Unlocking Drone Perception in Low AGL Heights: Progressive Semi-Supervised Learning for Ground-to-Aerial Perception Knowledge Transfer

Junjie Hu, *Member, IEEE*, Chenyou Fan, *Member, IEEE*, Mete Ozay, *Member, IEEE*, Hua Feng, Yuan Gao, *Member, IEEE*, and Tin Lun Lam, *Senior Member, IEEE*

***Abstract*— We explore the novel challenge of drone perception across varying low AGL (above ground level) heights, a task essential for dynamic tasks, unlike the fixed ground viewpoint in autonomous driving. Supervised learning for this incurs high annotation costs, and current semi-supervised methods struggle with viewpoint differences. In this paper, we introduce ground-to-aerial perception knowledge transfer and propose a progressive semi-supervised learning framework for drone perception using only labeled data from the ground viewpoint and unlabeled data from flying viewpoints. The framework hinges on four key components: 1) a dense viewpoint sampling strategy, segmenting the vertical flight height range into evenly distributed intervals; 2) nearest neighbor pseudo-labeling, inferring labels of the nearest neighbor viewpoint using a model learned on the preceding viewpoint; 3) MixView, generating augmented images among different viewpoints to mitigate viewpoint differences; and 4) a progressive distillation strategy, gradually learning until reaching the maximum flying height. To validate our approach, we create both synthesized and real-world datasets. Extensive experimental analyses reveal a remarkable relative accuracy improvement of 25.7% and 16.9% for the synthesized dataset and the real world, respectively. Code and datasets are available on https://github.com/FreeformRobotics/Progressive-Self-Distillation-for-Ground-to-Aerial-Perception-Knowledge-Transfer.**

Received 6 December 2023; revised 17 December 2024 and 12 February 2025; accepted 8 April 2025. Date of publication 15 May 2025; date of current version 1 July 2025. This work was supported in part by Shenzhen Science and Technology Program under Grant RCBS20231211090736065, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A151511, in part by Guangdong Natural Science Fund under Grant 2024A1515010252, and in part by Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects under Grant LGKCSDPT2024002/LGKCSDPT2024003. The Associate Editor for this article was K. Gao. *(Corresponding author: Tin Lun Lam.)*

Junjie Hu is with the School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen 518172, China (e-mail: hujunjie@cuhk.edu.cn; gaoyuan@cuhk.edu.cn).

Chenyou Fan is with the School of Artificial Intelligence, South China Normal University, Guangzhou 510631, China (e-mail: fanchenyou@scnu.edu.cn).

Mete Ozay is with Samsung Research and Development Institute, TW18 4QE Surrey, U.K. (e-mail: meteozay@gmail.com).

Hua Feng is with Guangzhou iMapCloud Intelligent Technology Company Ltd., Guangzhou 510095, China (e-mail: fenghua0thomershen@gmail.com).

Yuan Gao is with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen 518000, China, and also with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: gaoyuan@cuhk.edu.cn).

Tin Lun Lam is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China, and also with AIRS, Shenzhen 518000, China (e-mail: tllam@cuhk.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2025.3564977, provided by the authors.

Digital Object Identifier 10.1109/TITS.2025.3564977

***Index Terms*— Knowledge transfer, drones, semisupervised learning, semantic segmentation, multi-robot systems.**

I. Introduction

IN RECENT years, the integration of drones in various fields, including surveillance [1], infrastructure inspection [2], intelligence logistics [3], low-altitude economy [4], and intelligent transportation [5], has surged dramatically. An essential aspect of maximizing the utility of drones lies in their ability to perceive and understand their surroundings accurately. Particularly, the significance of drone perception becomes paramount when operating at low AGL (above ground level) heights, such as 5 to 10 meters, to execute sophisticated operations.

The contemporary machine learning paradigm, employing data-driven deep learning solutions, has demonstrated efficacy in various perception tasks like semantic segmentation [6], [7], [8], object detection [9], [10], and depth estimation [11], [12], [13], providing a potential avenue for addressing autonomous driving challenges. Recent studies on drone perception attempt to extend this paradigm from cars to drones, involving the preparation of labeled drone training sets and subsequent deep network learning for specific tasks [14], [15], [16].

Unfortunately, prior works often overlook a fundamental distinction in perception between drones and cars. In many instances, perception is assumed to occur at a specific flying height, e.g., 10 meters, as demonstrated in [17] and [18]. Practical scenarios, as illustrated in Fig. 1, reveal that drones may appear at arbitrary flight heights due to task-specific allocations. Given a maximum flight height h, an essential requirement for drone perception is the ability to accurately perceive from various flight heights within the range of $[0, h]$ meters. The fundamental challenge is the inaccuracy caused by viewpoint differences among different flight heights.

However, addressing this problem through supervised learning is challenging, necessitating an impractical amount of training data to match the performance of autonomous driving cars. Data acquisition for ground truths is costly and time-consuming, posing an obstacle to developing data-driven approaches. Moreover, semi-supervised methods [19], [20], [21] face limitations in handling substantial variations in flying viewpoints across different flying heights. Multi-view methods, such as structure from motion [22] or multi-view collaborative perception [23], [24], require multi-view input

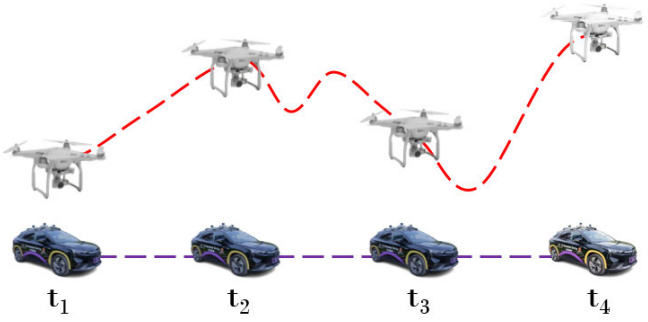


Fig. 1. An example of a trajectory comparison between a drone and a car. One essential difference is that a drone can optionally change its flight height at time instance $\mathbf{t}_j$, $j = 1, 2, 3, 4$.

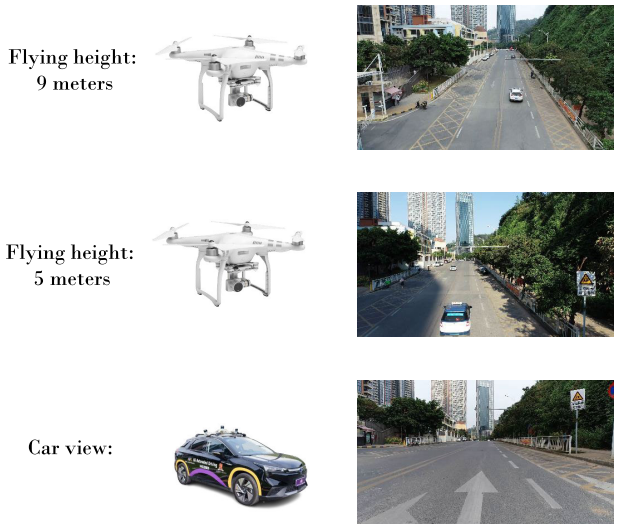


Fig. 2. Instances illustrating viewpoint variations, lighting changes, and the presence of dynamic objects, such as cars, at different altitudes in real-world scenarios.

images as inputs and are unsuitable for online semantic perception tasks predicting by a single agent from a single image. Moreover, they encounter difficulties coping with significant viewpoint differences, lighting changes, and dynamic objects in real-world scenarios, as illustrated in Fig. 2.

In this paper, we introduce the ground-to-aerial (GoA) perception knowledge transfer, a method that enables the transfer of perception knowledge from a UGV trained for ground-viewpoint perception to drone perception tasks, eliminating the need for additional data annotation at various flying heights beyond the ground viewpoint. To achieve this, a semi-supervised learning (SSL) approach is adopted, utilizing labeled data exclusively from the ground viewpoint while leveraging unlabeled data from aerial viewpoints. To address the challenges posed by viewpoint variations, the vertical flying range is first discretized into evenly spaced height intervals, with data sampled at each viewpoint. Although large viewpoint differences result in significant performance degradation, observations indicate that the nearest viewpoint yields similar accuracy. Based on this insight, nearest-neighbor pseudo-labeling is introduced, where predictions for viewpoint h_i are made using a network trained on its nearest preceding viewpoint h_{i-1}. Additionally, MixView is proposed as a data augmentation strategy that blends information from different viewpoints to generate augmented images, improving robustness to viewpoint shifts. Finally, a progressive SSL framework is developed to iteratively adapt the model from the ground viewpoint to higher altitudes, ultimately enabling perception across the full flight range.

To support our investigation, we generated a synthetic dataset using AirSim [25] spanning flying heights from 1 to 10 meters and collected a real-world dataset in an urban environment, covering flying heights from 1 to 9 meters. Both datasets encompass images taken from various flying heights. Our inquiry centers on drone perception for semantic segmentation, a fundamental task in robot perception. We assess our methodology on these datasets, offering a comprehensive analysis that includes both quantitative metrics and qualitative insights.

