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]

Project website: https://jiwei0921.github.io/MVUAV

Abstract

Thanks to the rapid progress in RGB & thermal imaging, also known as multispectral imaging, the task of multispectral video semantic segmentation, or MVSS in short, has recently drawn significant attentions. Noticeably, it offers new opportunities in improving segmentation performance under unfavorable visual conditions such as poor light or overexposure. Unfortunately, there are currently very few datasets available, including for example MVSeg dataset that focuses purely toward eye-level view; and it features the sparse annotation nature due to the intensive demands of labeling process. To address these key challenges of the MVSS task, this paper presents two major contributions: the introduction of MVUAV, a new MVSS benchmark dataset, and the development of a dedicated semi-supervised MVSS baseline - SemiMV. Our MVUAV dataset is captured via Unmanned Aerial Vehicles (UAV), which offers a unique oblique bird's-eye view complementary to the existing MVSS datasets; it also encompasses a broad range of day/night lighting conditions and over 30 semantic categories. In the meantime, to better leverage the sparse annotations and extra unlabeled RGB-Thermal videos, a semi-supervised learning baseline, SemiMV, is proposed to enforce consistency regularization through a dedicated Cross-collaborative Consistency Learning (C3L) module and a denoised temporal aggregation strategy. Comprehensive empirical evaluations on both MVSeg and MVUAV benchmark datasets have showcased the efficacy of our SemiMV baseline.

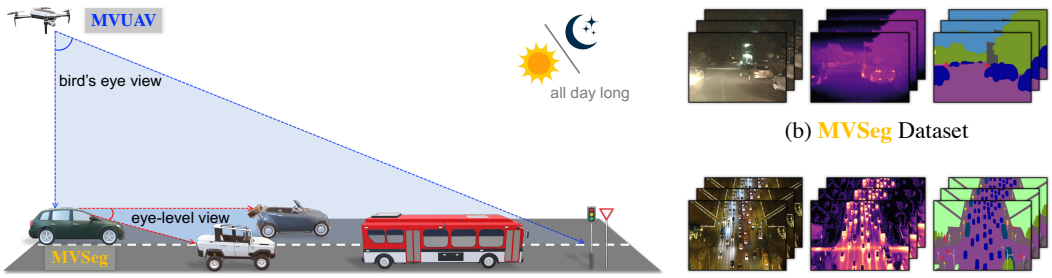
1 Introduction

Semantic segmentation is the process of categorizing each pixel in an image/video to a specific class label, which plays a vital role in visual scene understanding [1, 2]. Remarkable progresses has been made in the past several years, particularly in RGB-based semantic segmentation [3, 4, 5, 6, 7, 8, 9, 10, 11]. With the increasing accessibility of thermal sensors in capturing thermal radiation of objects with temperature above absolute zero, multispectral semantic segmentation (MSS) [12, 13, 14, 15], where both RGB and infrared thermal cameras are engaged, starts to gain notable traction. An exemplar illustration is shown in Fig. 1. This multispectral imaging setup excels specifically in scenes with unfavorable visual conditions, such as low-light or overexposure. On the other hand, the dynamic and ever-changing nature of real-world scenarios has propelled interests in video semantic segmentation (VSS) [16, 17, 18, 19, 20] with impressive segmentation performance. The task of multispectral (aka RGB-Thermal or RGB-T) video semantic segmentation (MVSS)[21] has emerged recently as a promising visual segmentation setup that has the best of both worlds.

Nevertheless, as being a relatively new task, there are still a number of hurdles in MVSS research. The most prominent one is the lack of quality datasets. One major MVSS benchmark presently

[1]University of Alberta [2]Yale University [3]Samsung AI Center-Mountain View (work done here). [†]Intern mentor.
[*]Equal contribution. Corresponding authors to: Wei Ji <wei.ji@yale.edu>, Jingjing Li <jingjin1@ualberta.ca>.

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



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

Figure 1: (a) Viewpoint diversity of the existing MVSeg dataset [21] 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.

available is MVSeg [21], a recently introduced dataset. It is worth mentioning that the multispectral videos in existing MVSS datasets, including MVSeg, are all taken from eye-level views, which clearly suggests the lack of data source diversity in MVSS. Even after securing good quality of RGB-T video footage, it is still a demanding exercise to make annotations. In practice, this is achieved by furnishing pixel-wise semantic labels on a selected subset of frames or key-frames[1]. Consequentially, there typically exists only *sparse* annotations in the MVSS datasets.

This paper aims to address the above-mentioned challenges. Specifically, we introduce MVUAV, a new MVSS dataset containing a diverse range of RGB-T videos captured by Unmanned Aerial Vehicles (UAVs) from an oblique bird's-eye viewpoint. As depicted in Fig. 1, this viewpoint provides a broader, holistic perspective free from the constraints of eye-level capture adopted by existing MVSS datasets such as MVSeg. In detail, our MVUAV dataset comprises 413 RGB-Thermal videos with 53,828 frame pairs in total. A subset of 2,183 key-frame pairs are meticulously annotated with pixel-wise semantic labels across 36 different classes of objects and stuffs, where the pixel annotation rate is 99.18%. Our dataset captures diverse real-world scenarios such as roads, streets, bridges, parks, seas, beaches, courts and schools; it also spans different lighting conditions from daytime to low-light and even pitch-dark scenarios. The MVUAV also presents some challenges, such as large scale variation, fine-grained scene parsing, moving object segmentation, and adverse illumination conditions, making it a valuable asset for evaluating various MVSS algorithms.

