

UAVid++: Higher-Quality Labels and Expanded Semantic Taxonomy for Aerial Semantic Segmentation

Vivian Chiciudean, Sergiu Nedevschi, and Florin Oniga

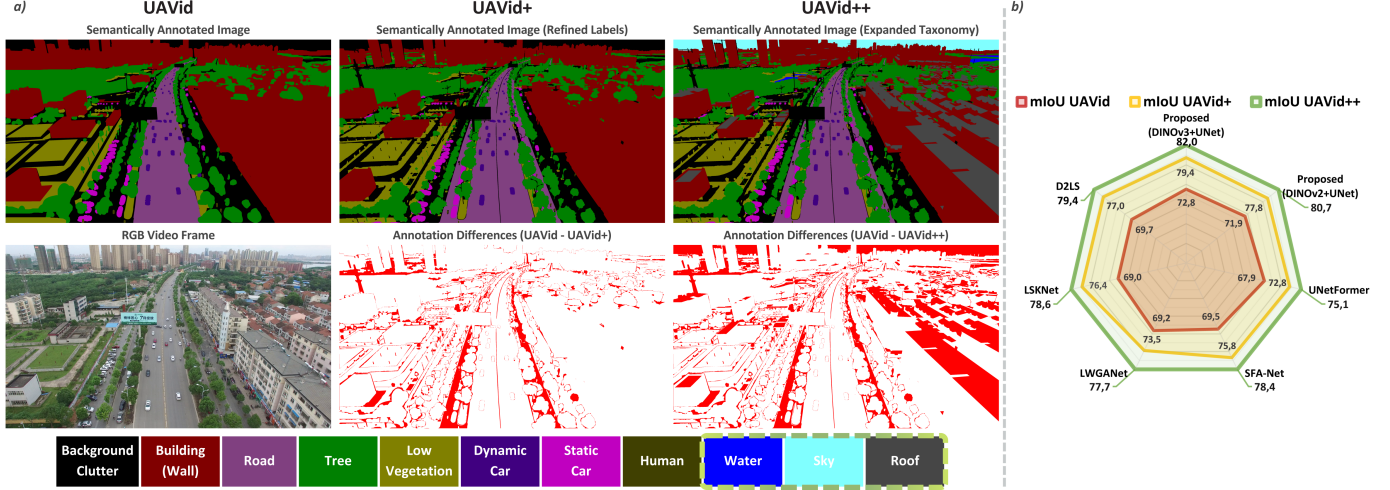


Fig. 1: **Overview of the UAVid++ dataset and evaluation results.** a) Comparison of UAVid, UAVid+ (corrected annotations), and UAVid++ (corrected annotations and expanded semantic taxonomy), including pixel-level annotation differences; b) Semantic segmentation performance of state-of-the-art methods and the proposed approach across the three dataset variants.

Abstract—High-quality annotated datasets are critical for semantic segmentation, particularly in aerial imagery, where the number of manually annotated frames is limited and precise labeling is often missing. To address this, we introduce UAVid++, a semantically refined and expanded version of the UAVid dataset for aerial semantic segmentation. The proposed dataset provides higher-quality annotations through manual corrections addressing labeling errors across the entire original UAVid dataset. In addition, we expand the taxonomy with three finely annotated semantic classes, Water, Sky, and Roof, selected to enrich the semantic space and separate visually distinct objects previously grouped together under the same label. We evaluated state-of-the-art aerial segmentation methods on UAVid, UAVid+ (with corrections, without expanded taxonomy), and the full UAVid++ dataset. Corrected annotations improved both training and evaluation performance by up to 7.4% mIoU, and adding the three classes further increased performance by up to 4.1%, demonstrating the combined impact of refined labels and the expanded semantic taxonomy. UAVid++ enables more discriminative evaluation of state-of-the-art methods and improves their relative ranking. We adapt frozen DINO-pretrained ViT backbones using trainable task-specific heads for UAV semantic segmentation. This strategy achieves strong segmentation performance and robust generalization while quantifying the gains from improved labels relative to pretrained backbone capacity. Relative to the best-performing state-of-the-art method, our configurations achieve a 2.7% mIoU improvement on UAVid++, up to 12.4% mIoU gains on out-of-distribution UAV datasets, and a 38.85% mIoU gain on a cross-domain dataset. Both the dataset and implementation will be pub-

licly available at <https://vivichiciudean.github.io/uavidplusplus/>.

Index Terms—UAV semantic dataset, annotation refinement, expanded taxonomy, pretrained ViT, generalization, UAVid++

I. INTRODUCTION

PERCEPTION in low-altitude UAV imagery has rapidly advanced within the field of computer vision, enabling applications such as autonomous navigation [1]–[3], urban infrastructure monitoring [4], [5], disaster assessment [6]–[9], emergency response [10]–[13], and environmental surveillance [14]–[18]. Within this domain, semantic segmentation [19] plays a central role in achieving fine-grained understanding and reliable perception.

Despite advances in supervised aerial segmentation networks [20]–[24], performance remains limited by the availability and quality of annotated datasets, as even minor label noise can substantially degrade segmentation accuracy [25]–[27]. While large-scale datasets can provide coarse annotations [28], obtained at relatively low cost and effort, these labels often lack the spatial precision needed to provide fine-grained guidance for precise supervision [29]. As a result, state-of-the-art models primarily depend on reliable annotations, with coarse labels offering only limited auxiliary benefit [29]–[31].

Producing high-quality annotations in UAV imagery is especially challenging. Urban scenes captured from low altitudes exhibit complex spatial arrangements, oblique viewpoints with lines of sight spanning hundreds of meters, rapid camera motion, and significant semantic variability. Frequent occlusions and small dynamic objects, such as vehicles and pedestrians,

The authors are with the Department of Computer Science, Technical University of Cluj-Napoca, Romania (e-mail: Name.Surname@cs.utcluj.ro)

This work was supported by the “Romanian Hub for Artificial Intelligence-HRIA” project, Smart Growth, Digitization and Financial Instruments Program, MySMIS no. 351416, Ministry of Investments and European Projects, Romanian Government.

further complicate precise labeling. These factors make accurate pixel-level annotation technically challenging yet critical for reliable segmentation.

Unlike the automotive domain, which benefits from large-scale datasets [32]–[34] with tens of thousands of images and over a hundred semantic classes, publicly available UAV datasets remain limited in size and semantic coverage [35]–[41]. Among them, UAVid [39] serves as a valuable benchmark for urban aerial perception, offering high-resolution oblique-view imagery with densely crowded scenes and explicitly annotated dynamic classes such as vehicles and pedestrians. These characteristics make it particularly suitable for evaluating fine-grained scene understanding in complex urban environments. However, a careful inspection reveals multiple annotation errors and substantial semantic intra-class variance, which can compromise benchmarking and lead to inaccurate performance estimates for state-of-the-art models [42].

To address these limitations, we introduce UAVid++, a semantically refined and expanded version of the UAVid dataset for aerial semantic segmentation. UAVid++ provides higher-quality annotations through manual corrections that address labeling errors and inconsistencies in the original UAVid dataset, along with extensive contour refinement to ensure precise pixel-level evaluation. Furthermore, we expand the semantic taxonomy by adding three classes (*Water*, *Sky*, *Roof*), selected to enrich the semantic space and separate visually distinct objects that were previously grouped under the same label. Together, these enhancements increase annotation fidelity, reduce semantic ambiguity, and strengthen the dataset’s suitability for evaluating generalization in aerial segmentation.

We evaluated state-of-the-art aerial segmentation methods [20]–[24] on UAVid, UAVid+ (corrected annotations, without expanded taxonomy), and the full UAVid++ dataset. Performance on the original UAVid was limited by coarse and inaccurate labels. UAVid+ improved both training and evaluation through refined annotations, while UAVid++ further improved segmentation performance by introducing three semantic classes, highlighting the impact of enhanced labels. Moreover, UAVid++ enables a more discriminative and informative evaluation of state-of-the-art architectures, improving their previous relative ranking.

We employ an adaptation strategy that combines frozen DINO-pretrained ViT backbones [43], [44] with trainable task-specific heads, tailored for aerial scenarios with limited labeled data. This design preserves the representational power of Vision Transformers [45], which generalize well from large-scale pretraining [43], [44] and achieve state-of-the-art results across diverse tasks [46], [47]. Inspired by recent advances [47]–[51], we explore three head variants to improve segmentation accuracy. The strategy surpasses prior state-of-the-art on UAVid++ and demonstrates stronger generalization across multiple aerial scenarios. Furthermore, the adaptation strategy provides an analytical tool for quantifying the gains from improved labels relative to pretrained backbone capacity.

In summary, we introduce UAVid++, a manually corrected and expanded UAV semantic segmentation dataset. The dataset is accompanied by a documented annotation protocol, which

defines the class-boundary rules and review procedure used during annotation. We provide a comprehensive evaluation of five state-of-the-art methods across the three dataset variants and observe gains of up to 9.7% mIoU on UAVid++ over UAVid. The corrected labels yield a more discriminative benchmark and revise the previous relative ranking of these methods, while the expanded semantic taxonomy supports more detailed scene understanding. In addition, we employ an adaptation strategy that combines frozen DINO-pretrained ViT backbones with trainable task-specific heads. This strategy delivers strong segmentation performance and robust generalization, including cross-domain settings, and serves as an analytical tool that quantifies the impact of annotation quality relative to backbone capacity, confirming that label fidelity outweighs backbone capacity under limited aerial training data.

Our main contributions are:

- UAVid++, an annotated dataset with consistent labeling, refined contours, and an expanded semantic taxonomy including three finely annotated classes to enhance semantic separability and reduce unlabeled regions.
- A comprehensive evaluation of the proposed dataset on state-of-the-art segmentation methods, demonstrating up to 7.4% mIoU gain from annotation corrections and 4.1% from the expanded taxonomy, relative to UAVid.
- A more informative benchmark provided by UAVid++, enabling a wider distribution of segmentation scores and a more discriminative evaluation of aerial segmentation methods, thereby improving their relative ranking.
- An adaptation strategy that combines frozen DINO-pretrained ViT backbones with trainable task-specific heads, outperforming state-of-the-art methods by 2.7% mIoU on UAVid++, by up to 12.4% mIoU on out-of-distribution UAV datasets, and by 38.85% mIoU on a cross-domain dataset.
- A quantification of the impact of annotation quality relative to backbone capacity, showing that improved annotations contribute substantially more to performance than using a larger frozen pretrained backbone.