In summary, our contributions include:

- To the best of our knowledge, this study marks the inaugural attempt to address drone perception under significant viewpoint changes induced by varying flight heights. Our goal is to enable precise drone perception across diverse flying altitudes.
- A novel semi-supervised learning framework designed to distill ground-to-aerial perception knowledge, leveraging labeled images from the ground viewpoint alongside unlabeled images from varying flight perspectives.
- Two intuitive yet effective methods: i) nearest neighbor pseudo-labeling and ii) MixView, designed to address the substantial viewpoint changes associated with varying flying heights.
- Two publicly available datasets curated for drone perception in semantic segmentation, including a synthetic dataset from AirSim with fixed lighting conditions and static objects, as well as a real-world dataset captured in an urban environment with changing lighting conditions and dynamic objects.
- Comprehensive quantitative and qualitative assessments to validate the efficacy of the proposed method, encompassing both simulated and real-world datasets. The evaluation involves comparisons with various SSL methods and ablation studies for a thorough analysis.

The remainder of this paper is organized as follows. Section II provides a comprehensive review of background information and related studies. Section III outlines our progressive semi-supervised learning framework. The datasets used for validation are detailed in Section IV. Section V presents a thorough examination of our methodology through extensive numerical evaluations. Finally, Section VI concludes this paper.

II. Related Work

A. Drone Perception in Low AGL Heights

Accurate drone perception at low AGL heights is paramount for precise navigation through complex environments, ensuring obstacle avoidance and enhancing safety. Similar to contemporary studies on self-driving cars, previous research approaches this challenge as a classic data-driven learning problem. This involves the collection of a training dataset

and the subsequent training of a deep neural network for perception tasks. Early works addressed issues such as indoor gate detection for drone navigation using convolutional neural networks [26]. Others aimed to directly learn navigation in GPS-denied indoor corridor environments [27]. This trend is also evident in object detection [28], [29], [30], as well as semantic segmentation [31], [32]. Approaches rooted in supervised learning have demonstrated advanced performance in tackling these challenges.

A fundamental distinction between drone perception and autonomous driving lies in the freedom of a drone to fly at arbitrary heights, whereas self-driving cars typically conduct perception solely from a ground viewpoint. Consequently, drone perception poses a greater challenge, requiring precision across various flying altitudes. While a direct approach would involve labeling data from diverse viewpoints and addressing the problem within a purely supervised learning framework, such an endeavor is impractical due to the considerable effort required for data labeling. Therefore, in this paper, we present a more pragmatic solution—a semi-supervised approach designed to facilitate drone perception under significant viewpoint variations.

B. Multi-Agent/View Perception

Numerous prior studies have explored collaborative perception among multiple agents or views. In [23] and [33], online collaboration methods were introduced to enhance the performance of a single agent's view when its captured images experienced degradation. Also, the collaborative bird's-eye-view (BEV) perception, utilizing LiDAR scans from multiple vehicles, has been investigated in [34] and [35].

More recently, an emerging approach based on few-shot learning [24], [36] has been developed to address multi-view perception in scenarios with limited data availability. However, these methods commonly assume minimal viewpoint differences between multi-view images. Additionally, they necessitate synchronous multi-view input images, rendering them unsuitable for our online drone semantic perception tasks, which predict from a single image.

C. Semi-Supervised Learning

Semi-supervised learning (SSL) is a dynamic area of research extensively explored in machine learning, aiming to develop highly discriminative models using a combination of labeled and unlabeled samples [37], [38]. While prior studies have demonstrated promising outcomes, particularly in the context of deep learning, we outline three fundamental approaches. (i) Generative Learning Approach: This method involves the generation of additional data from real data distribution using generative adversarial networks [39], [40] or variational auto-encoders [41]. (ii) Pseudo-Labeling Based Methods: These approaches first train a deep model on labeled data in a supervised manner and subsequently leverage it to predict pseudo-labels from unlabeled data [20], [42], [43], [44], [45]. (iii) Data Augmentation Methods: Techniques like MixUp [19] and MixMatch [46] propose blending two images and their corresponding labels at the pixel level, showcasing enhanced performance. In contrast, CutMix [47], ClassMix [48], and UniMatch [21] mix object-label information, making them particularly applicable to semantic segmentation.

Despite significant advancements in prior self-supervised learning (SSL) methods, their direct application to drone perception tasks remains challenging. The substantial differences between ground and drone viewpoints create obstacles, as pseudo-labeling methods often produce inaccurate pseudo-labels, and data augmentation techniques struggle to generate effective training pairs of images and semantic maps. Furthermore, generative approaches fail to produce realistic flying data due to the lack of labeled datasets from flying viewpoints.

To overcome these challenges associated with viewpoint differences in drone perception tasks, we propose nearest-neighbor pseudo-labeling and MixView, integrated within a progressive learning framework. The nearest-neighbor pseudo-labeling method mitigates the impact of significant viewpoint discrepancies by intuitively discretizing the altitude range into a set of intervals and assigning labels based on the most similar viewpoint. Therefore, it enables more accurate estimation of pseudo-labels. Meanwhile, MixView enhances data augmentation by effectively combining data from different viewpoints. The "manifold" assumption in SSL posits that the data space is constituted by multiple lower-dimensional manifolds, where data points sharing the same manifold should share the same label [38]. Viewpoint discrepancies, however, introduce misclassification in the data space. To address this challenge, MixView introduces a regularization mechanism to mitigate viewpoint differences, facilitating the repositioning of misclassified data points onto their correct manifolds. Together, these methods address the unique challenges of drone perception and demonstrate improved performance under diverse flying conditions.

D. Knowledge Distillation

Knowledge distillation (KD) originated as a technique to transfer knowledge from a well-trained expert model to a target local model in the domain of image recognition [49]. Since its inception, KD has found applications in knowledge transfer across various tasks, spanning depth estimation [50], depth completion [51], semantic segmentation [52], object detection [53], and even large language models [54].

Nevertheless, effective knowledge distillation relies on access to the original training set or a closely related alternative dataset. Given the significant viewpoint differences inherent in GoA distillation, we propose a novel progressive semi-supervised learning approach. This approach is designed to enhance the effectiveness of distillation in data-scarce scenarios where viewpoint variations pose notable challenges.

III. Technical Approach

A. Overview

Let h_1 to h_n denote a sequence of viewpoint heights, where h_1 represents the ground-level camera height and h_n corresponds to the maximum flying height of a drone. Our goal is to enable a drone to achieve accurate perception

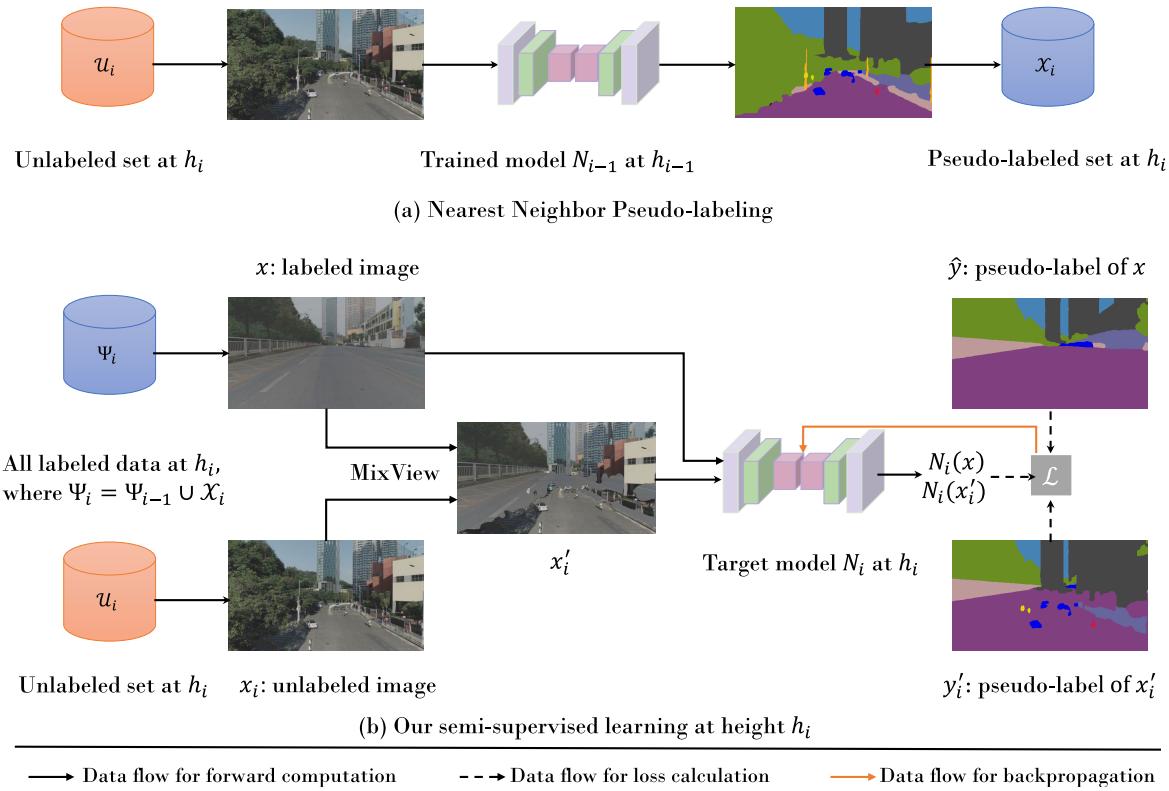


Fig. 3. Illustration of the proposed semi-supervised learning-based self-distillation method. The perception model N_i at height h_i is learned through distillation from a model N_{i-1} trained at the preceding height h_{i-1}. Additionally, the model is augmented with viewpoint-mixed images, which are synthesized by combining a sampled image from the current flying height and images from previous heights.

