1/2 (141,732)
SupOnly (RGB)	10.09	13.47	20.07	26.25
MT [22]	11.33	15.89	23.02	27.83
CCT [23]	11.75	16.11	23.72	28.71
CPS [25]	12.55	16.70	24.01	29.09
UniMatch [24]	13.36	17.21	24.10	29.21
Accel [77] (Video)	11.23	14.69	21.45	27.70
IFR [80]	13.11	17.03	24.91	29.87
SupOnly (RGBT)	11.28	14.88	21.31	27.60
MVNet [21]	13.07	16.86	23.36	29.77
SemiMV (Ours)	15.10	20.04	26.52	32.23

Table 4: Ablation study of the proposed SemiMV framework.

Methods	Information					mIoU
	RGB	Thermal	Labeled	Unlabeled	Video	
RGB	✓		✓			35.79
RGB-Thermal	✓	✓	✓			36.88
+C3L	✓	✓	✓	✓		40.73
+DMR	✓	✓	✓	✓	✓	42.39
+Dual-C3L	✓	✓	✓	✓	✓	43.04

Our Paper

WeChat