Both the dataset and implementation will be publicly released to ensure reproducibility and facilitate further research in aerial semantic segmentation.

II. RELATED WORK

A. UAV Semantic Segmentation Datasets

While the automotive domain benefits from large-scale benchmarks such as Cityscapes [28], Mapillary [33], and Waymo [34], which include up to 100,000 images covering more than 100 semantic classes, UAV datasets remain constrained in size and semantic scope. Existing low-altitude aerial datasets typically contain only a few hundred annotated images spanning a limited number of classes, with considerable variation in real-world coverage, viewpoints, and environmental conditions. A summary of representative datasets is provided in Table I. These datasets [35]–[41] cover diverse environments, including urban and rural settings, and differ in image resolution, camera orientation, flight altitude, and

TABLE I

Overview of publicly available datasets for semantic segmentation. Most aerial benchmarks provide coarse annotations and limited coverage of dynamic object classes.

Dataset	Domain	# Images	Resolution	# Classes	Scenes	Altitude (m)	Camera Orientation	Annotation Quality	Dynamic Entities
Cityscapes [28]	Automotive	5K	2K	30	Urban	–	–	Precise	Dense
Mapillary [33]		25K	2K	124	Urban	–	–		
Waymo [34]		100K	2K	28	Urban	–	–		
Aeroscapes [35]	Aerial	3269	720p	12	Urban	5-50	Oblique & Nadir	Coarse	Rare
WildUAV [36]		1539	4K	9	Rural	25-60	Oblique & Nadir		Rare
Ruralscapes [37]		1047	4K	12	Rural	–	Oblique		Rare
Dronescape [38]		556	4K	8	Urban & Rural	–	Oblique		Rare
UAVid [39]		420	4K	8	Urban & Rural	50	Oblique		Dense
UDD [40]		141	4K	6	Urban	60-100	Oblique & Nadir		Dense, but Unlabeled Humans
VDD [41]		400	4K	7	Urban	50-120	Oblique & Nadir		
IDD [41]		(270 UAVid 141 UDD 400 VDD)	4K	7	Urban & Rural	50-120	Oblique & Nadir		

semantic taxonomies. Despite this diversity, most UAV segmentation datasets suffer from coarse polygonal annotations and contain few labeled dynamic entities.

Among low-altitude aerial datasets, UAVid [39] is notable for its oblique-view images that densely capture hundreds of dynamic entities, such as vehicles and pedestrians, within a single frame. It comprises 42 video sequences, each containing 901 frames at 4K resolution, acquired at approximately 50 meters altitude across two countries. Each sequence includes annotations for 10 frames (one every 100), resulting in 420 labeled images with eight semantic classes. UAVid balances scene coverage, object diversity, and dynamic content, making it a valuable resource for urban aerial perception. However, UAVid, like most of the aforementioned aerial datasets, remains limited by shallow semantic taxonomies and inconsistent annotations. Reduced taxonomy depth amplifies visual ambiguity within classes and exacerbates class imbalance, making it difficult for models to differentiate closely related categories. Labeling errors and coarse contours further undermine reliable training and evaluation [25]–[27], [29], [30]. These problems are amplified by small dataset sizes [52], [53], where even minor label noise can alter benchmark rankings and mask model improvements [42].

The IDD dataset [41] constitutes one of the most heterogeneous benchmarks in terms of scene composition, as it aggregates imagery from three aerial datasets and consolidates them under a unified annotation framework. Specifically, it integrates 270 images from UAVid [39] (training and validation sets), 141 images from UDD [40], and 400 images from VDD [41], aiming to broaden environmental diversity while maintaining a common semantic labeling scheme. Its primary objective is to assemble a larger dataset using a coarse semantic taxonomy that prioritizes broad pixel coverage over detailed object delineation. The IDD taxonomy modifies that of UAVid by merging several semantically distinct categories, discarding classes such as pedestrians, and introducing additional coarse annotations. This label aggregation introduces several limitations. First, the annotations inherit inaccura-

cies from the source datasets, including missing instances, imprecise boundaries, and inconsistent labeling. Second, the annotation process frequently relies on large polygonal regions that overlap adjacent structures and leave substantial portions of the scene unlabeled or inaccurately delineated. These factors reduce annotation fidelity and compromise fine-grained evaluation, particularly for small dynamic objects that are prevalent in aerial imagery.

We select UAVid as our starting point because it provides densely crowded urban scenes with annotated humans. In contrast, IDD lacks labeled humans, discards certain UAVid classes through merging with UDD and VDD, and uses coarse polygonal approximations for the new classes. Nevertheless, limitations in UAVid’s existing annotations highlight the need for improved labeling with richer semantic taxonomies.

B. Architectures for UAV Semantic Segmentation

Early work in aerial semantic segmentation primarily relied on Fully Convolutional Networks (FCNs) [19], which efficiently learn local spatial patterns from high-resolution UAV imagery [35], [39], [54], [55]. However, FCNs struggle to capture long-range dependencies and global context, motivating a shift toward architectures that integrate both local and global information. Hybrid methods combining Vision Transformers (ViTs) [45] with convolutional backbones were developed to address this. UNetFormer [20] pairs a lightweight ResNet encoder [56] with a ViT-based decoder. AerialFormer [57] adopts a multi-resolution transformer tailored for aerial imagery. GLOTS [58] pairs a masked-image-modeling encoder with a multiscale global-local decoder and progressive upsampling to recover fine-grained detail. BiTSRS [59] adopts a bi-decoder design over a Swin Transformer [60] encoder to jointly capture global and local representations. SCSM [61] decomposes the scene into global representations and object distributions, using local-global semantic masks to suppress background interference and intra-class variance. LSKNet [23] dynamically adjusts large spatial receptive fields with selective kernels to capture long-range context. SFA-Net [22] improves

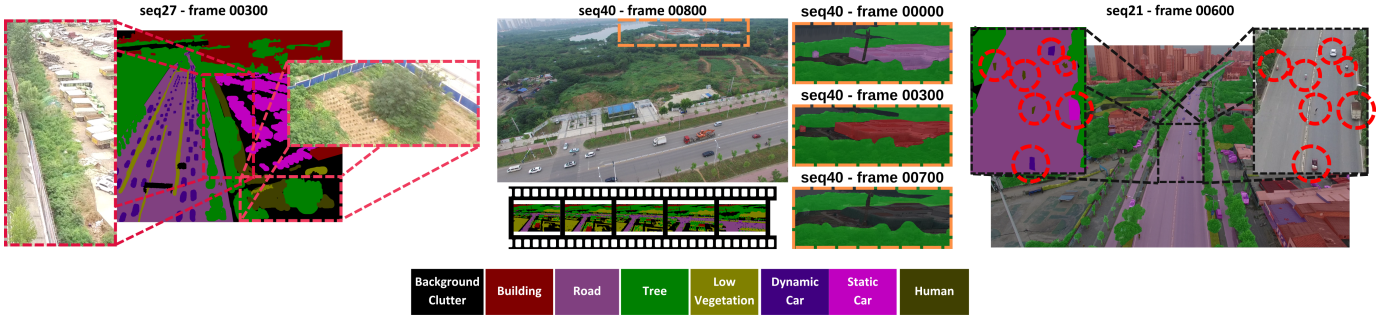


Fig. 2: **Annotation issues in UAVid:** Red rectangles highlight mislabeled components, orange rectangles indicate inconsistent labeling of the same object across frames, and red circles highlight objects with semantic contours misaligned with real-world edges.

boundary delineation via semantic feature adjustment modules. Building on these foundations, later work focused on enhancing attention mechanisms and efficiency. SegFormer [62] variants [63] have been applied to UAV imagery to balance accuracy and computational cost. LWGANet [21] employs lightweight group attention for efficient feature aggregation. More recent Transformer-based designs [24] incorporate adaptive semantic embeddings and class-aware dictionary learning, further strengthening feature representation and enabling fine-grained segmentation.

Despite these advances, most methods generalize poorly to out-of-distribution data, with performance degrading on camera perspectives or flight altitudes not seen during training. ViTs enable backbones pretrained on large web-based image datasets [43], [44], [64] to extract robust features for downstream tasks [46]–[48], [50], [51]. Knowledge distillation further offers a practical way to maintain robustness while enabling real-time deployment [65], [66]. These observations motivate the use of frozen DINO-pretrained ViT backbones with task-specific heads, where combining robust features with reliable annotations can enhance out-of-distribution generalization and segmentation performance, even on small UAV datasets.

III. UAVID++ DATASET

A. Why the UAVid++ Dataset?

UAVid [39] is an oblique-view aerial dataset comprising 42 video sequences of dense urban environments with numerous dynamic objects. It provides 420 manually annotated frames with eight semantic classes: *Background Clutter*, *Building*, *Road*, *Tree*, *Low Vegetation*, *Dynamic Car*, *Static Car*, and *Human*. Despite its utility as a benchmark for urban oblique-view aerial segmentation, a detailed pixel-level analysis of UAVid reveals annotation issues that undermine both semantic understanding and evaluation reliability.

Labeling errors are prevalent, with connected components frequently assigned incorrect semantic classes, leading to learning confusion. For instance, Fig. 2 (seq27, 00300) shows ground vegetation regions mislabeled as *Human*, highlighted by red rectangles. Annotation inconsistencies further undermine temporal coherence, as the same real-world structure may receive different labels or exhibit varying contour delineations

across consecutive frames or across video sequences. These inconsistencies are illustrated by the orange rectangles in Fig. 2 (seq40, 00800), where the same building is assigned different labels and contours across various frames. Moreover, many objects exhibit coarse polygonal annotations that fail to accurately capture true object boundaries, particularly for vehicles and pedestrians. Representative cases where the outlines are visibly misaligned with the real-world structure are highlighted by the red circles in Fig. 2 (seq21, 00600).

Additional labeling deficiencies occur along image boundaries, where 1 to 10 pixel bands present along most edges create *Background Clutter* fragments. Residual labeling errors also persist, appearing as small unlabeled or incorrectly labeled artifacts that compromise annotation integrity. In the official UAVid annotations, approximately 45,300 connected components smaller than six pixels are present across the dataset, with nearly 33,400 assigned to the *Background Clutter* class. These components predominantly correspond to mislabeled noise rather than meaningful scene content, highlighting pervasive annotation errors and further emphasizing the need for refined labels.