across a range of low AGL heights, from h_1 to h_n, using only labeled data at h_1 and unlabeled data at various flying heights. This challenge arises because, at higher altitudes, objects often appear significantly smaller and distorted in camera coordinates, making supervised learning the preferred approach for such scenarios. However, for low AGL heights, viewpoint differences significantly impact perception performance, necessitating a more adaptive solution.

To address this, we formulate the problem as a semi-supervised Ground-to-Aerial (GoA) knowledge transfer task. Specifically, we aim to distill the perception capabilities of a model N_1, trained on labeled ground-level data at h_1, to a drone operating at heights up to h_n. We propose a progressive learning strategy that begins with knowledge transfer from the ground-level model N_1 and iteratively extends to subsequent heights until reaching h_n.

As illustrated in Fig. 3, our progressive semi-supervised learning (SSL) framework facilitates the transfer of knowledge from height h_{i-1} to h_i. After n progressive steps, covering all heights from h_1 to h_n, we obtain the final model N_n. This model is capable of robustly handling images captured at any height within the specified range, effectively bridging the gap between ground and aerial perception.

B. Dense Sampling of Flying Height

One crucial aspect contributing to the efficacy of the proposed approach is the adoption of dense sampling of a drone's flying height. By addressing significant viewpoint differences through the gradual incorporation of intermediate viewpoints, these intermediates serve as supportive perspectives. Consequently, a higher number of assistant viewpoints is directly correlated with improved accuracy.

Formally, for a maximum flight height h, we advocate for sampling flight data at uniform intervals of h/n using the equation:

$$h_i = h/n \times i, \tag{1}$$

where $i \in [1, n]$. Eq. (1) divides the flight range $[0, h]$ into intervals $[h_1, h_2, \ldots, h_n]$, where $h_1 < h_2 < \ldots < h_n$.

We denote $\mathcal{X}_1$ as the labeled set of the ground viewpoint, and $\mathcal{U}_i$ and $\mathcal{X}_i$ represent the unlabeled set and its pseudo-labeled counterpart for the flying height h_i where $i > 1$, respectively, throughout this paper.

C. Nearest Neighbor Pseudo-Labeling

Fig. 4 quantifies the accuracy of N_1 at various flight heights. Evidently, the performance deteriorates with increasing viewpoint differences. However, a closer examination reveals a relatively small performance gap between adjacent viewpoints, such as those at 2 meters and 3 meters. Consequently, it is intuitive to leverage SSL between a viewpoint and its nearest neighbor. We introduce nearest neighbor pseudo-labeling, a method that applies pseudo-labeling to the nearest viewpoint.

Let (x_1, y_1) be a pair at h_1, and x_2 be an image at h_2, where $(x_1, y_1) \in \mathcal{X}_1$ and $x_2 \in \mathcal{U}_2$. A straightforward approach to infer a label at h_2 is by utilizing N_1 learned at h_1, expressed as:

$$\hat{y}_2 = N_1(x_2). \tag{2}$$

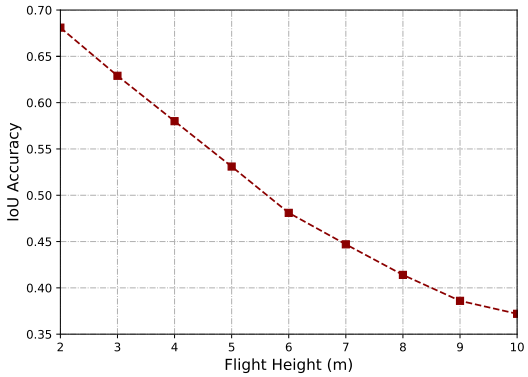


Fig. 4. The mean intersection-over-union (IoU) accuracy of a model trained on data collected from the ground viewpoint across various flight heights.

Here, $\hat{y}_i$ represents the one-hot label of x_i for all i, referred to as a "pseudo-label" as it serves as an approximation to the correct semantic map.

Eq. (2) allows an implicit entropy minimization to improve SSL and is effective due to small viewpoint differences. Then, we can learn a better model at h_2 by mixing data from the nearest neighbors of h_1 and h_2. The model N_2 is learned by minimizing the objective by

$$\min_{N_2} \frac{1}{|\mathcal{X}_1|} \sum_{(x_1,y_1)\in\mathcal{X}_1} \mathcal{H}(y_1, N_2(x_1)) + \frac{1}{|\mathcal{U}_2|} \sum_{x_2\in\mathcal{U}_2} \mathcal{H}(\hat{y}_2, N_2(x_2)), \tag{3}$$

where $\mathcal{H}$ is the cross entropy function. The pseudo-labels are obtained at h_i by

$$\hat{y}_i = N_{i-1}(x_i), \tag{4}$$

where N_{i-1} denotes a model trained at h_{i-1}. To avoid confusion, we simplify our notation by stating that the ground truths y_1 at the ground viewpoint are equivalent to $\hat{y}_1$ for the rest of the paper.

Then, the optimization objective for learning N_i can be written as

$$\begin{aligned} &\min_{N_i} \frac{1}{|\Psi_i|} \sum_{(x,\hat{y})\in\Psi_i} \mathcal{H}(\hat{y}, N_i(x)), \\ &\text{s.t. } \mathcal{X}_i = N_{i-1}(\mathcal{U}_i), \\ &\quad \Psi_i = \bigcup_{k=1}^{i} \mathcal{X}_k. \end{aligned} \tag{5}$$

D. MixView

Ideally, the recognition of objects, such as cars, should remain accurate across different flight heights, even when their appearance or shape varies among viewpoints. In this context, MixView is introduced to enforce the desired viewpoint invariance. It is hypothesized that correct object recognition should persist even when an object is manually relocated to a different viewpoint. Building on this premise, MixView is employed to amalgamate data samples from different viewpoints at the object level, thereby generating augmented data.

MixView can be viewed as a variant of ClassMix [48] tailored for handling SSL under viewpoint differences. In contrast to augmenting solely the unlabeled set, we enhance the original ClassMix by blending both the labeled and unlabeled sets to produce viewpoint-robust images. MixView is applied during the online training of N_i, enabling the model to progressively learn to recognize objects under varying viewpoints. The formulation of MixView is expressed as follows:

$$\begin{aligned} (x_i', y_i') &= MixView\left((x, \hat{y}), (x_i, \tilde{y}_i)\right), \\ x_i' &= m \odot x_i + (1 - m) \odot x, \\ y_i' &= m \odot \tilde{y}_i + (1 - m) \odot \hat{y}, \end{aligned} \tag{6}$$

where $(x, \hat{y}) \in \Psi_i$, $\tilde{y}_i = N_i(x_i)$, m is a binary mask that randomly exchanges half of the classes between x_i and x to generate a new image x_i'. $\odot$ denotes element-wise multiplication.

E. Progressive Self-Distillation

The progressive self-distillation learning strategy integrates the nearest neighbor pseudo-labeling and MixView. Starting with a labeled set $\mathcal{X}_1$ and a trained model N_1 at h_1, alongside unlabeled samples $\mathcal{U}_2, \mathcal{U}_3, \ldots, \mathcal{U}_n$ taken from $h_2, h_3, \ldots, h_n$, respectively, the objective is to progressively transfer perception abilities from N_1 to various flying heights.