Meanwhile, we seek to tackle the MVSS task from a relatively new semi-supervised perspective. As illustrated in Fig. 2, our strategy utilizes a small number of sparsely labeled RGB-Thermal videos, alongside massive amount of unlabeled videos. In the closely-related domain of semi-supervised RGB-based semantic image segmentation, the idea of consistency regularization has played important role in the eventual performance. Its efficacy stems from perturbation-invariant training, by enforcing consistent predictions despite of varied perturbations presented in unlabeled RGB images at different processing levels: input [22], feature [23, 24], or network [25]. There motivates us to explore consistency regularization in RGB-Thermal videos. In fact, MVSS seems to be an ideal setting, since RGB and thermal images essentially capture the same scenes from different sensory perspectives, that offers innate input perturbations. Further, processing the multimodal data through two parallel networks with distinct parameters also leads to valuable feature-level and network-level perturbations. Building on these insights, we introduce SemiMV, a novel semi-supervised MVSS framework. At its core is a Cross-collaborative Consistency Learning (C3L) module, where the RGB and thermal streams mutually offer pseudo-supervisions to each other. A pixel-wise reliability map is also generated, based on the learned cross-modal consistency, to guide the temporal fusion process and mitigate potential noise.

The main contributions of this paper are summarized as follows:

- A new benchmark dataset, MVUAV, is introduced. The RGB-T videos in the new dataset present complementary perspectives to the the existing MVSS datasets such as MVSeg that are typically from eye-level views, by capturing from an oblique bird's-eye viewpoint. It also provides pixel-level dense annotations with a rich set of 36 visual semantic categories.

[1]In video-based segmentation datasets (*e.g.,* Cityscapes [1] and MVSeg [21]), it is a common practice to *sparsely* annotate video frames, given similar content of consecutive frames and substantial cost saving in the amount of annotation effort.

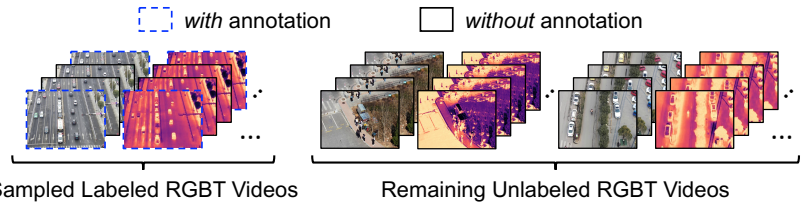


Figure 2: Illustration of training examples in semi-supervised MVSS setting, where a limited amount of sparsely labeled RGB-Thermal (RGBT) videos and massive unlabeled ones are utilized.

- We propose a simple yet effective semi-supervised MVSS baseline, SemiMV. Our baseline is to our knowledge the first in employing the consistency regularization idea tailored for the semi-supervised MVSS task. Our experimental evaluation on the MVSeg and MVUAV benchmark datasets demonstrates the efficacy of the SemiMV baseline.

2 Related Work

2.1 Semantic Segmentation in Diverse Modalities

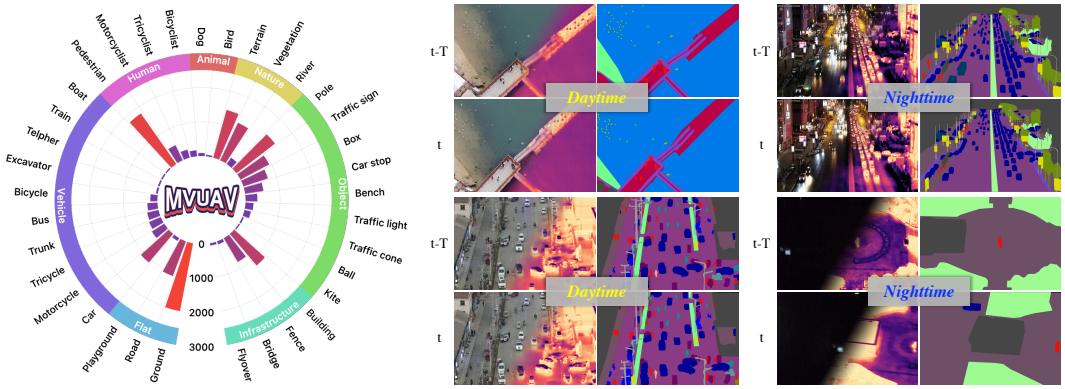
Semantic segmentation, a crucial task in computer vision, has evolved significantly over recent decades. This evolution spans four primary data modalities: RGB images, multispectral images, RGB videos, and multispectral videos, each uniquely contributing to the field's advancement and expanding its practical applications [26, 27, 28, 29, 19, 30, 31, 32].