Furthermore, the pixel-level distribution of UAVid semantic classes (Fig. 3a, red) shows that approximately 18.9% of the dataset is unlabeled, assigned to the *Background Clutter* class, while the most dominant class, *Building*, covers up to 27.6% of all pixels. Visual inspection of images containing large proportions of these two classes reveals substantial intra-class variability, which, combined with their prevalence, introduces a noticeable class imbalance. This imbalance limits complex scene understanding and reduces semantic diversity. In addition, visually heterogeneous regions sharing the same label further aggravate confusion, making class separability and model learning more challenging.

To overcome these challenges, we refine and extend UAVid into UAVid++, addressing the main sources of unreliability in training and evaluation. Annotation errors, inconsistent contours, and semantically coarse categories are systematically corrected to enhance fidelity and separability. The expanded taxonomy introduces three additional classes, enriching the semantic space and mitigating class imbalance. By improving label quality and category definitions, UAVid++ enables more effective training, fairer model comparison, and stronger

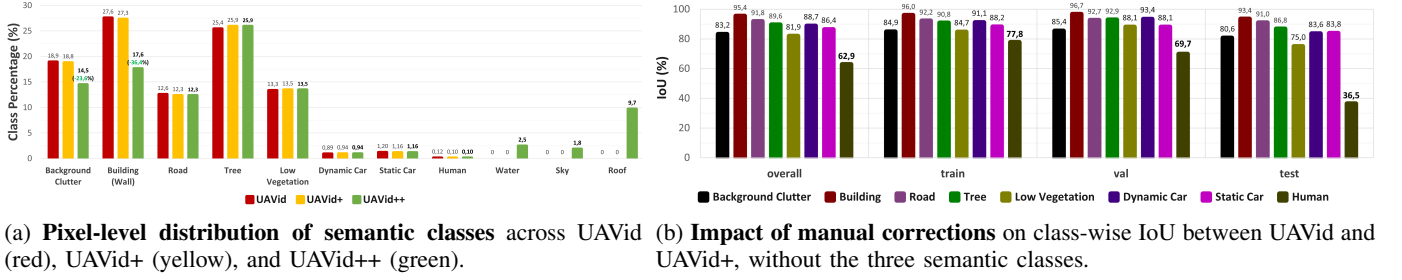


Fig. 3: Statistical analysis of UAVid, UAVid+, and UAVid++: pixel-level distribution and IoU differences.

generalization across diverse UAV imaging conditions, thereby advancing aerial semantic segmentation research.

Although UAVid++ inherits the scene diversity of the original UAVid dataset, its dense and consistent annotations provide a reliable foundation for adapting large-scale pre-trained representations to the aerial domain, supporting robust generalization beyond the diversity captured during training.

B. Annotation Protocol and Guidelines

1) *Annotation Tool*: To address the identified annotation deficiencies, we manually corrected all 420 semantic masks using GIMP [67] as the annotation tool. We used GIMP because UAVid [39] provides ground truth as rasterized segmentation masks rather than polygonal or vector data. Vector-based tools would require re-labeling each image from scratch, since they cannot directly edit existing pixel masks. GIMP allowed the annotators to work on the PNG label maps at 4K resolution, while inspecting the corresponding RGB frame. Within this setup, mislabeled regions were corrected, object boundaries were refined along real-world edges, and the three new classes were integrated. We organized the 42 UAVid video sequences into 12 scenes, each representing a real-world area captured by multiple flyover passes. This enabled consistent joint review and labeling of spatially co-located sequences across viewpoints and time. Annotations cover visible content up to several kilometers in line of sight when image quality permits, with vehicles and humans annotated at distances of up to several hundred meters.

2) *Annotation Corrections*: Corrections targeted all errors: incorrect class assignments, inter-frame and inter-sequence inconsistencies, coarse object contours, image boundary artifacts, and labeling noise. The result is a consistent, enhanced, manually annotated dataset with accurate boundaries and semantic coherence, referred to as UAVid+.

3) *Expanded Semantic Taxonomy*: Based on the pixel-level analysis of the UAVid semantic classes, we introduce three additional semantic categories: *Water*, *Sky*, and *Roof*. These classes were selected not only to improve separability among existing categories, but also to support broader perception tasks beyond semantic segmentation. *Water* includes all visible water surfaces, enabling reliable detection of reflective regions that may challenge perception systems. *Sky* captures all visible sky regions, providing a consistent representation of high-intensity areas that commonly induce saturation and dynamic-range issues. *Roof* separates building roofs from facades,

offering structural cues that facilitate downstream tasks such as surface-normal and depth estimation. Collectively, these additions expand the semantic taxonomy, enhance scene expressiveness, and improve applicability for multi-task aerial perception. Together with the annotation corrections, these efforts yield the final UAVid++ dataset.

4) *Class-Boundary Definitions*: For the inherited semantic classes, our corrections followed the class-boundary definitions of UAVid [39]. For the three new classes, we apply the following definitions. The *Roof* class covers all upper surfaces of a building, namely pitched, mansard, flat, and stepped roofs, as well as roof-mounted equipment, such as the air-conditioning units and solar panels present in the dataset. The *Roof* and *Building (Wall)* boundary is defined at the lower edge of the roof overhang for pitched roofs and at the top edge of the parapet or cornice for flat roofs. Vertical outer surfaces below this boundary are labeled as *Building (Wall)*. The *Sky* class covers all visible sky regions, including clouds and haze. Backlit scene entities retain their foreground label up to the visible outline. The *Water* class includes all visible water surfaces, encompassing the lakes, ponds, swimming pools, and puddles present in the dataset, regardless of reflectance or mirroring. Objects floating on or partially submerged in water retain their original class labels down to the waterline. Wet ground without pooled water is not labeled as *Water*.

5) *Quality Control*: Each annotated image was rigorously corrected and refined at the pixel level, including the addition of the three new semantic classes. Annotation was performed by a team of two trained annotators, with each responsible for fully segmenting entire scenes to ensure a comprehensive understanding of the real-world objects present. Every frame was cross-checked by two additional reviewers to enforce consistency and factual correctness. This protocol minimized annotator bias and produced an improved dataset for benchmarking. On average, each image required approximately eight hours for annotation corrections, with an additional three hours to incorporate the three new semantic classes.

6) *Inter-Annotator Agreement*: Following the inter-annotator protocol established in Cityscapes [28], we assessed annotation reliability on a 5% image subset (randomly sampled across the 12 scenes), independently re-annotated by both trained annotators without access to each other's labels. We used pairwise mIoU between the two annotations as the agreement measure. Initial inter-annotator agreement was 90.6% mIoU. After an independent image-level review by both reviewers, agreement reached 94.7% mIoU. Most

disagreements occurred for objects that are inherently ambiguous in oblique UAV views: the *Human* class (small object scale and frequent occlusion), *Low Vegetation* versus *Tree* (similar appearance between tall bushes and trees), and distant objects in the upper image region (minor contour differences).

C. Dataset Statistics

We present the distribution statistics of the two UAVid++ variants in Fig. 3a, with yellow indicating UAVid+ and green representing the final UAVid++ dataset. Introducing the three additional semantic classes significantly decreases the fraction of pixels labeled as *Background Clutter* and *Building*, resulting in enhanced class separability and richer semantic representation. Overall, approximately 700 million pixels were modified across the 420 annotated images to achieve higher labeling fidelity and incorporate the expanded semantic taxonomy, representing 19.6% of the entire dataset. Furthermore, the number of connected components for vehicles and pedestrians increased by 15.3% (from 30,700 to 35,400), indicating improved separation within the *Dynamic Car*, *Static Car*, and *Human* classes. After manually cleaning residual pixel-level noise, components smaller than six pixels were nearly eliminated. Despite the removal of approximately 45,300 noise-like components, the UAVid++ dataset contains 170,000 connected components in total, with 53,800 newly added ones.

To analyze structural properties of the three dataset variants, we computed connected components for each semantic class using 8-neighborhood connectivity. Components were grouped into seven bins according to pixel area, from regions of at most five pixels to those exceeding 10,000 pixels (Fig. 4). Figure 4a illustrates substantial residual labeling noise in the original UAVid dataset, visible as numerous small (≤ 5 pixels) components highlighted in red. These artifacts are most pronounced in the *Background Clutter* class, but occur across all semantic classes, totaling 45,376 noise-like connected components. The dataset contains 13,926 *Dynamic Car*, 8,882 *Static Car*, and 7,967 *Human* components, of which approximately 600 are noise-like. Overall, the UAVid dataset comprises 161,444 connected components across all classes. Figure 4b illustrates the improvements introduced in UAVid+. The manual removal of noise-like components reduces the number of connected components to 116,070, while refined object boundaries introduce new components and yield a more balanced area distribution, producing a total of 139,721. The counts of *Dynamic Car*, *Static Car*, and *Human* components increase to 15,269, 10,288, and 9,885 respectively. UAVid++ (Fig. 4c) further enhances UAVid+ by adding the *Water*, *Sky*, and *Roof* classes. Counts for *Dynamic Car*, *Static Car*, and *Human* components remain unchanged, while the new categories enrich scene structure. Consequently, the total number of connected components increases to 169,902, surpassing the original UAVid despite the removal of noise-like components.