The model N_i is trained using the labeled set $\Psi_i = \bigcup_{k=1}^{i} \mathcal{X}_k$ and augmented data from $\mathcal{U}_i$. The final model N_n is trained iteratively across all flying heights by:

$$\begin{aligned} &\min_{N_n} \mathcal{L}, \\ &\text{s.t. } \mathcal{X}_i = N_{i-1}(\mathcal{U}_i), \\ &\quad \Psi_i = \bigcup_{k=1}^{i} \mathcal{X}_k, \\ &\quad (x_i', y_i') = MixView\left((x, \hat{y}), (x_i, \tilde{y}_i)\right), \end{aligned} \tag{7}$$

where

$$\mathcal{L} = \sum_{i=2}^{n} \left(\frac{1}{|\Psi_i|} \sum_{(x,\hat{y})\in\Psi_i} \mathcal{H}(\hat{y}, N_i(x)) + \lambda \frac{1}{|\mathcal{U}_i|} \sum_{x_i'\in\mathcal{U}_i} \mathcal{H}(y_i', N_i(x_i')) \right) \tag{8}$$

and $\lambda \in [0, 1]$ is a hyperparameter varying from 0 to 1 during the optimization.

The detailed implementation of our method is given in Algorithm 1. As shown, the inputs to our algorithm consist of a labeled set at the ground viewpoint h_1, an unlabeled set spanning various flying heights from h_2 to h_n, and a baseline model N_1 trained on h_1, which remains fixed during subsequent operations. We then apply our progressive learning framework, which sequentially applies nearest-neighbor pseudo-labeling followed by MixView for each flying height. This framework gradually adapts the model's perception capabilities to accommodate varying flying heights, ensuring enhanced performance across the altitude range.

Algorithm 1 Algorithm of Our Progressive SSL Framework

Input: $\mathcal{X}_1$: A labeled set of image pairs (x_1, y_1) at the ground viewpoint h_1; $\mathcal{U}_i$: An unlabeled set from the flying height h_i, $\forall i = 2, 3, \ldots, n$; N_1: The baseline model for ground viewpoint perception.

Hyperparameters: Initial learning rate: 0.01, weight decay: $1e^{-4}$, number of training steps: *iterations.*

Output: *N_n: The final model trained on images captured from viewpoints h_1 to h_n.*

Freeze N_1;
▷ *% Progressive Learning %*
for *$i = 2$ to n* ***do***
  ▷ *% Applying the nearest neighbor pseudo-labeling %*
  $\mathcal{X}_i = N_{i-1}(\mathcal{U}_i)$
  $\Psi_i = \bigcup_{k=1}^{i} \mathcal{X}_k$
  Initialize N_i;
  for *j = 1 to iterations* ***do***
    Set gradients of N_i to 0;
    Select $(x, \hat{y})$ from Ψ_i and x_i from $\mathcal{U}_i$;
    $\tilde{y}_i = N_i(x_i)$;
    ▷ *% Applying MixView %*
    $(x_i', y_i') = MixView\big((x, \hat{y}), (x_i, \tilde{y}_i)\big)$;
    Calculate the loss $\mathcal{L}$ with Eq. (8);
    Backpropagate $\mathcal{L}$;
    Update N_i;
  end for
  ▷ *% Updating $\mathcal{X}_i$ %*
  $\mathcal{X}_i = N_i(\mathcal{U}_i)$
end for

IV. Datasets for Drone Perception

In this section, we present the datasets utilized for evaluating our proposed method. Given the absence of a pre-existing dataset suitable for our specific problem, we designed two datasets, AirSim-Drone and AIRs-Street, to comprehensively assess our method in both simulated and real-world scenarios.

A. AirSim-Drone Dataset

The AirSim-Drone dataset, as shown in Fig. 5 (a), is synthetically generated using Microsoft AirSim, capturing data from both ground and aerial perspectives. The drone's flight route adheres to a predetermined map, maintaining fixed environmental conditions. Data is collected at varying flight heights from 2 to 10 meters, with an additional random flight test sequence ('uav_random') for evaluating generalization performance.

This dataset comprises nine semantic attributes, including "Plant," "Building," "Road," "Sky," "Car," "Ground," "Fence," "Pole," and "Others." Each image features a resolution of 768×432. Table I provides detailed information, such as flying height, image count, and labeled semantic maps. It is important to note that the number of captured images in each sequence may vary slightly due to the manual determination of start and end times for video recordings. Ground truth semantic annotations are available for all RGB images, with only the ground viewpoint's ground truths used for training and others reserved for validation."

TABLE I
The Detailed Information of Our AirSim-Drone Dataset

Sequences	Flight height	All samples	Labeled samples
car01	1	1947	1947
uav02	2	1057	1057
uav03	3	1057	1057
uav04	4	1043	1043
uav05	5	1046	1046
uav06	6	1047	1047
uav07	7	1048	1048
uav08	8	1048	1048
uav09	9	1037	1037
uav10	10	1037	1037
uav_random	2 to 10	1143	1143

TABLE II
The Detailed Information of Our AIRs-Street Dataset

Sequences	Flight height	All samples	Labeled samples
car01	1	770	117
uav02	2	750	25
uav03	3	717	25
uav04	4	548	25
uav05	5	591	25
uav06	6	475	25
uav07	7	428	25
uav08	8	657	25
uav09	9	623	25

B. AIRs-Street Dataset

The AIRs-Street dataset, as shown in Fig. 5 (b), is collected in an urban street scenario. To precisely gauge ground distances, we employ a DJI Phantom Pro 4 equipped with a 3D Time-of-Flight (TOF) sensor for data acquisition. Manual control is exercised to ensure the drone follows a predefined route, introducing minimal deviations in flying height. Care is taken to maintain a slow forward speed, limiting flight height deviations to within ± 0.5 meters.

Similar to the AirSim-Drone dataset, images are obtained at different flight heights ranging from 1 meter to 9 meters with a 1-meter interval, resulting in a total of 9 sequences. The captured images have a resolution of 1920×1080.

Due to variations in flying speed, the number of captured images differs among flying heights. Nonetheless, all sequences share the same start and endpoint. For the ground viewpoint sequence, 117 frames are uniformly selected for annotation and training. Each flight sequence involves the uniform selection of 25 frames for annotation and evaluation over the video length.[1] We manually annotate pixel-wise attribute labels for 13 semantic classes, including "Plant," "Building," "Road," "Sky," "Car," "Sidewalk," "Pedestrian," "Motorcycle," "Wall," "Fence," "Traffic Sign," "Traffic Light," and "Others." Additional details are provided in Table II.

V. Experimental Evaluation

A. Implementation Details

1) Implementation of Our Method: Due to GPU memory constraints, we resize the original image resolution from 768×432 to 384×216 for AirSim-Drone and from 1920×1080 to 480×270 for AIRs-Street, respectively. For semantic

[1]Owing to lack of additional labeled images, we directly use these labeled flight data for evaluation.

(a) AirSim-Drone (b) AIRs-Street

Fig. 5. Examples of selected images from our datasets. (a) shows nine images of the flight heights from 1 meter to 9 meters in AirSim-Drone, and similarly, (b) shows nine images of the flight heights from 1 meter to 9 meters in AIRs-Street.

segmentation network architecture, we employ the pre-trained DeepLabV3+ [58] on the CityScape dataset [59], modifying the output layers to align with the semantic categories in our datasets. Training is conducted for 500 epochs for each flying height, utilizing Algorithm 1 and an NVIDIA GeForce RTX 2080 Ti. We employ the SGD optimizer [60] with a learning rate of 0.01 and a weight decay of 0.0001, and apply the polynomial learning rate decay throughout training. All experiments are implemented using PyTorch [61].

2) Baseline Methods: We identify the following methods as baselines for our comparative analysis. All methods are derived from the same model built on DeepLabv3+ with the ResNet-101 backbone.

- Ground-only: As the drone's perception capability is initiated by distilling knowledge from the model trained at the ground viewpoint, we designate the trained model at the ground viewpoint as a baseline, denoted as Ground-only.
- Pseudo-Labeling [55]: This approach employs the ground-only model to predict pseudo-labels for all unlabeled flying data and subsequently trains the model using both labeled and pseudo-labeled data.
- ClassMix [48]: As a representative method of SSL, ClassMix performs data augmentation by randomly exchanging objects between two images.
- S4MC [45]: Representing the current state-of-the-art on PASCAL VOC 2012 with 50% labeled data using DeepLabv3+ network.
- UniMatch [21]: Representing the current state-of-the-art on PASCAL VOC 2012 with 12.5% labeled data using DeepLabv3+ network.
- CorrMatch [56]: demonstrating second best performance on PASCAL VOC 2012 with 6.25% labeled data using DeepLabv3+ network.
- PrevMatch [57]: Representing the current state-of-the-art on PASCAL VOC 2012 with 6.25% labeled data using DeepLabv3+ network.

The hyper-parameters of Pseudo-Labeling and ClassMix are set to be the same as our method. For S4MC, UniMatch, CorrMatch, and PrevMatch, we utilize their original implementations to ensure fair comparisons.