RGB semantic segmentation (RSS) has undergone remarkable advancements. FCN [33], as a milestone work, uses a fully convolutional network for per-pixel prediction. Subsequent innovations, such as dilation convolution [34], pyramid pooling [35], and attention mechanisms [36, 37, 38], have enriched segmentation by integrating contextual information and capturing global contexts. Recently, transformer-based methods [39, 40, 41, 42] have gained prominence with the advent of Visual Transformer [43]. For a detailed overview, readers are advised to refer to comprehensive survey papers [44]. *Multispectral semantic segmentation (MSS)* is an emergent field that aims to enhance model robustness in adverse lighting conditions by integrating RGB and thermal imagery [45]. The primary challenge in MSS lies in effectively merging RGB and thermal data. Various strategies have been developed, such as concatenation [46, 47], element-wise summation [48, 49], bridging-then-fusing [14], attention-weighted fusion [50], explicit extract and fusion [13], and cross-modal fusion [15]. Additionally, there is a related area of research focused on multimodal RGBD-based segmentation [51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66], which addresses some limitations of purely RGB-based segmentation. On the other hand, *video semantic segmentation (VSS)* is gaining traction with the growing importance of dynamic video processing. VSS models [18, 67, 16, 68, 69, 70] have focused on leveraging temporal contexts of video frames. Notably, some studies [71, 72, 73] utilize optical flow [74] to warp features from neighbouring frames for feature alignment and aggregation. Recently, attention-based methods [17, 75, 76, 77] have been proposed to selectively retrieve information from past frames for improving current frame segmentation, yielding promising results. *Multispectral video semantic segmentation (MVSS)* is a nascent area that combines the strengths of multispectral and temporal contexts. A pioneering study [21] introduces the large-scale MVSeg dataset to benchmark this field, and proposes a baseline model for learning a joint representation from multispectral video input. Our paper contributes a new MVUAV dataset to enrich the dataset diversity in MVSS.

2.2 Semi-supervised Learning in Segmentation

Semantic segmentation has a major challenge in real-world scenarios where only limited pixel-level labels are available due to high expense of human labor. Recently, semi-supervised semantic segmentation has surged in popularity to utilize a plethora of unlabeled data.

Semi-supervised RSS has gained extensive research interest by extending the powerful semi-supervised learning (SSL) techniques. Notably, consistency regularization and self-training have demonstrated great success. Consistency regularization enforces the consistency of the predictions with various perturbations, *e.g.*, input perturbation via augmenting input images [22], feature perturbation



(a) Taxonomic system (b) Dataset visual examples

Figure 3: Illustrations of the proposed MVUAV dataset. (a) Taxonomic system and its histogram distribution showing the number of annotated frames across different categories. (b) Examples of multispectral UAV videos and corresponding annotations in both daytime and nighttime scenarios.

using multiple decoders [23] or feature dropout [24], and network perturbation through different initialization [25]. Meanwhile, self-training methods [78, 79] generate pseudo segments for the unlabeled images and train subsequent models with both human-annotated and pseudo-labeled data. To leverage temporal video information, several semi-supervised VSS methods [78, 80] have been devised. For example, Zhuang *et al.* [80] propose inter-frame feature reconstruction to leverage the ground-truth labels to supervise the model training on unlabeled frames. Despite their effectiveness, these semi-supervised models are limited to single-mode RGB inputs. *To our best knowledge, our work is the first to deal with semi-supervised MVSS problem with RGB-Thermal video inputs.*

3 The MVUAV Dataset

This section describes the construction of the MVUAV dataset and analyzes its statistical results.

Dataset Collection. The main principle of data acquisition is to provide a comprehensive collection of calibrated RGB and thermal infrared video sequences from a new UAV perspective, and furnish high-quality semantic annotations. Toward this objective, we initially gather about 500 RGB-T video sequences from a latest aerial tracking dataset VTUAV [81] (altitudes ranging from 5-20m). The RGB-Thermal videos are captured at diverse environments such as parks, streets, and beaches, under various lighting conditions and seasons. To ensure the quality of our dataset, we remove unqualified videos or frames that are blurry, misaligned, or repetitive. After this selection process, the MVUAV dataset consists of 413 high-quality multispectral UAV videos, with 53,828 paired frames in total.

Dataset Annotation. We then provide pixel-wise semantic labels to the multispectral UAV videos. We employ Labelme toolkit to annotate the multispectral UAV videos. The annotation process presents many challenges. First, the dataset has many complex scenes at adverse conditions, *e.g.*, nighttime, darkness, overexposure, making it difficult to identify target objects from RGB images alone. Second, it is crucial to maintain annotation consistency across video frames, otherwise objects might be misidentified in different frames. Third, the UAV-view presents a more complex and crowded visual field compared to the eye-level view in MVSeg, making object identification and silhouette distinction more challenging. To overcome these challenges, we first overlay thermal heatmaps onto corresponding RGB images to visually aid annotators in identifying objects in complex scenes. Meanwhile, video frames from the same video are annotated by the same person to ensure temporal consistency. Three inspectors review and correct the annotations to make sure temporal consistency, with the assistance of video-level annotation visualization. Due to these challenges, the annotation and quality control process takes about 90 minutes per frame on average. We visualize some representative examples from the MVUAV dataset in Fig. 3.

Dataset Statistics. Table 1 outlines some critical attributes of the MVUAV dataset and related semantic segmentation datasets with different modalities. Our MVUAV dataset comprises 413 RGB-Thermal videos at a frame rate of 25 fps, including 54k image pairs in total and 2,183 annotated image pairs. The new MVUAV dataset can act as a valuable asset to complement the existing MVSeg dataset for more thorough evaluations of various MVSS models. Additionally, the MVUAV dataset

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%

features high-resolution imagery (1920×1080), which is helpful for fine-grained scene parsing. This dataset also provides detailed annotations for a rich set of semantic categories, covering 8 root categories and 36 sub-classes (including background) as shown in Fig. 3, at a high pixel annotation rate of 99.18%. This can facilitate detailed and comprehensive scene understanding. *Additional dataset details and discussions with related UAV-view datasets [88, 89, 90, 91] are provided in the supplementary material.*

Dataset Splits. The dataset is partitioned into training, validation, and test sets, which contain 275, 35, and 103 videos respectively, with 1,464, 171, and 548 annotated masks. One frame is annotated for every 25 frames. Additionally, the test set data involves both daytime and nighttime scenes, consisting of 72 videos with 351 annotated frames and 31 videos with 197 annotated frames, respectively.