To quantify the impact of manual corrections, we computed the mean Intersection over Union (mIoU) between UAVid and UAVid+, excluding the three newly added classes. Direct comparison with UAVid++ is not meaningful due to the expanded

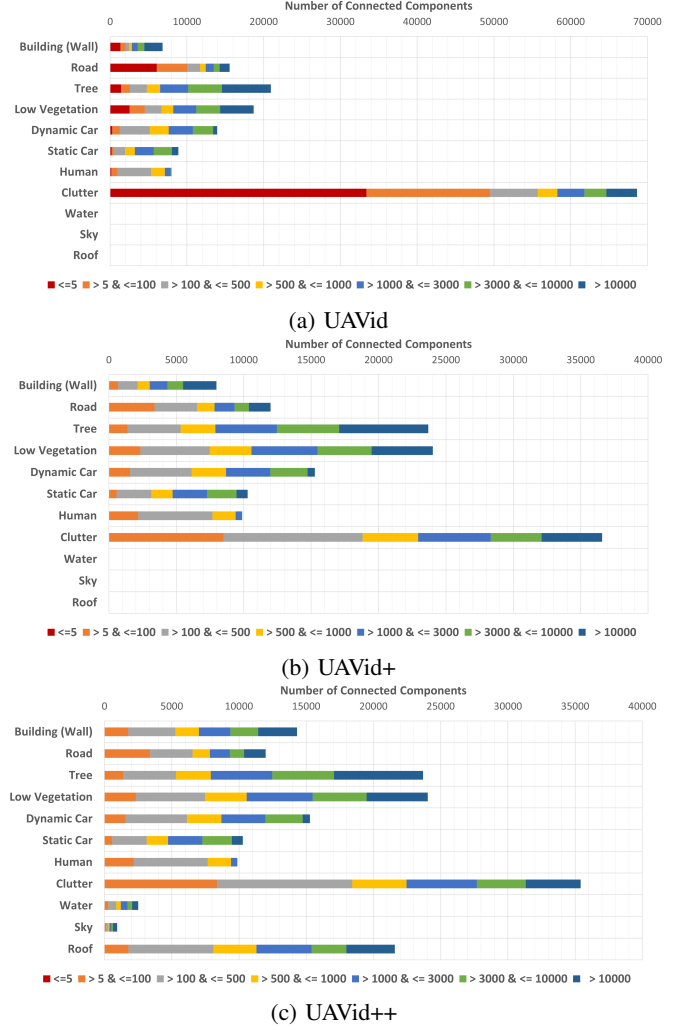


Fig. 4: **Area-based analysis of connected components in UAVid (a), UAVid+ (b), and UAVid++ (c).** UAVid+ reduces small noise-like components (≤ 5 pixels) while increasing instances of *Dynamic Car*, *Static Car*, and *Human*. UAVid++ further increases the total count by adding three new semantic classes. Please zoom in.

semantic taxonomy. The results indicate notable differences (Fig. 3b), particularly for the *Human* class, where the 36.5% mIoU on the test set reflects the extensive modifications made during annotation, underscoring their magnitude. Large pixel-area classes, such as *Building*, *Road*, and *Tree*, exhibited smaller mIoU changes, as contour-level refinements affect only a minor fraction of their total area. Overall, the mIoU between UAVid and UAVid+ is 84.98%.

Figure 1a illustrates the impact of corrected and semantically expanded annotations, comparing the original UAVid semantic labels with UAVid+ and UAVid++. The bottom part of the figure highlights the differences relative to the original UAVid annotations in red. UAVid+ substantially reduces labeling noise, refines object contours, and improves semantic consistency. UAVid++ extends UAVid+ by adding three new semantic classes: *Water*, *Sky*, and *Roof*, enhancing semantic richness while preserving the prior corrections.

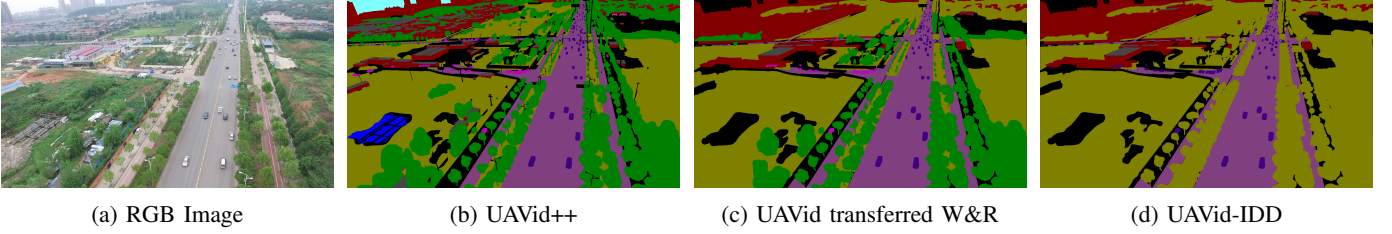


Fig. 5: **Visual comparison of UAVid++ (b), UAVid-IDD (d), and UAVid with transferred Water and Roof labels from IDD (c).** (c) and (d) depict coarse *Roof* annotations (top-left, please zoom in), missing *Water* labels (bottom-left), and no annotation corrections.

D. Comparison with IDD Labels

We quantitatively compare UAVid++ with the IDD [41] *Water* and *Roof* annotations of UAVid by counting patches larger than five pixels. UAVid-IDD contains only 299 *Water* and 4,449 *Roof* coarse segments. For the same 270 images, UAVid++ contains 1,049 *Water* and 12,292 *Roof* segments. Across all 420 images, UAVid++ provides 2,505 *Water* and 21,583 *Roof* segments, yielding five times more regions with higher-quality labels. Visual inspection shows that UAVid-IDD annotations rely on coarse polygons that frequently overlap adjacent entities and are missing in large portions of the scenes (Fig. 5). Overall, UAVid-IDD contains only 270 images with coarser annotations, five times fewer labeled *Water* and *Roof* regions, a reduced semantic taxonomy, no contour refinement for the new classes, and an unassessed impact of the reduced taxonomy and inherited UAVid annotation errors.

IV. ADAPTING DINO ViTs FOR AERIAL SEGMENTATION

We employ an adaptation strategy that combines frozen DINO-pretrained ViT backbones with trainable task-specific heads for aerial semantic segmentation under limited annotated data. The approach is motivated by two key observations. First, foundation models such as DINO-pretrained ViTs [43], [44], [64] learn rich, generalizable representations from large-scale web data. Although not trained for aerial imagery, these features can effectively adapt to aerial domains when combined with high-quality annotations despite substantial viewpoint and appearance shifts. Second, conventional end-to-end training on limited annotated data is prone to overfitting and poor generalization, particularly in the presence of annotation errors or shallow semantic taxonomies.

A. Proposed Approach

The backbone consists of a frozen DINO-pretrained ViT, with experiments conducted across multiple variants from the DINOv2 [43] and DINOv3 [44] families, covering model scales from millions up to several billion parameters. All backbones were pretrained exclusively on large-scale web image datasets, without any specific exposure to aerial imagery. For the DINOv3 variant, we employed models trained on the LVD-1689M dataset. Backbone weights are kept fixed during training, preserving their general visual representations while reducing computational cost. Extending these backbones, we employ three head architectures. (i) A linear probing head

comprising a single layer that directly maps the frozen backbone features to the semantic label space. (ii) A convolutional head composed of two sequential Conv-BatchNorm-ReLU blocks, where the first preserves the number of channels and the second reduces them by half, followed by a 1×1 convolution mapping the features to the semantic space. (iii) A UNet-inspired head [68] consisting of three downsampling stages that progressively halve both spatial resolution and channel dimensionality. Given the high channel dimensionality of the backbone (4096 for ViT-7B, 1024 for ViT-L), this design keeps the head compact, comparable in size to the convolutional head, while effectively aggregating contextual information across multiple feature layers. The downsampling is followed by a bottleneck compressing features to $C/16$, and three upsampling stages restore spatial resolution and channel dimensionality. Skip connections from the corresponding downsampling stages preserve fine-grained details, and a final 1×1 convolution projects the features to the semantic space. This architecture achieves a sixteen-fold compression while maintaining symmetry and balanced representational capacity, producing an effective task-specific head capable of leveraging the backbone’s features.

B. Loss Function

To enhance segmentation performance, we adopt the composite loss defined in Eq. (1), which integrates class-balanced Cross-Entropy loss [68], [69], Dice loss [70], and a weighted Boundary-Aware loss [71] to emphasize accurate contour delineation. This formulation jointly mitigates class imbalance and enhances boundary precision.

$$\mathcal{L} = \mathcal{L}_{\text{CE}}^w + \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Boundary}}^w \quad (1)$$

C. Training and Inference

We adopt the standard $1K \times 1K$ tiling protocol used in recent state-of-the-art aerial segmentation methods [20]–[24], ensuring an objective comparison. As an additional benefit, this reduces the memory requirements of ViT-based backbones while preserving the spatial detail needed for small-object segmentation. Processing full 4K frames would either require aggressive downsampling, which can degrade small-object details, or substantially increase computational cost.

During training and validation, each image is divided into eight patches (1088×1088 for DINOv3 and 1092×1092 for DINOv2) to increase the effective dataset size and match the

TABLE II

Semantic segmentation performance of state-of-the-art methods on UAVid, UAVid+, and UAVid++. Cross-evaluation between UAVid and UAVid+ exposes annotation issues in the original training and test splits. UAVid++ further increases mIoU through its expanded semantic taxonomy. Green values indicate gains over UAVid.

Dataset Used		SOTA Methods (mIoU, %)				
Training Set	Test Set	UNetFormer [20]	LWGANet [21]	SFA-Net [22]	LSKNet [23]	D2LS [24]
UAVid	UAVid	67.87	69.18	69.52	69.03	69.68
	UAVid+	70.38(+2.5↑)	72.33(+3.1↑)	72.64(+3.1↑)	72.65(+3.6↑)	72.72(+3.0↑)
UAVid+	UAVid	68.38	68.01	70.16	70.63	70.94
	UAVid+	72.83(+5↑)	73.52(+4.3↑)	75.85(+6.3↑)	76.44(+7.4↑)	76.96(+7.3↑)
UAVid++	UAVid++	75.07(+7.2↑)	77.66(+8.5↑)	78.43(+8.9↑)	78.61(+9.6↑)	79.35(+9.7↑)

patch embedding configuration. For testing, images are tiled in the same manner, but the prediction patches are recombined to enable high-resolution evaluation while preserving spatial continuity.

V. EXPERIMENTAL EVALUATION

A. Experimental Setup

1) *Datasets*: We evaluate our method and prior state-of-the-art approaches on three dataset variants: the original UAVid [39], the corrected UAVid+, and the expanded UAVid++. The experiments follow the official UAVid data split protocol [39] to ensure fair comparison.

For UAV generalization assessment, we consider two out-of-distribution datasets, UDD [40] and VDD [41], which cover a diverse range of camera orientations and altitudes between 50 and 120 meters. For the semantic labels of UDD and VDD, we use the IDD [41] annotations, which define seven categories (*Building*, *Road*, *Vegetation*, *Vehicle*, *Water*, *Roof*, and *Clutter*) that are conceptually aligned with UAVid++. Notably, these annotations do not include the *Sky* or *Human* classes. To ensure fair comparison, inference results were post-processed by merging *Low Vegetation* and *Tree* into *Vegetation*, *Dynamic Car* and *Static Car* into *Vehicle*, as well as *Sky* and *Human* into *Background Clutter*.