B. Evaluation Metrics

For an equitable evaluation of semantic segmentation, we utilize metrics commonly employed in previous studies, specifically calculating the mean intersection-over-union (IoU) value. The mean IoU, also known as the Jaccard index, serves to quantitatively measure the overlap percentage between the target mask and the prediction output. It is computed as follows:

$$IoU(y, \hat{y}) = \sum_{k=1}^{c} \frac{|M(\hat{y}_k) \cap M(y_k)|}{|M(\hat{y}_k) \cup M(y_k)|} \tag{9}$$

Here, c represents the number of categories, and $M(y_k)$ and $M(\hat{y}_k)$ denote the sets of elements in the binary masks of y and $\hat{y}$ for class k, respectively.

In addition to mean IoU accuracy, we capture statistics, namely the mean and standard deviation (std) of the results. Ideally, we anticipate a high mean value and low std, signifying the model's consistent accuracy across different flight heights. A method is considered unsuccessful if it falls short of the performance of the trained model at the ground viewpoint (i.e., Ground-only).

C. Results on AirSim-Drone

The results of various approaches are presented in Table III. A closer examination of Ground-only reveals a noticeable trend of performance degradation from uav02 to uav10, indicating a 45.7% relative accuracy drop attributed to significant viewpoint differences.

In contrast, all previous SSL methods proved ineffective, performing even worse than Ground-only. Specifically, Pseudo-Labeling exhibited lower accuracy, with a mean of 0.457. While ClassMix, S4MC, UniMatch, CorrMatch, and PrevMatch were specifically designed to enhance semantic segmentation, they suffered substantial deterioration due to viewpoint differences. The failure of these SSL methods was anticipated as the presence of clear viewpoint differences in data samples prevented correct label predictions from unlabeled data.

On the other hand, our method demonstrated promising results for each sequence, showcasing only a 20.7% accuracy

TABLE III
THE MEAN IOU ACCURACY OF OUR APPROACH AND OTHER METHODS ON THE AIRSIM-DRONE DATASET

Methods	uav02	uav03	uav04	uav05	uav06	uav07	uav08	uav09	uav10	mean,	std
Ground-only	0.672	0.616	0.561	0.519	0.479	0.446	0.418	0.388	0.365	0.496	0.105
Pseudo-Labeling [55]	0.639	0.578	0.530	0.485	0.444	0.404	0.369	0.337	0.323	0.457	0.110
ClassMix [48]	0.209	0.197	0.190	0.183	0.180	0.178	0.176	0.174	0.173	0.184	**0.011**
S4MC [45]	0.403	0.330	0.277	0.237	0.217	0.205	0.200	0.189	0.190	0.250	0.074
UniMatch [21]	0.502	0.372	0.314	0.277	0.253	0.236	0.226	0.217	0.210	0.290	0.095
CorrMatch [56]	0.560	0.509	0.441	0.405	0.362	0.308	0.297	0.266	0.268	0.380	0.107
PrevMatch [57]	0.564	0.524	0.460	0.421	0.390	0.370	0.353	0.326	0.314	0.414	0.087
Ours	**0.680**	**0.650**	**0.625**	**0.609**	**0.594**	**0.578**	**0.564**	**0.551**	**0.539**	**0.599**	0.047

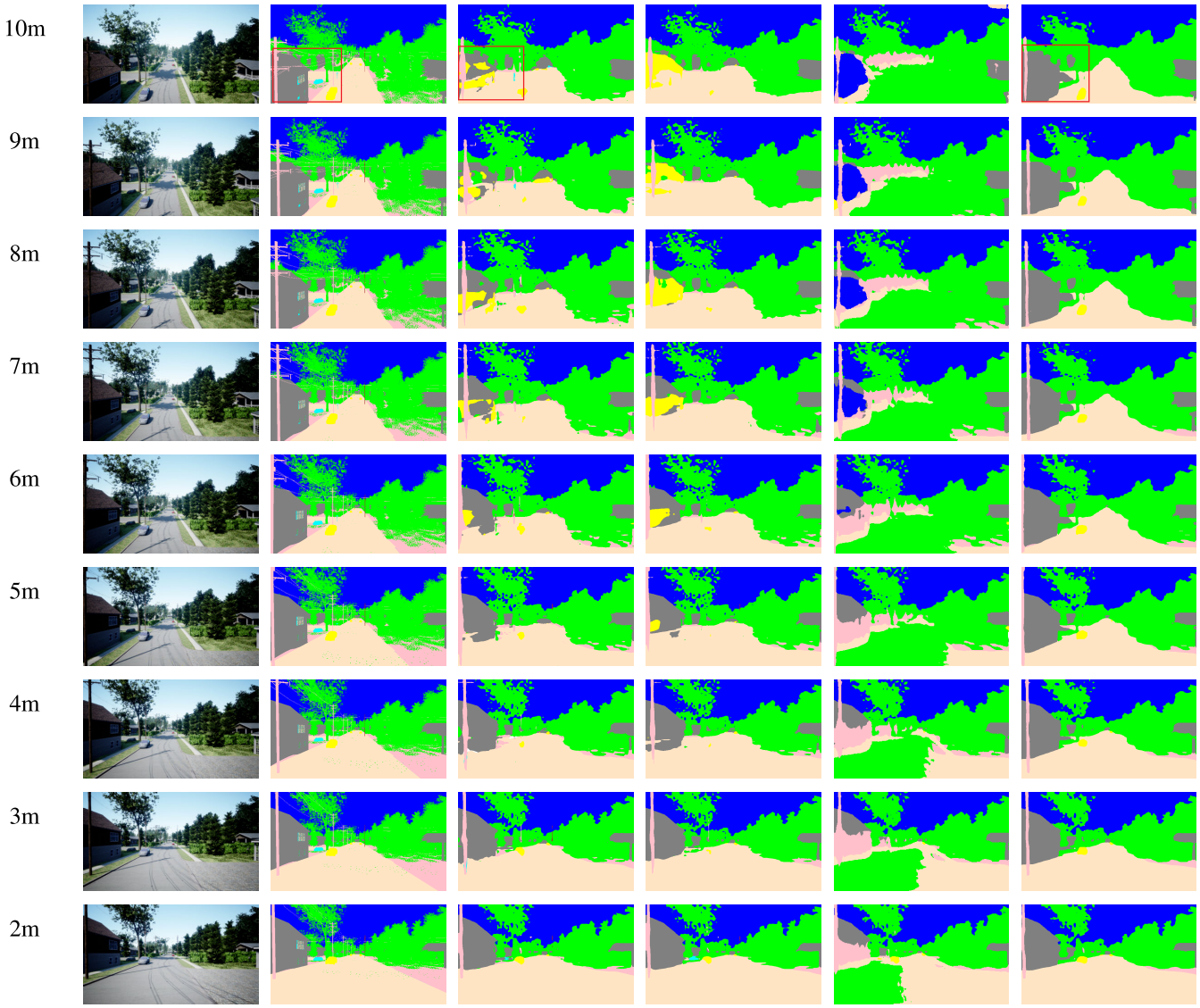


Fig. 6. Qualitative comparisons between our method and other approaches on the AirSim-Drone dataset. We show the results for different flight heights from 2 to 10 meters. We draw red boxes on the Ground truth and results of Ground-only and Ours at 10m for better visualization.

drop from uav02 to uav10. Compared to the original drop of 45.7%, the improvement achieved by our method is significant. We obtained a mean of 0.599 and a standard deviation of 0.047, indicating a 20.8% and 55.2% improvement from Ground-only's 0.496 and 0.105, respectively.

Fig. 6 displays the estimated semantic maps generated by different methods at varying flying heights. UniMatch exhibited inaccuracies and failures at higher flying heights. In the results of Ground-only and Pseudo-Labeling, the impact of viewpoint differences on semantic segmentation is evident, with false positives gradually increasing from 2 meters to 10 meters in the left bottom areas of images. In contrast, our method consistently produced correct results.