4 Methodology

4.1 Problem Definition

The training setting of the semi-supervised MVSS task is illustrated in Fig. 2, where we have a small set of labeled multispectral videos with sparse annotations and a larger corpus of unlabeled multispectral videos. Following the common practice in VSS [75, 17] and MVSS [21], we use video clips as input units. Each video clip contains a sequence of t frame pairs and only the final frame pair in the labeled video clip has semantic annotations. Formally, we denote a multispectral video clip as $\mathcal{V} = \{(I_i^R, I_i^T)\}_{i=1}^{t}$, where (I_i^R, I_i^T) represents the i-th frame pair with spatial resolution of $H \times W$. The labeled set is denoted as $\mathcal{D}^L = \{(\mathcal{V}_n^L, y_n)\}_{n=1}^{n_L}$, which comprises n_L clips, and y_n is the pixel-level semantic labels for the final frame pair of each clip, in a space of C classes. The unlabeled set is denoted as $\mathcal{D}^U = \{\mathcal{V}_n^U\}_{n=1}^{n_U}$, including n_U unlabeled multispectral video clips. Additionally, we use an evaluation set, $\mathcal{D}^V = \{(\mathcal{V}_n^V, y_n)\}_{n=1}^{n_V}$. The goal of Semi-MVSS is to develop a segmentation model that can effectively learn from both $\mathcal{D}^L$ and $\mathcal{D}^U$, and exhibit robust generalization to $\mathcal{D}^V$.

4.2 Proposed SemiMV Framework

Fig. 4 depicts the overall architecture of our SemiMV. The network takes a multispectral video clip as input, which contains a Query frame pair at time step t, and M Memory pairs selected from past frames. In the semi-supervised MVSS setting, only the Query pairs from the labeled video set have ground-truth semantic annotations.

Supervised Training. We first feed the RGB and Thermal pairs into two parallel segmentation networks (Net^R and Net^T), *e.g.*, DeepLabv3+, which generate initial segmentation predictions P_i^R and P_i^T ($i \in [t-M, \cdots, t]$). *For the labeled Query images*, common supervised training is employed on the outputs of two networks using ground-truth segmentation maps, represented by:

$$\mathcal{L}_{sup} = \mathbb{E}_{(I_t^R, I_t^T, y)\in\mathcal{D}^L}(l_{ce}(P_t^R, y) + l_{ce}(P_t^T, y)). \tag{1}$$

Here (I_t^R, I_t^T) is a Query pair in the labeled set and y is the corresponding ground-truth map. l_{ce} denotes the cross-entropy loss function.

Cross-collaborative Consistency Learning. The key challenge in semi-supervised MVSS lies in how to mine effective supervision from *unlabeled* RGB-Thermal videos to complement label-guided training. In semi-supervised RSS, consistency regularization has achieved notable success, benefiting from perturbation-invariant training to enforce consistent predictions across various perturbations of unlabeled RGB images at different processing levels—input [22], feature [23, 24], or network [25].

This success motivates us to consider applying consistency regularization to RGB-thermal videos. To explore this, a simple yet effective Cross-collaborative Consistency Learning (C3L) module is devised, which leverages the unique properties of RGB-Thermal videos to perform perturbation-invariant training. Our intuition is that: RGB and thermal images inherently capture the same scene from distinct sensory perspectives, *i.e.,* visible light and thermal infrared, which provide innate input perturbations; further, processing these multimodal data through two parallel networks with distinct parameters introduces valuable feature-level and network-level perturbations. With these insights, our C3L is devised to apply mutual pseudo supervision between RGB and thermal streams to effectively utilize unlabeled RGB-Thermal frame pairs.

Specifically, we first compute a pair of one-hot pseudo-labels from the initial probabilistic segmentation predictions, P_i^R and P_i^T, using the *argmax* function: $Y_i^R, Y_i^T = argmax(P_i^R), argmax(P_i^T)$. Then, the pseudo-labels are exploited to provide pseudo supervisions to the other stream, defined as:

$$\mathcal{L}_{c3l} = \mathbb{E}_{(I_i^R, I_i^T) \in \mathcal{D}^U \bigcup \mathcal{D}^L} (l_{ce}(P_i^R, Y_i^T) + l_{ce}(P_i^T, Y_i^R)). \tag{2}$$

Ideally, the C3L loss is expected to enhance the robustness of the model by fostering cross-modal consistency between unlabeled RGB-Thermal pairs. However, our experiments indicate a decrease in segmentation performance with this approach, as discussed in Sec. 5.4. We conjecture this decrease is primarily due to the inherent limitations of particular sensors, where the pseudo label generated from either RGB or thermal images alone might be incomplete, causing confused training and error accumulation. To counteract this, we need to consider effective cross-modal collaboration in the C3L framework. In terms of the design of cross-modal collaboration, a lot of fusion strategies [48, 49, 50, 13, 15] have been proposed in the MSS field as discussed in related works. Therefore, our work does not claim a new fusion strategy, but introduces the insights to highlight the importance of cross-modal collaboration for semi-supervised consistency regularization in MVSS. In C3L, we employ the Cross-modal Fusion block (CMF)[2] from [13] in our implementation and discuss other design choices in the supplements. By introducing cross-modal collaboration, the pseudo labels are refined by engaging the complementary information from the alternate stream, leading to significant improved performance.