To further assess generalization beyond low-altitude UAV scenarios, we also consider a cross-domain dataset. We use OpenEarthMap [72], a global land-cover benchmark composed of 5,000 high-altitude aerial and satellite images from 44 countries across six continents, with a ground sampling distance of 0.25 to 0.5 m and eight semantic classes: *Bareland*, *Rangeland*, *Developed space*, *Road*, *Tree*, *Water*, *Agriculture land*, and *Building*. For a fair comparison, we map the predictions of models trained on UAVid++ onto the OpenEarthMap taxonomy as follows. *Road*, *Tree*, and *Water* map directly to their UAVid++ counterparts. The *Building* class in OpenEarthMap is mapped to *Roof* in UAVid++, because OpenEarthMap labels only roof surfaces, even when parts of vertical facades are visible. *Rangeland* and *Agriculture land* are both mapped to *Low Vegetation*, as they correspond to low-lying vegetated surfaces. *Bareland* and *Developed space* are mapped to *Background Clutter*, as they represent unstructured ground and miscellaneous surfaces without a dedicated category. UAVid++ classes without a semantic equivalent in OpenEarthMap, namely *Sky*, *Building (Wall)*, *Static Car*, *Dynamic Car*, and *Human*, are

mapped to *Background Clutter*, since OpenEarthMap provides no corresponding category.

2) *Implementation Details*: Starting from the original UAVid split [39], the training, validation, and test sets contain 200, 70, and 150 images, respectively. Given the limited amount of training data and the high resolution of the frames ($4K \times 2K$), all images were partitioned into non-overlapping $1K \times 1K$ tiles. Each image yields eight tiles, producing 1,600 training samples and 560 validation samples. The same tiling strategy was applied at test time, after which prediction patches were recombined to generate full-resolution outputs. All experiments were conducted on a single NVIDIA A100 GPU with 40 GB of memory. For prior state-of-the-art methods, we followed the original implementation settings to ensure fair replication and evaluation. To maintain consistency across all experiments, both training and inference employed a unified tiling strategy using $1K \times 1K$ crops, identical for the proposed approach and all baselines. The proposed method was trained using standard data augmentation (geometric transformations, photometric adjustments, coarse dropout) applied to the image crops, optimized with AdamW [73] at a base learning rate of 3×10^{-5} , weight decay of 0.001, and a batch size of 8 for 40 epochs. The learning rate schedule employed a linear warm-up during the first five epochs, followed by cosine annealing [74] for the remainder.

3) *Metrics*: Semantic segmentation performance was evaluated using per-class Intersection over Union (IoU), with overall performance quantified as the mean Intersection over Union (mIoU) across all classes.

B. Evaluation of Prior Methods

We evaluated five state-of-the-art semantic segmentation methods [20]–[24] on the three dataset variants, with mIoU results reported in Table II. To assess the impact of annotation corrections in both training and test sets, we trained and evaluated state-of-the-art models on various configurations of UAVid and UAVid+. Cross-evaluation revealed that correcting the test set alone improved performance by up to 3.6%, while incorporating training set corrections further increased mIoU, yielding total gains of up to 7.4% (LSKNet [23]). Moreover, on the original UAVid dataset, performance plateaued at approximately 69% mIoU with minimal variation across methods (67.9–69.7%), reflecting limited annotation quality and consistency. In contrast, UAVid+ produced a broader range of results

TABLE III

Semantic segmentation performance for multiple backbone-head configurations trained on the **UAVid**, **UAVid+**, and **UAVid++** dataset variants. Results show consistent gains as annotation quality and semantic richness increase, with improvements observed across all architectures, including very large pretrained backbones (e.g., ViT-7B), highlighting the importance of high-quality fine-tuning data.

Dataset Used		Proposed Method (mIoU, %)				
Training Set	Test Set	ViT-g [43] UNet	ViT-H+ [44] Linear	ViT-H+ Conv	ViT-H+ UNet	ViT-7B UNet
UAVid	UAVid	71.91	63.17	72.03	72.23	72.82
	UAVid+	75.02 (+3.1↑)	65.95 (+2.8↑)	75.34 (+3.3↑)	75.74 (+3.5↑)	76.26 (+3.4↑)
UAVid+	UAVid	72.70	63.19	72.56	73.30	73.45
	UAVid+	77.83 (+5.9↑)	66.29 (+3.1↑)	77.96 (+5.9↑)	78.67 (+6.4↑)	79.41 (+6.6↑)
UAVid++	UAVid++	80.68 (+8.8↑)	70.13 (+6.9↑)	80.86 (+8.8↑)	81.42 (+9.2↑)	82.04 (+9.2↑)

TABLE IV

Semantic segmentation performance of the proposed method on UAVid++ using various frozen backbones and the trainable UNet-inspired head. Green values indicate gains over the best-performing state-of-the-art method.

Strategy	Method	mIoU (%)	IoU per Class (%)										
			Wall	Road	Tree	L.Veg.	D.Car	S.Car	Human	Clutter	Water	Sky	Roof
SOTA	D2LS [24]	79.35	86.0	86.5	86.6	74.2	81.7	73.8	43.1	71.8	89.7	94.5	84.9
Proposed (DINOv2 [43])	ViT-L UNet	79.27	87.0	87.5	85.2	73.2	79.9	77.5	35.3	73.2	92.4	94.9	85.7
	ViT-g UNet	80.68	88.1	88.3	86.1	74.1	<u>81.8</u>	80.6	38.5	74.6	93.1	95.4	87.1
Proposed (DINOv3 [44])	ViT-L UNet	80.82	88.3	88.9	86.6	75.1	81.6	80.8	36.4	74.9	93.5	95.4	87.7
	ViT-H+ UNet	81.42 (+2.1↑)	88.9	89.5	86.6	75.1	81.5	81.7	38.7	76.0	93.8	95.6	88.4
	ViT-7B UNet	82.04 (+2.7↑)	89.0	89.8	86.8	75.4	82.6	82.8	<u>41.1</u>	76.5	94.0	95.7	88.7

(72.8-76.9%), allowing more accurate assessment of model robustness. Expanding to the full UAVid++ dataset with three additional semantic classes provided further gains of up to 4.1% mIoU (LWGANet [21]), leading to a total improvement of up to 9.7% over the UAVid best-performing state-of-the-art method (D2LS [24]). Furthermore, LSKNet [23], initially ranked fourth out of five models on UAVid and considered inferior, rises to second place when trained and evaluated on UAVid+ and UAVid++, surpassing LWGANet [21] and SFA-Net [22]. This demonstrates that improved annotations are essential for accurately assessing the performance of state-of-the-art methods. Overall, both the corrected annotations and the expanded semantic taxonomy enhance training and evaluation, enabling better differentiation between models and improving segmentation accuracy.

C. Evaluation of the Proposed Approach

1) *Quantifying the Impact of Annotation Quality Relative to Pretrained Backbone Capacity:* Large pretrained models can offer a powerful backbone for aerial segmentation, where annotated data is scarce and domain adaptation is often the most effective strategy for achieving high performance and robust out-of-distribution generalization. However, the limited size of available datasets makes the quality of semantic annotations critical, as even models with billions of parameters can be significantly hindered if the training and evaluation data is noisy or imprecise. Table III reports semantic segmentation performance across multiple backbone-head configurations when trained on the three dataset variants. The results show a consistent improvement in mIoU as the training data becomes progressively refined and semantically enriched. When models are trained and evaluated on the original UAVid annotations,

most configurations achieve performance around 72% mIoU. Using the evaluation set of the refined UAVid+ consistently improves performance across architectures, indicating that improved annotations provide a more reliable evaluation. More substantial gains appear when training directly on UAVid+, where several configurations improve by more than 6% mIoU relative to the baseline training setup. The largest improvements are observed when both training and evaluation are performed on UAVid++, with the best-performing configuration achieving 82.04% mIoU, corresponding to a 9.2% improvement over the UAVid baseline. These gains remain consistent even for large pretrained backbones with billions of parameters (ViT-7B), highlighting that high-capacity models still depend strongly on the quality and annotation granularity of fine-tuning data.

The observed performance differences highlight the importance of annotation fidelity relative to model capacity. Increasing the backbone size, from ViT-H+ with 840M parameters to ViT-7B with 6.7B parameters, yields only a 0.6% mIoU gain for both UAVid (72.23% vs 72.82%) and UAVid++ (81.42% vs 82.04%). In contrast, holding the backbone fixed and improving the dataset produces a 9.2% mIoU gain for both backbones, from 72.23% to 81.42% for ViT-H+ and from 72.82% to 82.04% for ViT-7B. Both backbones benefit almost identically from the improved annotations. This demonstrates that scaling the frozen pretrained backbone does not reduce sensitivity to annotation quality. Even at billion-parameter scale, annotation fidelity remains the primary driver of segmentation performance in the small-dataset regime typical of UAV imagery.

2) *Influence of Pretrained Model Capacity under Fixed Annotation Quality:* We further investigate the effect of var-