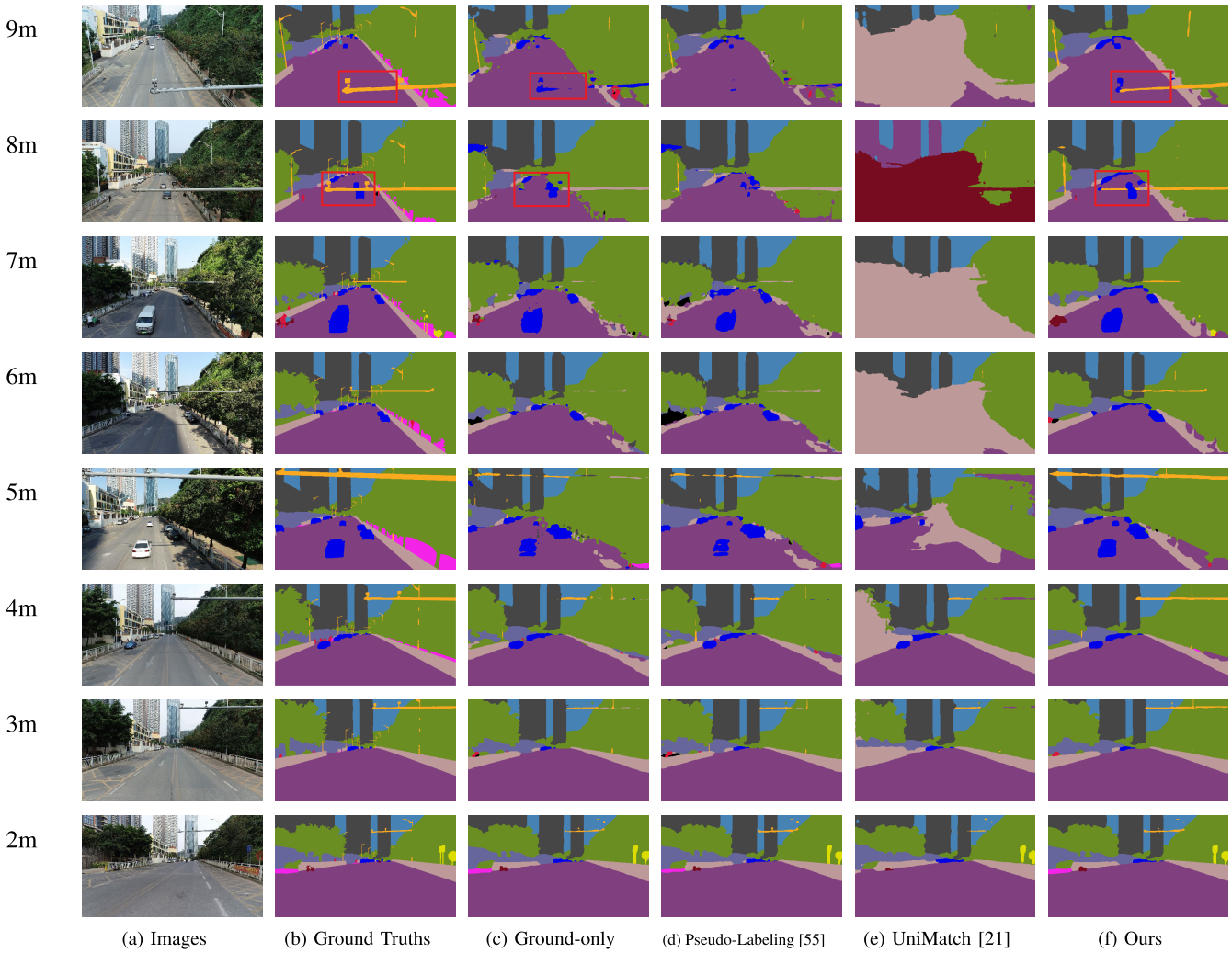


Fig. 7. Qualitative comparisons between our method and other approaches on the AIRs-Street dataset. We show the results for different flight heights from 2 to 9 meters. We draw red boxes on the Ground truth and results of Ground-only and Ours at 8m and 9m for better visualization.

TABLE IV
GENERALIZATION PERFORMANCE ON THE AIRSIM-DRONE TEST SET

Methods	Accuracy
Ground-only	0.417
Pseudo-Labeling [55]	0.400 (–4.1%)
ClassMix [48]	0.179 (–57.1%)
S4MC [45]	0.267 (–36.0%)
UniMatch [21]	0.302 (–27.6%)
CorrMatch [56]	0.342 (–18.0%)
PrevMatch [57]	0.388 (–7.0%)
Ours	**0.524** (**+25.7%**)

Table IV outlines the generalization performance on the test sequence. As evident, all SSL methods failed on the task, whereas our method excelled with a remarkable 25.7% performance improvement over Ground-only.

D. Results on AIRs-Street

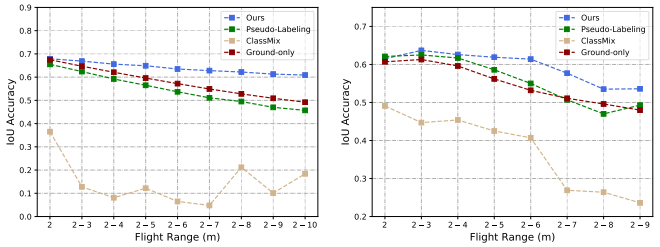
The real-world AIRs-Street dataset poses greater challenges compared to the synthesized AirSim-Drone dataset. It encompasses a more extensive array of semantic attributes, fewer labeled ground viewpoint images, and introduces dynamic elements like moving cars and non-rigid objects such as pedestrians. Additionally, varying lighting conditions due to sunlight intensity changes further contribute to the dataset's complexity. Consequently, drone perception performance is expected to degrade due to these inherent challenges, independent of viewpoint differences.

Table V presents the quantitative outcomes for various methods. The results illustrate a 39.2% accuracy decline for Ground-only from uav02 to uav09. ClassMix, S4MC, UniMatch, CorrMatch, and PrevMatch exhibited suboptimal results for each sequence compared to Ground-only, while Pseudo-labeling achieved the same mean accuracy as Ground-only. Notably, our method consistently outperformed Ground-only from uav02 to uav09. Similar to the AirSim-Drone results, the accuracy improvement tends to rise with ascending flying heights. For instance, there is a mere 6.0% boost for uav03 and a substantial 32.2% boost for uav09. Overall, our method achieved a mean of 0.561 and a standard deviation of 0.062, indicating a performance improvement of 16.9% and 66.7% from Ground-only, which had a mean of 0.480 and a standard deviation of 0.100, respectively.

Fig. 7 offers qualitative comparisons between our method and other approaches. The qualitative results align with those observed in AirSim-Drone. All methods exhibit gradual

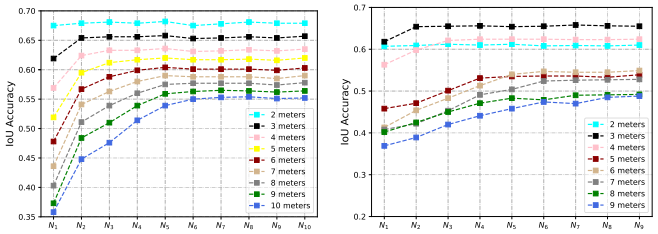
TABLE V
THE MEAN IOU ACCURACY OF OUR APPROACH AND OTHER METHODS ON THE AIRS-STREET DATASET

Methods	uav02	uav03	uav04	uav05	uav06	uav07	uav08	uav09	mean	std
Ground-only	0.607	0.618	0.563	0.458	0.413	0.408	0.402	0.369	0.480	0.100
Pseudo-Labeling [55]	**0.611**	0.623	0.546	0.467	0.406	0.421	0.386	0.381	0.480	0.100
ClassMix [48]	0.282	0.260	0.244	0.238	0.233	0.233	0.229	0.230	0.244	**0.019**
S4MC [45]	0.395	0.388	0.326	0.250	0.241	0.229	0.231	0.260	0.290	0.070
UniMatch [21]	0.417	0.376	0.323	0.246	0.223	0.223	0.201	0.212	0.278	0.083
CorrMatch [56]	0.515	0.466	0.365	0.323	0.261	0.254	0.292	0.256	0.342	0.100
PrevMatch [57]	0.514	0.497	0.454	0.439	0.355	0.378	0.399	0.352	0.424	0.062
Ours	0.610	**0.655**	**0.624**	**0.539**	**0.549**	**0.528**	**0.492**	**0.488**	**0.561**	0.062



(a) Results on the AirSim-Drone. (b) Results on the AIRs-Street.

Fig. 8. Results for different methods while varying the maximum flying heights. The x-axis denotes the flight heights and the y-axis shows the mean accuracy on the corresponding range.



(a) Results on the AirSim-Drone. (b) Results on the AIRs-Street.

Fig. 9. Visualization of the application of our self-distillation method on each flight height. N_i denotes the trained model for the maximum flight height h_i.

deterioration from a flying height of 2 meters to 9 meters. Notably, false positives tend to increase for the comparison methods, whereas our method consistently predicts pixel attributes more accurately, particularly for cars.

E. Adaptability to Various Flight Heights

We establish distinct height ranges, commencing from h_2 up to the maximum flight height h_i, where h_i varies from 2, 3, . . . , 10 for AirSim-Drone and 2, 3, . . . , 9 for AIRs-Street, respectively. Subsequently, we assess the performance of our methods and baselines within these designated ranges.

Fig. 8 presents the outcomes on the two datasets. Notably, ClassMix exhibits erratic behavior on AirSim-Drone and consistently performs poorly on both datasets. Our method consistently outperforms the other two approaches across various settings, except at the maximum flight height of 2 meters on AIRs-Street.

We further visualize the accuracy progression within our progressive learning framework for each flight height in Fig. 9, denoting N_i as the trained model applied to the maximum flight height h_i (e.g., N_5 is learned from h_1 to h_5). As depicted in Fig. 9 (a), our method exhibits substantial improvement for significant viewpoint differences (e.g., uav10) while maintaining accuracy for smaller viewpoint differences (e.g., uav02) on AirSim-Drone. This trend is similarly observed in Fig. 9 (b).