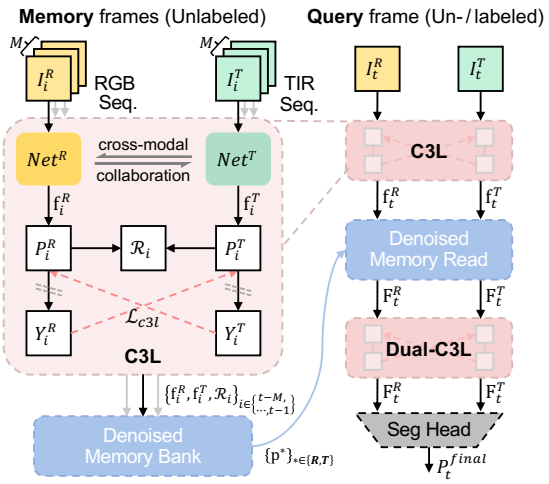


Figure 4: Overview of proposed method. For simplicity, the supervised losses are omitted. The C3L loss $\mathcal{L}_{c3l}$ (Eq. 2) aims to learn from unlabeled RGB-Thermal pairs. The DMR is responsible for integrating temporal information from the denoised memory bank to update query features. A dual-C3L loss (Eq. 7) is further applied to regularize updated query features. Finally, a segmentation head predicts the final mask P_t^{final}. The dotted \\ means stop gradient.

Denoised Memory Read. Next we consider how to integrate temporal information from past video frames. Among solutions in the related VSS and MVSS fields, attention-based memory read methods [21, 75, 17] have yield promising results, which selectively retrieve information from past (Memory) frames for improving current (Query) frame segmentation. In semi-supervised MVSS, due to the absence of ground-truth supervisions for past frames, Memory features are prone to be unreliable.

To deal with this issue, we further introduce a reliability estimation strategy, which can be easily integrated into existing temporal aggregation modules [21, 75] to mitigate potential noise. Here we utilize prototypical memory read [21] for efficient temporal aggregation. Our intuition is that reliable RGB and thermal features tend to yield consistent predictions. Conversely, discrepancies in these predictions can, to a certain extent, suggest potential unreliability. To quantify this, we design a normalized bidirectional Kullback–Leibler (KL) divergence function to estimate the pixel-wise

[2]CMF [13] is built on the gating mechanism that enables the model to emphasize useful features in one modality and compensate its missing information from the other.

reliability map as:

$$\mathcal{R}_i = 1 - \frac{1}{2}\big(\mathcal{N}(\sum_{c\in\mathcal{C}} P_i^R(c)\log\ \frac{P_i^R(c)}{P_i^T(c)}) + \mathcal{N}(\sum_{c\in\mathcal{C}} P_i^T(c)\log\ \frac{P_i^T(c)}{P_i^R(c)})\big). \tag{3}$$

Here, the KL divergence function is performed pixel-wisely (we omit the pixel scripts for simplicity); $\mathcal{N}(\cdot)$ is a min-max normalization function performed spatially to normalize divergence values to the range of 0 to 1. The final reliability map $\mathcal{R}_i$ is with the dimension of $H \times W \times 1$.

To efficiently store reliable memory features with minimal memory usage, we then establish a denoised prototype-based memory bank [21]. Concretely, for each memory feature $\mathbf{f}_i^* \in \mathbb{R}^{H\times W\times D}$ extracted from Net^R and Net^T, where $* \in \{R, T\}$ indicates the image modality and D is the channel dimension, we generate C denoised class-level prototype features by spatially aggregating denoised features belonging to each category:

$$\mathbf{p}_i^* = \mathcal{G}(\mathbf{f}_i^* \times \mathcal{R}_i, Y_i^*) \in \mathbb{R}^{C\times D}, \tag{4}$$

Here $\times$ means pixel-wise multiplication to down-weight unreliable memory features based on the reliability map $\mathcal{R}_i$. $\mathcal{G}$ is the aggregation operation, which spatially averages the features belonging to each class based on pseudo-label Y_i. With this process, we obtain a condensed and denoised prototype-based memory bank $\{\mathbf{p}^* \in \mathbb{R}^{MC\times D}\}_{*\in\{R,T\}}$.

Subsequently, we use the attention mechanism to selectively retrieve relevant semantic information from the denoised memory bank, thereby refining query features. Taking RGB query feature $\mathbf{f}_t^R \in \mathbb{R}^{H\times W\times D}$ as an example, the updated RGB query feature $\mathbf{F}_t^R$ is derived as follows:

$$\mathbf{w}^* = \textit{Softmax}(\bar{\mathbf{f}}_t^R \otimes \textit{transpose}(\bar{\mathbf{p}}^*)), * \in \{R, T\}, \tag{5}$$

$$\mathbf{F}_t^R = \phi([\mathbf{w}^R\mathbf{p}^R, \mathbf{w}^T\mathbf{p}^T, \mathbf{f}_t^R]). \tag{6}$$

Here, $\bar{\mathbf{f}}_t^R$ and $\bar{\mathbf{p}}^*$ indicate L_2 normalized features, $\otimes$ denotes matrix multiplication, $[\cdot,\cdot,\cdot]$ means feature concatenation, and $\phi(\cdot)$ is a convolutional operation to adjust channel size.

The denoised memory read module finally outputs two enhanced query features $\mathbf{F}_t^R, \mathbf{F}_t^T \in \mathbb{R}^{H\times W\times D}$, which are enriched with useful denoised temporal contexts from unlabeled past frames.