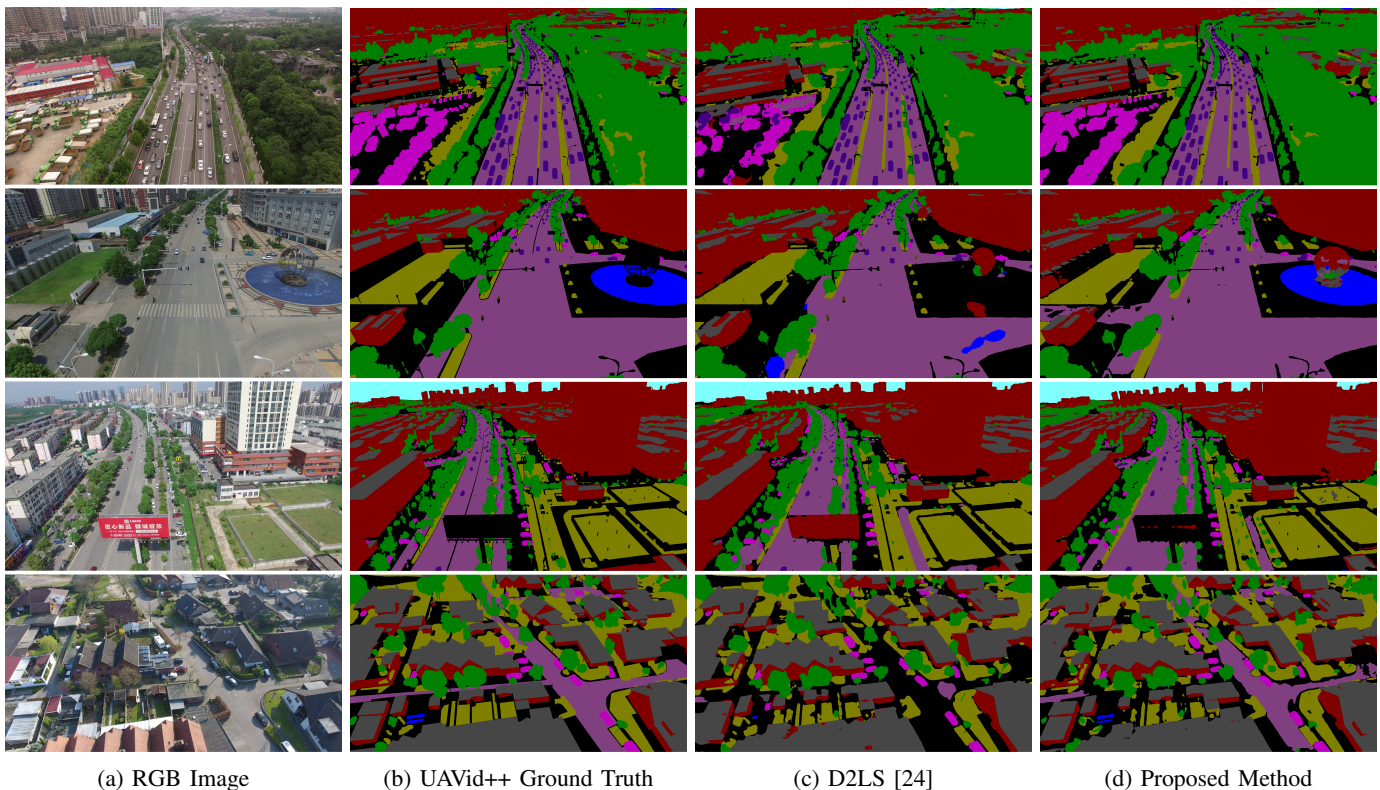


Fig. 6: **Qualitative comparison of the proposed method (d), the UAVid++ ground truth (b), and the best-performing state-of-the-art approach (c).** The proposed method achieves semantic segmentation with clearer distinction in ambiguous or visually similar regions.

TABLE V

Head design ablation study on UAVid++: Frozen DINO-pretrained backbones combined with the proposed trainable heads. The UNet-inspired head consistently achieves the best performance across all backbones. Green values indicate gains over the linear probing approach.

Strategy	Backbone	Head (mIoU, %)		
		Linear	Conv	UNet
DINOv2 [43]	ViT-L	68.40	77.75	79.27 (+10.9↑)
	ViT-g	71.83	80.12	80.68 (+8.8↑)
DINOv3 [44]	ViT-L	72.42	80.41	80.82 (+8.4↑)
	ViT-H+	70.13	80.86	81.42 (+11.3↑)
	ViT-7B	74.95	81.99	82.04 (+7.1↑)

ious pretrained backbones combined with the UNet-inspired head on UAVid++, as summarized in Table IV, which reports detailed per-class IoU results. While performance generally improves with backbone capacity, with the largest DINOv3 model (ViT-7B, 6.7B parameters) achieving +2.7% mIoU over the state-of-the-art [24], competitive results are also obtained with smaller backbones (ViT-H+ 840M, ViT-L 300M, ViT-g 1.1B), demonstrating the robustness of the approach.

3) *Effect of Segmentation Head Design:* Importantly, backbone capacity alone does not explain the observed performance gains. Ablation studies on head design, summarized in Table V, show that replacing linear probing with the UNet-inspired head improves mIoU by up to 11.3%. While the convolutional head achieves competitive results, the UNet-

inspired head consistently delivers the highest accuracy across all backbones, demonstrating the critical role of task-specific hierarchical feature refinement. These findings confirm that the gains over prior state-of-the-art methods result not only from high-capacity pretrained backbones but also from the task-specific design of the head, which effectively leverages the backbone’s representational power for aerial semantic segmentation.

4) *Qualitative Comparison with SOTA:* A qualitative comparison between the proposed method (Fig. 6d) and the best-performing state-of-the-art approach [24] (Fig. 6c) on UAVid++ (Figs. 6a and 6b) is presented in Fig. 6. The proposed method produces better scene understanding with fewer semantic confusions, particularly in differentiating *Static Car* from *Dynamic Car*, demonstrating stronger contextual reasoning. Additional improvements are observed in *Water*, *Road*, and *Building (Wall)*, where visually similar or ambiguous regions are more accurately separated. These qualitative trends align with the quantitative gains reported in Table IV.

D. UAVid++ Taxonomy Ablation Study

We conducted an ablation study to assess the contribution of the newly introduced semantic categories (*Water*, *Sky*, *Roof*). For this analysis, we used the best-performing state-of-the-art model [24]. Starting from UAVid+, we add each class independently: first only *Roof*, then only *Sky*, and finally *Water*. This isolates the individual effect of each category and

TABLE VI

Ablation study on UAvid++ semantic taxonomy using prior SOTA [24]: Gradually adding the three new semantic classes does not significantly affect performance on original classes, confirming that the new categories capture distinct, non-overlapping concepts.

Dataset	mIoU (%)	IoU per Class (%)										
		Building (Wall)	Road	Tree	L.Veg.	D.Car	S.Car	Human	Clutter	Water	Sky	Roof
(0) UAvid	69.68	89.7	82.1	81.7	66.5	72.0	61.7	33.3	70.4	-	-	-
(1) UAvid+	76.96	92.4	86.4	86.4	74.0	81.9	72.1	44.8	77.6	-	-	-
(1.1) $\hookrightarrow$ with <i>Roof</i>	76.54 (-0.4↓)	86.0	86.5	86.5	74.1	78.9	69.8	44.2	77.8	-	-	85.0
(1.2) $\hookrightarrow$ with <i>Sky</i>	78.81 (+1.8↑)	92.6	86.7	86.5	74.2	82.0	72.2	44.6	76.0	-	94.6	-
(1.3) $\hookrightarrow$ with <i>Water</i>	78.11 (+1.1↑)	92.6	86.1	86.6	74.1	82.4	73.5	44.7	73.9	89.1	-	-
(2) UAvid++	79.35 (+2.4↑)	86.0	86.5	86.6	74.2	81.7	73.8	43.1	71.8	89.7	94.5	84.9

tests whether their inclusion alters performance on the original semantic taxonomy. As reported in Table VI, adding *Sky* and *Water* increases overall mIoU, with minor shifts observed in the *Background Clutter* category. The *Roof* class slightly impacts mIoU, producing a small adjustment in the *Building* category. However, visual inspection indicates that these fluctuations do not correspond to a direct decline in model performance. Instead, they stem from inherent ambiguities at category boundaries, where visually similar elements naturally overlap, underscoring the challenge of accurately annotating transitional regions in aerial imagery. Overall, the inclusion of these new classes enriches the semantic space by separating visually distinct elements that were previously aggregated under broader labels, thereby enhancing scene expressiveness and the dataset’s suitability for multi-task aerial perception.

E. Generalization Capabilities

We assess generalization under two regimes. First, out-of-distribution generalization within the UAV domain is evaluated on two datasets captured at varying low altitudes and camera orientations. Second, cross-domain generalization is evaluated on a semantic segmentation dataset of high-altitude aerial and satellite imagery, which exhibits a substantial domain shift relative to the low-altitude oblique UAV setting used for training. For all experiments, we trained exclusively on the UAvid++ training set.

1) *Out-of-Distribution UAV Generalization:* For this experiment, we evaluated on the full UDD [40] and VDD [41] datasets using the IDD [41] annotations. The best-performing state-of-the-art method [24] exhibits a noticeable performance drop (Table VII) under variations in altitude or camera orientation, revealing limited robustness beyond its training distribution. In contrast, for the proposed method, the intermediate-sized ViT-H+ [44] backbone provides rich pre-trained representations, while the UNet-inspired head further refines hierarchical features for aerial segmentation. This combination delivers substantial improvements, achieving gains of 12.4% mIoU on UDD and 8.9% on VDD, with approximately 3% mIoU of the improvement attributable to the UNet-inspired head over linear probing.

Figure 7 illustrates qualitative comparisons between the proposed method (Fig. 7d, ViT-H+ UNet) and the state-of-the-art approach [24] (Fig. 7c), trained on the UAvid++ training set and evaluated on UDD and VDD (Figs. 7a and 7b) using the IDD labels. The proposed method not only generalizes

strongly to unseen conditions, but also exhibits higher visual fidelity compared with the original ground truth annotations. Notable improvements are observed in classes such as *Water* and *Road*, consistent with the quantitative gains reported in Table VII.

2) *Cross-Domain Generalization:* For this experiment, we used the full OpenEarthMap [72] dataset. This constitutes a challenging evaluation scenario, as the model must simultaneously handle shifts in viewpoint (oblique to nadir), altitude, spatial resolution, scene composition, and geographic distribution, without any fine-tuning on the target high-altitude aerial domain. The models were trained exclusively on low-altitude oblique UAV imagery, while the frozen backbone was pretrained on web-scale imagery.

We evaluated the same ViT-H+ UNet model trained on UAvid++, without any retraining, and compared it against D2LS [24] trained under identical data, class mapping, and post-processing settings (Table VIII). Under this substantial domain shift, the proposed pipeline achieved 62.95% mIoU, outperforming D2LS (24.1% mIoU) by a margin of 38.85% mIoU, with consistent improvements across all classes. The largest gains are observed for structured classes: *Roof* increases from 14.9% IoU to 79.1%, and *Road* from 7.2% to 49.5%. Significant gains are also observed on *Water* (27.1% to 71.4%), *Tree* (29.8% to 63.3%), and *Low Vegetation* (37.3% to 66.4%). The smallest improvement is on *Background Clutter* (28.1% to 47.8%), which is the most visually heterogeneous class. These results indicate that the proposed adaptation approach remains discriminative under severe domain shifts, whereas an end-to-end model trained solely on oblique UAV imagery generalizes poorly to the higher-altitude nadir domain.

The qualitative results in Fig. 8 are consistent with the quantitative findings. D2LS (Fig. 8c) produces large unstructured regions and frequently misses or misclassifies road and roof areas, whereas the proposed method (Fig. 8d) preserves building layouts and road structures while closely matching the ground truth (Fig. 8b) across all scenes. Figure 8e further highlights that, although OpenEarthMap does not annotate vehicles, the proposed method still segments both dynamic and static cars under the full UAvid++ label set. Since these objects lack ground truth annotations (Fig. 8b), post-processing maps them to the *Background Clutter* class. Consequently, this reduces the overall mIoU, despite providing valid and meaningful additional information.