Additionally, due to the absence of supervision for flying viewpoints, we observe a gradual slowdown in accuracy improvement during the application of our self-distillation. As the viewpoint discrepancy gradually increases, addressing more challenging cases necessitates additional supervision—a facet we earmark for future exploration in our work.

F. Ablation Studies

We conduct several ablation studies to comprehensively analyze each component of our progressive SSL framework.

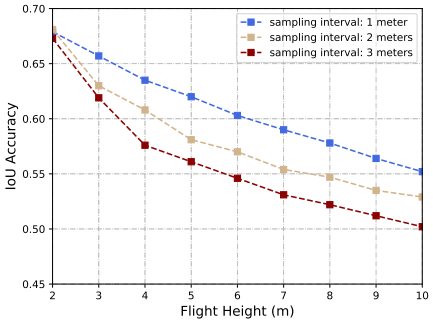
TABLE VI
SPLIT OF AIRSIM-DRONE WITH DIFFERENT SAMPLING INTERVALS

Intervals	Labeled view	Unlabeled view
1 meter	$h1$	$h2, h3, h4, h5, h6, h7, h8, h9, h10$
2 meters	$h1$	$h3, h5, h7, h9$
3 meters	$h1$	$h4, h7, h10$

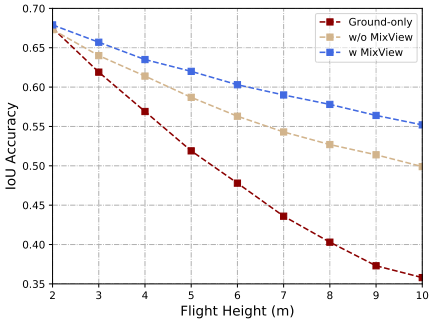
Specifically, we explore 1) the impact of different sampling intervals of viewpoints, 2) the outcomes with and without utilizing MixView, and 3) the results with and without utilizing nearest neighbor pseudo-labeling. All ablation studies are executed on AirSim-Drone, and detailed information is provided as follows:

1) Analysis With Different Sampling Intervals: We assess three settings of the sampling interval—1 meter, 2 meters, and 3 meters. Table VI provides detailed information on data sampling, and the experimental results are illustrated in Fig. 10 (a). Here, blue, yellow, and red colors represent results for sampling intervals of 1, 2, and 3 meters, respectively. The findings demonstrate that denser sampling of viewpoints contributes to enhanced model performance. Specifically, we achieve 13.8%, 18.3%, and 23.8% mean accuracy boosts from the Ground-only model, respectively.

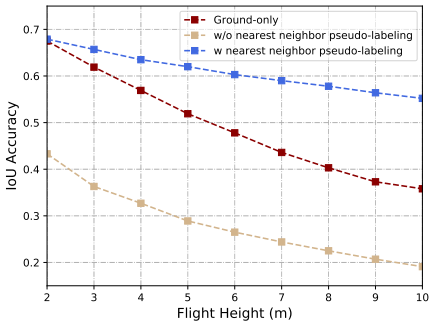
2) Analysis With and Without Using MixView: Within our progressive learning framework, we introduce MixView, a method to amalgamate images from different viewpoints for augmented training samples. Experimental results illustrate that MixView contributes to mitigating viewpoint differences. In Fig. 10 (b), where red, yellow, and blue colors represent results for the ground-only model, our method without MixView (w/o), and our method with MixView (w), respectively. Notably, even without employing MixView, the progressive distillation integrated with nearest neighbor pseudo-labeling outperforms the Ground-only model by



(a) Results for different sampling intervals.



(b) Results with and without using the MixView.



(c) Results with and without using the nearest neighbor pseudo-labeling.

Fig. 10. Results of three ablation studies. We show the effect of different sampling intervals of viewpoints in (a), and performance with and without using MixView in (b), and results with and without using the nearest neighbor pseudo-labeling in (c), respectively.

16.5%. Furthermore, MixView provides an additional performance boost of 7.3%.

3) Analysis With and Without Using Nearest Neighbor Pseudo-Labeling: In our original setting, when training a model at h_i, we predict pseudo-labels using the trained model at h_{i-1} from data captured at h_i. N_i is then trained with pseudo-labeled samples of $h_1, \ldots, h_i$ and unlabeled samples of h_i. Skipping this step removes the effect of nearest viewpoint pseudo-labeling, and the model N_i is trained with pseudo-labeled samples of $h_1, \ldots, h_{i-1}$ and unlabeled samples of h_i. The results are shown in Fig. 10 (c), where red, yellow, and blue colors denote results for the ground-only model, our method without using nearest neighbor pseudo-labeling (w/o), and our method using nearest neighbor pseudo-labeling (w), respectively. Without utilizing nearest neighbor pseudo-labeling, the mean accuracy degrades up to 47.8% from the Ground-only model.

G. Summary

Considering a model trained at the ground viewpoint as a baseline (i.e., a Ground-only model), we show through experiments on both a synthesized dataset and a real-world dataset that:

- The performance of a Ground-only model gradually deteriorates from lower flying height to higher flying height as viewpoint discrepancy gradually increases.
- The viewpoint difference leads the previous SSL-based methods to malfunction. In our experiments, all baseline methods failed on the task, as their performance was even worse than that of the Ground-only model.
- Our method shows a substantial performance boost compared to the Ground-only model. The mean relative accuracy improvement is 25.7% and 16.9% for the fixed AirSim-Drone and challenging AIRs-Street, respectively.
- Most importantly, we find that the performance improvement is significant for large viewpoint differences. We obtained 47.7% and 32.2% relative accuracy improvement for the flight height of 10 meters, i.e., uav10 of AirSim-Drone, and the flight height of 9 meters, i.e., uav09 of AIRs-Street, respectively.
- Our method demonstrates good adaptability to various height ranges. It outperformed other SSL-based approaches in different height ranges.

VI. Conclusion

In this paper, we have explored drone perception at low altitudes without incurring additional flight data labeling costs. The key requirement for effective drone perception, as argued, lies in its ability to perceive from diverse flight heights, posing a fundamental challenge due to significant viewpoint differences. Given labeled ground viewpoint images and unlabeled images from various flying perspectives, we formulated the problem as a semi-supervised learning task and proposed a progressive learning framework that enables gradual adaptation from the ground viewpoint to the maximum flying height. To validate the approach, a synthesized dataset spanning flight heights from 1 to 10 meters and a real-world dataset covering heights from 1 to 9 meters were created. The results on both datasets demonstrate promising results across various flight heights. However, due to the lack of annotated data at higher altitudes, our method tends to exhibit reduced accuracy in such scenarios and is currently applicable only to lower AGL heights. In the future, several promising directions can be pursued, including generating and utilizing additional annotated datasets for flying viewpoints and addressing drone perception challenges under diverse environmental conditions, such as varying weather scenarios. As a preliminary exploration, we hope to inspire further research within the community.

References

[1] A. V. Savkin and H. Huang, "Navigation of a UAV network for optimal surveillance of a group of ground targets moving along a road," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9281–9285, Jul. 2022.

[2] J. A. Besada et al., "Drone mission definition and implementation for automated infrastructure inspection using airborne sensors," *Sensors*, vol. 18, no. 4, p. 1170, Apr. 2018.

[3] F. Betti Sorbelli, C. M. Pinotti, and G. Rigoni, "On the evaluation of a drone-based delivery system on a mixed Euclidean-manhattan grid," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1276–1287, Jan. 2023.

[4] A. Bajaj, B. Philips, E. Lyons, D. Westbrook, and M. Zink, "Determining and communicating weather risk in the new drone economy," in *Proc. IEEE 92nd Veh. Technol. Conf. (VTC-Fall)*, Nov. 2020, pp. 1–6.

[5] F. Ho, R. Geraldes et al., "Decentralized multi-agent pathfinding for uav traffic management," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 997–1008, Feb. 2020.

[6] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang, "PASS: Panoramic annular semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4171–4185, Oct. 2020.

[7] K. Yang, X. Hu, Y. Fang, K. Wang, and R. Stiefelhagen, "Omnisupervised omnidirectional semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1184–1199, Feb. 2022.

[8] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[10] F. Ceola, E. Maiettini, G. Pasquale, G. Meanti, L. Rosasco, and L. Natale, "Learn fast, segment well: Fast object segmentation learning on the iCub robot," *IEEE Trans. Robot.*, vol. 38, no. 5, pp. 3154–3172, Oct. 2022.

[11] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1043–1051.

[12] J. Hu et al., "Boosting light-weight depth estimation via knowledge distillation," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.*, 2023, pp. 27–39.

[13] J. Hu, C. Fan, L. Zhou, Q. Gao, H. Liu, and T. L. Lam, "Lifelong-MonoDepth: Lifelong learning for multidomain monocular metric depth estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 796–806, Jan. 2025.