Dual-C3L. In order to make full use of the unlabeled data and to further regularize the *memory-updated features*, we add C3L loss on the updated query features as well, called Dual-C3L loss:

$$\hat{\mathcal{L}}_{c3l} = \mathbb{E}_{(I_t^R, I_t^T)\in\mathcal{D}^U \bigcup \mathcal{D}^L}(l_{ce}(\hat{P}_t^R, \hat{Y}_t^T) + l_{ce}(\hat{P}_t^T, \hat{Y}_t^R)), \tag{7}$$

where $\{\hat{P}_t^R, \hat{P}_t^T\}$ are updated predictions inferred from updated query features $\{\mathbf{F}_t^R, \mathbf{F}_t^T\}$, and $\{\hat{Y}_t^R, \hat{Y}_t^T\}$ are corresponding pseudo labels. Accordingly, for the labeled query pairs, an additional supervision loss, $\hat{\mathcal{L}}_{sup}$, is also applied on the updated predictions, similar to Eq. 1.

Final Prediction and Training Objective. To infer the final output, the updated query features $\{\mathbf{F}_t^R, \mathbf{F}_t^T\}$ are concatenated together, followed by a 3×3 convolutional layer as segmentation head to predict the final mask P_t^{final}. A supervised cross-entropy loss is also applied to P_t^{final}, as:

$$\mathcal{L}_{sup}^{final} = \mathbb{E}_{(I_t^R, I_t^T, y)\in\mathcal{D}^L} l_{ce}(P_t^{final}, y). \tag{8}$$

The overall training objective of the proposed SemiMV framework is thus defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \hat{\mathcal{L}}_{sup} + \mathcal{L}_{sup}^{final} + \lambda(\mathcal{L}_{c3l} + \hat{\mathcal{L}}_{c3l}), \tag{9}$$

where λ is the trade-off weight to balance the supervised losses and pseudo losses from C3L.

5 Experiments

5.1 Datasets and Evaluation Metric

Our experiments are conducted on both MVSeg and MVUAV datasets. *MVSeg* [21] is a street scene dataset with 26 semantic classes. It consists of 452, 84, and 202 videos in its training, validation, and testing subsets, respectively. Annotations are provided for every 15 frames, resulting in 2,241 training, 378 validation, and 926 testing semantic masks. *MVUAV* provides bird's-eye view scenes with 36 semantic classes. The dataset splits are mentioned in Sec. 3. We follow the partition protocols of [25] and divide the whole training set via randomly sub-sampling 1/2, 1/4, 1/8, and 1/16 training videos as the labeled set, and treat the remaining videos as the unlabeled set. The numbers of annotated videos/frames for each training partition, denoted as (#videos, #frames), are included in Table 2 and Table 3. By convention [21], we adopt the mean Intersection over Union (mIoU) for evaluation.

5.2 Implementation Details

The model is implemented on the Pytorch and trained using two NVIDIA A100 GPUs. To be consistent with previous work [25, 21], we adopt DeepLabv3+ [34] with ResNet50 as backbone, for both RGB and thermal streams. For the RGB stream, we initialize the network parameters using weights pretrained on ImageNet [92]. For the thermal stream, we randomly initialize the network parameters, and generate 3-channel thermal images as inputs by repeating the 1-channel thermal images. All training images are resized to 320×480. The model is optimized by Adam with batch size of 2, and the learning rate is 2e-4 which is annealed following the poly LR policy. The network converges around 200 epochs. We follow [21] to adopt three reference frames ($M = 3$) at a sample rate of 3 as Memory. λ is empirically set as 1. The network training involves two-stages: the first stage is backbone warming-up trained with only annotated query frames (50 epochs), and the second stage is main-training of SemiMV trained with all videos (150 epochs). During testing, the SemiMV processes each frame sequentially, inferring results in just 19.1 ms per frame.

5.3 Comparison with the State-of-the-Arts

Since our SemiMV is the first work to address semi-supervised MVSS, to provide a reference level, we reimplement some related approaches in *semi-supervised RSS* (MT [22], CCT [23], CPS [25], and UniMatch [24]), *semi-supervised VSS* (IFR [80]), *VSS* (Accel [77]), and *MVSS* (MVNet [21]) fields, using their official codes. These models use ResNet50 [93] as feature extractor.

Results on MVSeg Dataset. Table 2 lists the segmentation results on the MVSeg dataset. We show the fully-supervised performance achieved by training only on the labeled data using different modalities, in the first row of each block. We observe that the semi-supervised models consistently surpass their supervised baselines across all data partitions, highlighting the significance of semi-supervised learning in semantic segmentation. In particular, our SemiMV improves the SupOnly (RGBT) baseline by large margins, *i.e.*, +2.22%, +5.67%, +6.16%, +5.98%, under 1/16, 1/8, 1/4, 1/2 partition protocols, respectively. This demonstrates the effectiveness of our SemiMV in engaging unlabeled multispectral videos to enhance the generalization capabilities of segmentation models. Moreover, compared to MVNet [21], which utilizes sparsely labeled videos alone, our SemiMV consistently performs better. This is attributed to the simultaneous usage of both unlabeled past frames and extensive unlabeled multispectral videos in our SemiMV framework.