TABLE VII

Out-of-distribution generalization performance of the proposed method (ViT-H+ UNet) compared to the best-performing state-of-the-art method (D2LS [24]). Both models were trained on **UAVid++** and evaluated on the unseen datasets **UDD** and **VDD**.

Training Set	Test Set	Method	mIoU (%)	IoU per Class (%)						
				Building	Road	Vegetation	Vehicle	Water	Roof	Clutter
UAVid++	UDD [40]	D2LS [24]	55.46	55.0	42.5	83.5	65.6	58.3	51.5	31.8
		ViT-H+ Linear	64.58 (+9.1↑)	65.6	50.8	81.4	63.3	74.0	77.2	39.7
		ViT-H+ UNet	67.89 (+12.4↑)	64.3	49.8	81.8	70.2	79.8	84.6	44.6
	VDD [41]	D2LS [24]	60.14	36.6	50.7	84.2	73.3	58.3	77.4	40.4
		ViT-H+ Linear	65.98 (+5.8↑)	52.4	64.1	85.6	71.5	70.2	72.1	45.8
		ViT-H+ UNet	69.00 (+8.9↑)	39.3	67.5	85.3	77.4	82.3	71.4	59.7

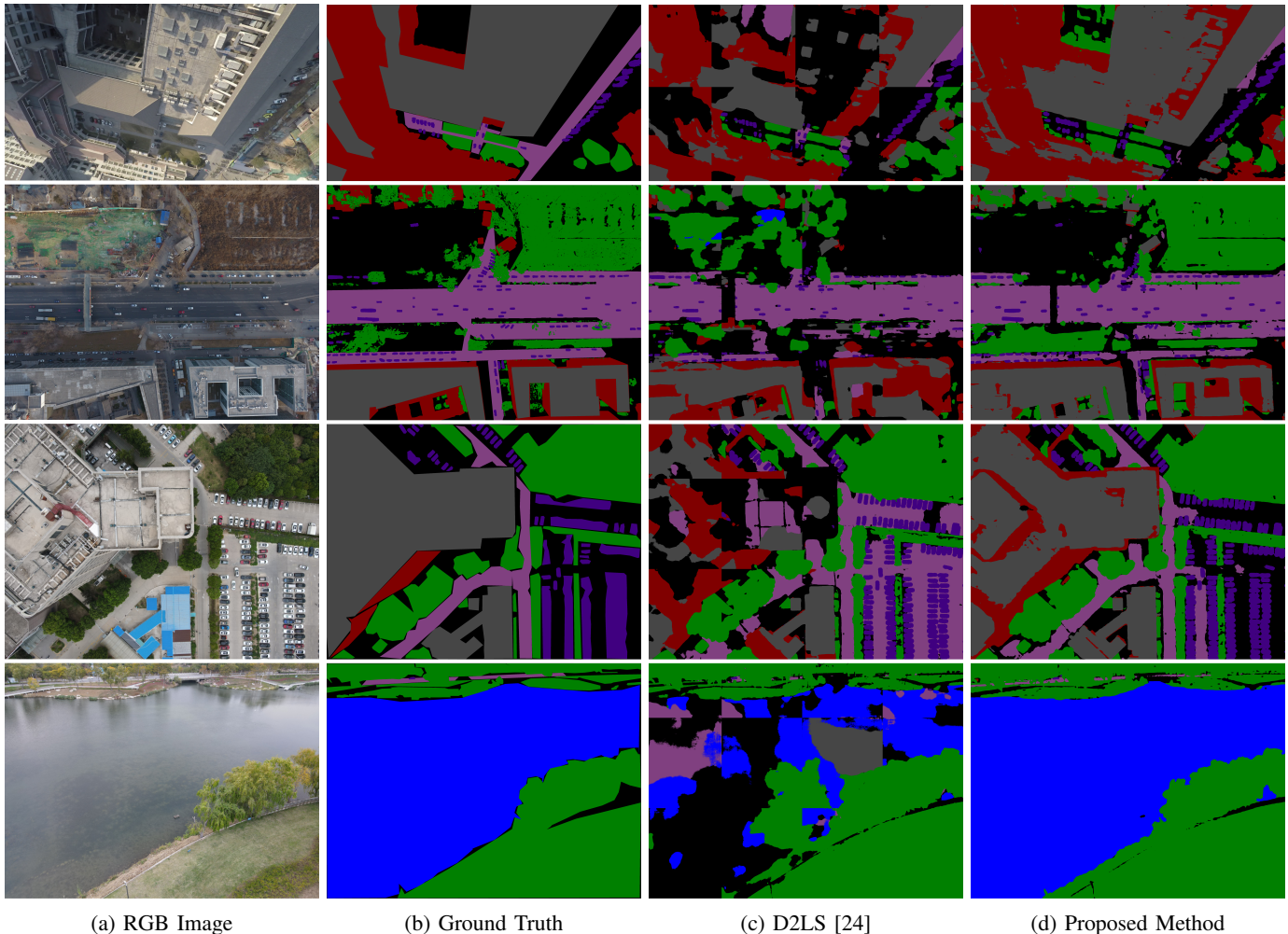


Fig. 7: **Qualitative evaluation of the proposed method on out-of-distribution UAV datasets (UDD [40] and VDD [41]):** The proposed approach maintains robust semantic segmentation across diverse camera orientations and altitudes.

F. Discussion

Our manual correction and semantic taxonomy expansion effort is not intended as a refinement strategy for large-scale datasets [33], [34]. Rather, it addresses a current gap in low-altitude UAV semantic segmentation: the absence of a more reliable benchmark for dense urban aerial scenes. At the typical scale of UAV semantic datasets, where only a few hundred frames are annotated, dense pixel-level refinement is feasible and supports three main benefits, as demonstrated

by our experiments. First, it enables a more discriminative evaluation of state-of-the-art methods. Second, it provides reliable ground truth for adapting large pretrained backbones to the aerial domain. Third, it establishes a robust benchmark for assessing weakly- and self-supervised methods, thereby paving the way for future pseudo-label generation.

Although the proposed adaptation approach leverages large pretrained ViT backbones (ViT-7B, 6.7B parameters; ViT-g, 1.1B; ViT-H+, 840M; ViT-L, 300M), foundation models alone fail to surpass prior methods when paired with a linear probing

TABLE VIII

Cross-domain generalization performance of the proposed method (ViT-H+ UNet) compared to the best-performing state-of-the-art method (D2LS [24]). Both models were trained on UAVid++ and evaluated on unseen high-altitude aerial and satellite data, OpenEarthMap [72], without retraining and under an identical class mapping and post-processing. mIoU and per-class IoU are computed over the six categories shared by the UAVid++ and OpenEarthMap taxonomies. Green values indicate gains over D2LS.

Training Set	Test Set	Method	mIoU (%)	IoU per Class (%)					
				Roof	Road	Low Veg.	Tree	Water	Clutter
UAVid++	OpenEarthMap [72]	D2LS [24]	24.1	14.9	7.2	37.3	29.8	27.1	28.1
		ViT-H+ UNet	62.95 (+38.85↑)	79.1	49.5	66.4	63.3	71.4	47.8

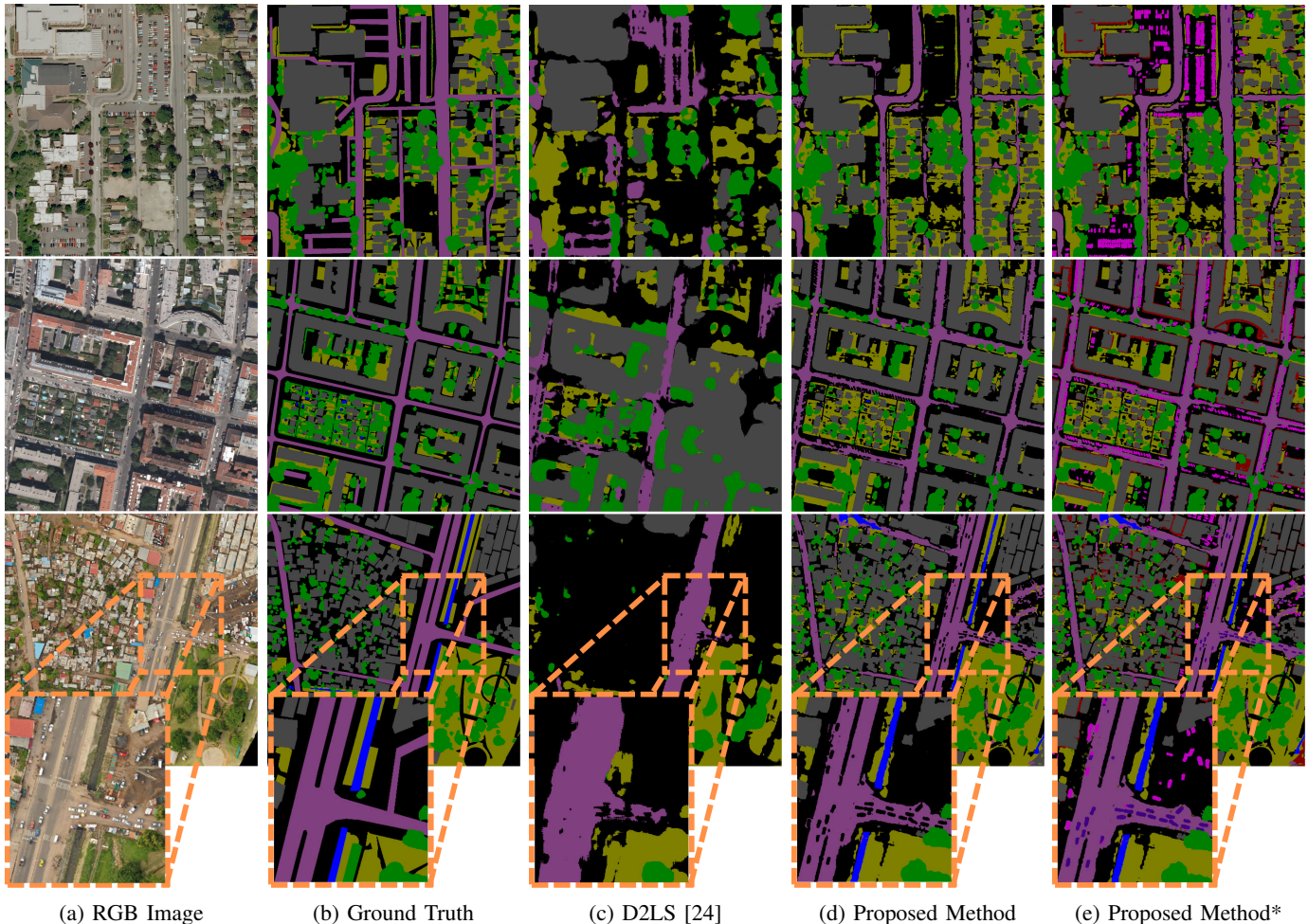


Fig. 8: Qualitative evaluation of the proposed method on the unseen high-altitude aerial and satellite data (OpenEarthMap [72]): The proposed approach maintains robust semantic segmentation under a severe UAV-to-satellite shift. (e) Proposed Method* shows the proposed method predictions using the full UAVid++ taxonomy, where classes absent from OpenEarthMap, such as Dynamic and Static Car, are still segmented. All models are trained exclusively on the UAVid++ training set.

head. The observed improvements stem from the combination of the pretrained backbone with either a convolutional or UNet-inspired head. This enables the model to fully exploit the backbone’s representational capacity while maintaining strong out-of-distribution generalization. The task-specific heads introduce only a small number of additional parameters relative to the backbone. For example, with ViT-H+, linear probing adds 15.3K parameters, the convolutional head 22M, and the UNet-inspired head 26M parameters, the latter consistently

achieving the best results.