[14] P. Zhu et al., "VisDrone-DET2018: The vision meets drone object detection in image challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Jan. 2019, pp. 437–468.

[15] P. Zhu et al., "Visdrone-vid2019: The vision meets drone object detection in video challenge results," in *Proc. Int. Conf. Comput. Vis. Workshop (ICCV)*, 2019, pp. 227–235.

[16] Y. Cao et al., "VisDrone-DET2021: The vision meets drone object detection challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2847–2854.

[17] X. Guo, J. Hu, J. Chen, F. Deng, and T. L. Lam, "Semantic histogram based graph matching for real-time multi-robot global localization in large scale environment," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8349–8356, Oct. 2021.

[18] A. Gawel, C. D. Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1687–1694, Jul. 2018.

[19] H. Zhang, M. Cissé, Y. Dauphin, and D. López-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2017, pp. 1–13.

[20] S. Liao, X. Jiang, and Z. Ge, "Weakly supervised multilayer perceptron for industrial fault classification with inaccurate and incomplete labels," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 1192–1201, Apr. 2022.

[21] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7236–7246.

[22] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 4104–4113.

[23] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4105–4114.

[24] C. Fan, J. Hu, and J. Huang, "Few-shot multi-agent perception with ranking-based feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11810–11823, Oct. 2023.

[25] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-fidelity visual and physical simulation for autonomous vehicles," *Field Service Robot.*, vol. 5, pp. 621–635, Nov. 2017.

[26] S. Jung, S. Hwang, H. Shin, and D. H. Shim, "Perception, guidance, and navigation for indoor autonomous drone racing using deep learning," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2539–2544, Jul. 2018.

[27] R. P. Padhy, S. Verma, S. Ahmad, S. K. Choudhury, and P. K. Sa, "Deep neural network for autonomous UAV navigation in indoor corridor environments," *Proc. Comput. Sci.*, vol. 133, pp. 643–650, Jul. 2018.

[28] S. Vaddi, C. Kumar, and A. Jannesari, "Efficient object detection model for real-time UAV applications," 2019, *arXiv:1906.00786*.

[29] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: A survey," *Image Vis. Comput.*, vol. 104, Dec. 2020, Art. no. 104046.

[30] P. Zhu et al., "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022.

[31] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020.

[32] S. Girisha, M. M. M. Pai, U. Verma, and R. M. Pai, "Semantic segmentation of UAV aerial videos using convolutional neural networks," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, Jun. 2019, pp. 21–27.

[33] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6876–6883.

[34] K. Yang, D. Yang, J. Zhang, H. Wang, P. Sun, and L. Song, "What2comm: Towards communication-efficient collaborative perception via feature decoupling," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7686–7695.

[35] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake LiDAR in 3D detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9243–9252.

[36] C. Fan, J. Hu, and J. Huang, "Few-shot multi-agent perception," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1712–1720.

[37] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," 2021, *arXiv:2103.00550*.

[38] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2019.

[39] A. Odena, "Semi-supervised learning with generative adversarial networks," 2016, *arXiv:1606.01583*.

[40] Q. H. Cap, H. Uga, S. Kagiwada, and H. Iyatomi, "LeafGAN: An effective data augmentation method for practical plant disease diagnosis," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 1258–1267, Apr. 2022.

[41] D. P. Kingma and S. Mohamed, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3581–3589.

[42] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–12.

[43] S. Lu and Y. Wen, "Semi-supervised condition monitoring and visualization of fused magnesium furnace," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 3471–3482, Oct. 2022.

[44] C. Fan, J. Hu, and J. Huang, "Private semi-supervised federated learning," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 2009–2015.

[45] M. Kimhi, S. Kimhi, E. Zheltonozhskii, O. Litany, and C. Baskin, "Semi-supervised semantic segmentation via marginal contextual information," in *Proc. Trans. Mach. Learn. Res.*, Jan. 2023, pp. 1–24.

[46] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.

[47] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[48] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1368–1377.
[49] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[50] J. Hu, C. Fan, M. Ozay, H. Jiang, and T. L. Lam, "Dense depth distillation with out-of-distribution simulated images," *Knowl.-Based Syst.*, vol. 284, Jan. 2024, Art. no. 111312.
[51] J. Hu, C. Fan, X. Guo, L. Zhou, and T. L. Lam, "Self-supervised single-line LiDAR depth completion," *IEEE Robot. Autom. Lett.*, vol. 8, no. 11, pp. 7320–7327, Nov. 2023.
[52] C. Wang, D. Chen, J.-P. Mei, Y. Zhang, Y. Feng, and C. Chen, "SemCKD: Semantic calibration for cross-layer knowledge distillation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 6305–6319, Jun. 2023.
[53] C. Chen, G. Yao, L. Liu, Q. Pei, H. Song, and S. Dustdar, "A cooperative vehicle-infrastructure system for road hazards detection with edge intelligence," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 5186–5198, May 2023.
[54] K. J. Liang et al., "MixKD: Towards efficient distillation of large-scale language models," in *Proc. Int. Conf. Learn. Represent.*, Nov. 2020, pp. 1–16.
[55] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.
[56] B. Sun, Y. Yang, L. Zhang, M.-M. Cheng, and Q. Hou, "CorrMatch: Label propagation via correlation matching for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3097–3107.
[57] W. Shin, H. Joon Park, J. Sob Kim, and S. Won Han, "Revisiting and maximizing temporal knowledge in semi-supervised semantic segmentation," 2024, *arXiv:2405.20610*.
[58] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[59] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[60] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.
[61] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NIPS*, Dec. 2019, pp. 8024–8035.

Junjie Hu (Member, IEEE) received the M.S. and Ph.D. degrees from the Graduate School of Information Science, Tohoku University, Sendai, Japan, in 2017 and 2020, respectively. He is currently an Assistant Professor with the School of Artificial Intelligence, the Chinese University of Hong Kong, Shenzhen, and a Research Scientist with Shenzhen Institute of Artificial Intelligence and Robotics for Society. He has published more than 40 research papers in top-tier international journals and conference proceedings in AI and robotics, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON ROBOTICS, CVPR, ICCV, AAAI, IJCAI, IEEE ROBOTICS AND AUTOMATION LETTERS, ICRA, and IROS. His research interests include machine learning, computer vision, and robotics. He serves as a reviewer for *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, CVPR, ICCV, ECCV, ICRA, IROS, IEEE ROBOTICS AND AUTOMATION LETTERS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, and *Journal of Field Robotics*.

Chenyou Fan (Member, IEEE) received the B.S. degree in computer science from Nanjing University, China, in 2011, and the M.S. and Ph.D. degrees from Indiana University, USA, in 2014 and 2019, respectively. He was a Research Scientist with Shenzhen Institute of Artificial Intelligence and Robotics for Society. He is currently an Associate Professor with South China Normal University. His research interests include machine learning and computer vision. He served on the Program Committee for CVPR, NeurIPS, ACM MM, and top AI journals for more than 20 times.

Mete Ozay (Member, IEEE) received the B.Sc., M.Sc., Ph.D. degrees in mathematical physics, information systems, and computer engineering and science from METU, Turkey. He has been a Visiting Ph.D. Student and a fellow at Princeton University, USA, a Research Fellow at the University of Birmingham, U.K., an Assistant Professor at Tohoku University, Japan, and a Research Scientist and Engineer at Samsung Research, U.K. His current research interests include pure and applied mathematics, theoretical computer science and neuroscience.

Hua Feng received the B.S. degree from Dongguan University of Technology, Dongguan, China, in 2017, and the master's degree from Wuyi University, Jiangmen, China, in 2022. From 2021 to 2022, he was a Research Intern at Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS). He is currently an Algorithm Engineer at Guangzhou iMapCloud Intelligent Technology Company Ltd. His research interests include robotics, machine learning, visual SLAM, and UAV navigation.

Yuan Gao (Member, IEEE) received the Ph.D. degree in the area of machine learning and robotics from Uppsala University, Sweden, under the supervision of Prof. Ginevra Castellano and Prof. Danica Kragic. He is currently a Research Scientist at Shenzhen Institute of Artificial intelligence and Robotics for Society and a part-time Assistant Professor at Chinese University of Hong Kong, Shenzhen. His research interests include machine behavior analysis, reinforcement learning, robotics, and general machine learning.

Tin Lun Lam (Senior Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2010. He is currently an Assistant Professor with The Chinese University of Hong Kong, Shenzhen, China, and the Director of the Center for the Intelligent Robots, Shenzhen Institute of Artificial Intelligence and Robotics for Society. He has published two monographs and more than 100 research papers in top-tier international journals and conference proceedings in robotics. His research interests include multi-robot systems, field robotics, and collaborative robotics. He received the IEEE/ASME T-MECH Best Paper Award in 2011 and the IEEE/RSJ IROS Best Paper Award on Robot Mechanisms and Design in 2020.