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

Results on MVUAV Dataset. Table 3 reports the comparison results on the MVUAV dataset. Benefiting from the integration of rich multispectral video information and the powerful capability of semi-supervised learning to utilize unlabeled data, our SemiMV achieves the highest performance. The improvement of our SemiMV over the supervised baseline SupOnly (RGBT) are +3.82%, +5.16%, +5.21%, +4.63% under 1/16, 1/8, 1/4, 1/2 partition protocols, respectively. Overall, our architecture can adapt to various settings and scenarios by achieving superior performance on both MVSeg and MVUAV datasets. This showcases the robustness of our SemiMV framework.

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

Visual Comparison. Fig. 5 illustrates the qualitative results of different methods on the MVSeg dataset under 1/4 partition protocol. As seen, the MVNet and our SemiMV reconciling both multispec-

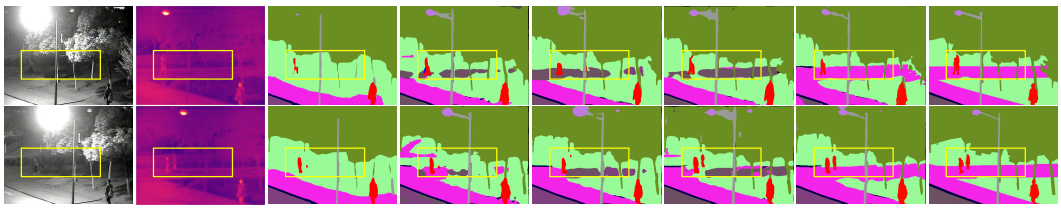


Figure 5: Qualitative results on MVSeg dataset. We highlight the details with the yellow boxes.

tral and temporal contexts can better identify objects in low-light places, for example, the pedestrians within the yellow boxes, while other methods struggle. Moreover, the results of our method are closer to the ground truths than MVNet, attributing to the combined benefits of semi-supervised learning and multispectral video data. *More results are provided in the supplementary materials.*

5.4 Ablation Studies

In this section, we conduct ablation studies on the MVSeg dataset under 1/4 partition setup.

Effect of each component. In Table 4, we evaluate the performance improvements achieved by systematically integrating key components into our framework. Initially, we compare two supervised baselines: RGB *vs.* RGB-Thermal, each trained with labeled RGB frames or RGB-Thermal frame pairs, respectively. It is observed that incorporating a thermal infrared branch brings a notable performance gain of 1.09%, verifying the value of multispectral information in semantic segmentation. As the proposed C3L and DMR are gradually incorporated, increased performance is consistently observed. In particular, C3L enhances the supervised baseline by a remarkable amount of +3.85% (36.88% $\rightarrow$ 40.73%), demonstrating its efficacy in harnessing unlabeled RGB-Thermal frame pairs. The addition of our DMR further boosts performance by +1.66% mIoU, benefiting from the exploitation of denoised and context-rich unlabeled past frames. Notably, our Dual-C3L loss added on updated query features further elevates the mIoU score to 43.04%, amplifying the combined strengths of our C3L and DMR modules in leveraging unlabeled multispectral videos for Semi-MVSS.

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

Analysis on C3L. We then delve into the designs of C3L in Table 5. To effectively utilize unlabeled RGB-Thermal frame pairs, our C3L leverages cross pseudo labels as supervision signals and introduces cross-modal collaboration to enable the generation of reliable pseudo labels. To verify their significance, we develop two variants, C3L_1^- w/o cross supervision and C3L_2^- w/o cross collaboration. In C3L_1^-, the mean-teacher output of each stream is used to generate pseudo label for itself without using cross supervisions. This leads to a reduction of 1.61% mIoU compared to the standard C3L. We conjecture this drop is due to the accumulation of self-errors, whereas our C3L, with its cross supervision, has the ability to mutually correct potential errors, thereby enhancing overall accuracy. Moreover, it is worth noting that the C3L_2^- variant, which applies cross pseudo supervision without the benefit of cross-modal collaboration, leads to a deterioration in result. This indicates that cross-modal collaboration is critical and indispensable for the effective functioning of our Semi-MVSS framework.

Table 5: Ablation analysis of our C3L module.

Ablation analysis on C3L	Labeled	Unlabeled	mIoU
Baseline	✓		36.88
C3L_1^- *w/o cross supervision*	✓	✓	39.12
C3L_2^- *w/o cross collaboration*	✓	✓	36.67
C3L	✓	✓	40.73

Analysis on DMR. Table 6 presents our DMR-Proto and an alternative module - DMR-STM as in [75]. DMR-STM performs an all-to-all attention mechanism for matching between query and memory frames. Our results reveal that, 1) engaging temporal contexts from unlabeled past frames is indeed useful, as both DMR modules yield increased mIoU scores; 2) the proposed reliability estimation strategy

Table 6: Ablation analysis of our DMR module.

Methods	w/o temporal	DMR-STM		DMR-Proto	
		w/o denoised	with denoised	w/o denoised	with denoised
mIoU	40.73	41.47	42.25	41.85	42.39
Δ	-	(+0.74)	(+1.52)	(+1.12)	(+1.66)

effectively filters out unreliable features, enhancing the performance of both DMR variants. Considering computational efficiency, DMR-Proto is used for temporal integration in our SemiMV network. *Additional model analyses are provided in the supplementary materials.*

6 Conclusion

This study introduces MVUAV, a new multispectral video semantic segmentation dataset obtained through UAVs from oblique bird's-eye viewpoint. The dataset, accompanied with precise semantic annotations, serves as a complementary resource to the MVSeg dataset, offering a broader perspective for evaluating MVSS models. Additionally, this paper pioneers the development of SemiMV, the first semi-supervised MVSS framework tailored to utilize both labeled and unlabeled multispectral videos effectively. Comprehensive empirical results affirm the effectiveness of our approach and highlight the promising potential of semi-supervised leaning in MVSS.

Broader Impacts. The proposed MVUAV dataset offers significant value in enhancing the performance of semantic segmentation and propelling further research in the field of MVSS. We envisage that the most proximate impacts of this dataset will be positive, providing a valuable asset for researchers and developers in the field. Our semi-supervised MVSS method makes robust scene parsing available without intensive human annotation efforts, which saves a lot of costs. Note that the proposed MVUAV dataset is strictly prohibited from being used to identify or invade the privacy of any individual and is made available solely for academic purposes. In addition, we discuss potential limitations with some feasible solutions in the supplements.

Reproducibility Statement. Our source code and dataset, along with easy-to-follow instructions, are publicly available on our project website.

7 Acknowledgements

This work was partially supported by the the Alberta Innovates CASBE - NSERC Alliance and the NSERC Discovery (RGPIN-2019-04575) grants, and Samsung AI Center-Mountain View.

References

[1] 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.

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

[3] Qi Bi, Shaodi You, and Theo Gevers. Generalized foggy-scene semantic segmentation by frequency decoupling. In *CVPR*, pages 1389–1399, 2024.

[4] Qi Bi, Shaodi You, and Theo Gevers. Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In *AAAI*, pages 801–809, 2024.

[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.

[6] 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.

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

[8] Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In *ACM MM*, 2024.

[9] Qi Bi, Shaodi You, and Theo Gevers. Interactive learning of intrinsic and extrinsic properties for all-day semantic segmentation. *IEEE Transactions on Image Processing*, 32:3821–3835, 2023.

[10] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, 21:617–630, 2024.

[11] Munan Ning, Donghuan Lu, Yujia Xie, Dongdong Chen, Dong Wei, Yefeng Zheng, Yonghong Tian, Shuicheng Yan, and Li Yuan. Madav2: Advanced multi-anchor based active domain adaptation segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13553–13566, 2023.

[12] 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*, 24:2526–2538, 2021.

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

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020.

[24] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, pages 7236–7246, 2023.

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

[26] Qi Bi, Jingjun Yi, Hao Zheng, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning generalized medical image segmentation from decoupled feature queries. In *AAAI*, pages 810–818, 2024.

[27] 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.

[28] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *AAAI*, pages 6030–6038, 2024.

[29] 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.

[30] 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.

[31] Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *ICML*, pages 60906–60926, 2024.

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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *AAAI*, pages 819–827, 2024.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[44] Irem Ulku and Erdem Akagündüz. A survey on deep learning-based architectures for semantic segmentation on 2d images. *Applied Artificial Intelligence*, pages 1–45, 2022.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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*, 18(3):1000–1011, 2020.

[50] 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.

[51] 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.

[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.

[53] 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.

[54] 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.

[55] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020.

[56] 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, 2020.

[57] 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.

[58] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Xiang Ruan. Self-supervised pretraining for rgb-d salient object detection. In *AAAI*, pages 3463–3471, 2022.

[59] 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.

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

[61] 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.

[62] 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.

[63] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, pages 235–252, 2020.

[64] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, pages 892–904, 2023.

[65] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *International Journal of Computer Vision*, pages 1–78, 2024.

[66] 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.

[67] 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*, pages 522–539, 2022.

[68] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *CVPR*, pages 5997–6005, 2018.

[69] Xiaoqi Zhao, Shijie Chang, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *International Journal of Computer Vision*, pages 1–19, 2024.

[70] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pages 1553–1563, 2021.

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

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

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

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, pages 695–714, 2020.

[79] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, pages 4248–4257, 2022.

[80] Jiafan Zhuang, Zilei Wang, and Yuan Gao. Semi-supervised video semantic segmentation with inter-frame feature reconstruction. In *CVPR*, pages 3263–3271, 2022.

[81] 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.

[82] 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.

[83] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108–119, 2020.

[84] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020.

[85] Haitao Xiong, Wenjie Cai, and Qiong Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, 113:103628, 2021.

[86] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *ICCV*, pages 8115–8124, 2023.

[87] Connor Lee, Matthew Anderson, Nikhil Raganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Cart: Caltech aerial rgb-thermal dataset in the wild. *arXiv preprint arXiv:2403.08997*, 2024.

[88] 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.

[89] 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.

[90] 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.

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

[92] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We outline our main contributions at the end of Sec. 1. The claims made match the dataset description in Sec. 3, methodology in Sec. 4, and experimental results in Sec. 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Sec. 4 provides a detailed explanation of the proposed SemiMV.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The proposed method is illustrated in Sec. 4, and our implementation details are provided in Sec. 5.2. The MVSeg dataset used in the experiments are public available from previous works. Besides, we will release the code and our new dataset on our website.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The authors will make the newly proposed dataset and source code, along with detailed instructions, publicly available on our project website.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details are given in Sec. 3 and Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We show the reliability of the proposed method by conducting multiple running experiments on the SemiMV, as in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details are given in Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirmed that we conform the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the potential societal impacts in Sec. 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We prohibit people from using our MVUAV in any manner to identify or invade the privacy of any person. Additionally, our MVUAV dataset is made freely available solely for academic purposes.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original papers for all assets used in our research, which are publicly available to the research community.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly proposed MVUAV dataset is detailed in Sec. 3. We will also release it along with user-friendly usage guidelines.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.