Experimental results demonstrate that the proposed adaptation approach enables effective generalization in two distinct settings. The first is to other UAV datasets, evaluated on UDD and VDD, and the second is to high-altitude aerial and satellite imagery, evaluated on OpenEarthMap. However, further improving robustness under larger domain shifts will require expanding the adaptation dataset, which we plan to address in future work.

Importantly, the adaptation method is not intended for

real-time, on-board UAV inference. The frozen pretrained backbone is instead used for benchmarking, for evaluating the effect of annotation quality, and for assessing generalization across unseen aerial domains. Because the method generalizes well to out-of-distribution UAV datasets and cross-domain high-altitude aerial and satellite imagery, future work will investigate two deployment-oriented directions. The first is pseudo-label generation for large volumes of unannotated UAV videos, which could expand the available training data without additional manual annotation. The second is knowledge distillation [65], which could transfer the segmentation capabilities of the large model to compact networks that better satisfy the payload, power, and latency constraints of on-board UAV deployment. In both applications, the annotations provided by UAVid++ will ensure that the resulting models are built upon a reliable foundation.

VI. CONCLUSIONS

We introduce UAVid++, a manually corrected semantic variant of the UAVid dataset, enriched with three finely annotated classes: *Water*, *Sky*, and *Roof*. The annotation corrections and expanded taxonomy yield cumulative improvements of up to 9.7% mIoU for prior state-of-the-art methods and enable a broader and more informative benchmarking. Building on this, we employ an adaptation strategy that combines frozen DINO-pretrained ViT backbones with task-specific heads for aerial datasets with limited annotations. Trained on UAVid++, the strategy surpasses the best-performing state-of-the-art method by 2.7% mIoU and exhibits strong generalization, reaching gains of up to 12.4% mIoU on out-of-distribution UAV datasets and 38.85% mIoU on a cross-domain dataset. Our analysis demonstrates that even large pretrained models rely critically on accurate annotations for effective adaptation to task-specific applications.

REFERENCES

- [1] F. Betti Sorbelli, "Uav-based delivery systems: A systematic review, current trends, and research challenges," *Journal on Autonomous Transportation Systems*, vol. 1, no. 3, pp. 1–40, 2024.
- [2] J. Lee, T. Miyanishi, S. Kurita, K. Sakamoto, D. Azuma, Y. Matsuo, and N. Inoue, "Citynav: A large-scale dataset for real-world aerial navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 5912–5922.
- [3] Y. Chang, Y. Cheng, U. Manzoor, and J. Murray, "A review of uav autonomous navigation in gps-denied environments," *Robotics and Autonomous Systems*, vol. 170, p. 104533, 2023.
- [4] J. Zhu, S. Peng, L. Wang, H. Tan, Y. Liu, M. Zhang, and S. Yan, "Lod-loc v2: Aerial visual localization over low level-of-detail city models using explicit silhouette alignment," 2025. [Online]. Available: <https://arxiv.org/abs/2507.00659>
- [5] H. Shakhathreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE access*, vol. 7, pp. 48 572–48 634, 2019.
- [6] S. M. S. M. Daud, M. Y. P. M. Yusof, C. C. Heo, L. S. Khoo, M. K. C. Singh, M. S. Mahmood, and H. Nawawi, "Applications of drone in disaster management: A scoping review," *Science & Justice*, vol. 62, no. 1, pp. 30–42, 2022.
- [7] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89 644–89 654, 2021.
- [8] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment," *Scientific data*, vol. 10, no. 1, p. 913, 2023.
- [9] S. A. H. Mohsan, N. Q. H. Othman, Y. Li, M. H. Alsharif, and M. A. Khan, "Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends," *Intelligent service robotics*, vol. 16, no. 1, pp. 109–137, 2023.
- [10] C. Kyrkou and T. Theodoridis, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles," in *CVPR workshops*, 2019, pp. 517–525.
- [11] A. M. R. Bernal, J. Cleland-Huang, and W. Scheirer, "Psych-occlusion: Using visual psychophysics for aerial detection of occluded persons during search and rescue," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 3383–3395.
- [12] A. M. Russell Bernal, W. Scheirer, and J. Cleland-Huang, "Nomad: A natural, occluded, multi-scale aerial dataset, for emergency response scenarios," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 8584–8595.
- [13] Y. Wan, Y. Zhong, A. Ma, and L. Zhang, "An accurate uav 3-d path planning method for disaster emergency response based on an improved multiobjective swarm intelligence algorithm," *IEEE Transactions on Cybernetics*, vol. 53, no. 4, pp. 2658–2671, 2022.
- [14] H. Xu, L. Wang, W. Han, Y. Yang, J. Li, Y. Lu, and J. Li, "A survey on uav applications in smart city management: Challenges, advances, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 8982–9010, 2023.
- [15] G. Zhang, Y. Liu, X. Yang, H. Huang, and C. Huang, "Trafficnight: An aerial multimodal benchmark for nighttime vehicle surveillance," in *European Conference on Computer Vision*. Springer, 2024, pp. 36–48.
- [16] H. Nguyen, K. Nguyen, A. Pemasiri, F. Liu, S. Sridharan, and C. Fookes, "Ag-vpreid: A challenging large-scale benchmark for aerial-ground video-based person re-identification," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1241–1251.
- [17] H. Qin, T. Xu, T. Li, Z. Chen, T. Feng, and J. Li, "Must: The first dataset and unified framework for multispectral uav single object tracking," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16 882–16 891.
- [18] S. Byun, I.-K. Shin, J. Moon, J. Kang, and S.-I. Choi, "Road traffic monitoring from uav images using deep learning networks," *Remote Sensing*, vol. 13, no. 20, p. 4027, 2021.
- [19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [20] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.
- [21] W. Lu, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "Lwganet: A lightweight group attention backbone for remote sensing visual tasks," 2025. [Online]. Available: <https://arxiv.org/abs/2501.10040>
- [22] G. Hwang, J. Jeong, and S. J. Lee, "Sfa-net: Semantic feature adjustment network for remote sensing image segmentation," *Remote Sensing*, vol. 16, no. 17, p. 3278, 2024.
- [23] Y. Li, X. Li, Y. Dai, Q. Hou, L. Liu, Y. Liu, M.-M. Cheng, and J. Yang, "Lsknet: A foundation lightweight backbone for remote sensing: Y. li et al." *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1410–1431, 2025.
- [24] X. Zou, Y. Li, S. Zhang, K. Li, S. Wang, P. Tao, J. Xing, and C. Lang, "Dynamic dictionary learning for remote sensing image segmentation," 2025. [Online]. Available: <https://arxiv.org/abs/2503.06683>
- [25] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical image analysis*, vol. 65, p. 101759, 2020.
- [26] D. Acuna, A. Kar, and S. Fidler, "Devil is in the edges: Learning semantic boundaries from noisy annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 075–11 083.
- [27] C. Agnew, C. Eising, P. Denny, A. Scanlan, P. Van De Ven, and E. M. Grua, "Quantifying the effects of ground truth annotation quality on object detection and instance segmentation performance," *IEEE Access*, vol. 11, pp. 25 174–25 188, 2023.
- [28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
- [29] A. Das, Y. Xian, Y. He, Z. Akata, and B. Schiele, "Urban scene semantic segmentation with low-cost coarse annotation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 5978–5987.

- [30] A. Zlateski, R. Jaroensri, P. Sharma, and F. Durand, "On the importance of label quality for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1479–1487.
- [31] Y. Liu and J. Wang, "Can less be more? when increasing-to-balancing label noise rates considered beneficial," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17467–17479, 2021.
- [32] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
- [33] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kotschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.
- [34] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.
- [35] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1499–1508.
- [36] H. Florea, V.-C. Miclea, and S. Nedevschi, "Wilduav: Monocular uav dataset for depth estimation tasks," *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2021.
- [37] A. Marcu, V. Licaret, D. Costea, and M. Leordeanu, "Semantics through time: Semi-supervised segmentation of aerial videos with iterative label propagation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [38] A. Marcu, M. Pirvu, D. Costea, E. Haller, E. Slusanschi, A. N. Belbachir, R. Sukthankar, and M. Leordeanu, "Self-supervised hypergraphs for learning multiple world interpretations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 983–992.
- [39] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.
- [40] Y. Chen, Y. Wang, P. Lu, Y. Chen, and G. Wang, "Large-scale structure from motion with semantic constraints of aerial images," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 347–359.
- [41] W. Cai, K. Jin, J. Hou, C. Guo, L. Wu, and W. Yang, "Vdd: Varied drone dataset for semantic segmentation," *Journal of Visual Communication and Image Representation*, vol. 109, p. 104429, 2025.
- [42] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," 2021. [Online]. Available: <https://arxiv.org/abs/2103.14749>
- [43] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: <https://arxiv.org/abs/2304.07193>
- [44] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jegou, P. Labatut, and P. Bojanowski, "Dinov3," 2025. [Online]. Available: <https://arxiv.org/abs/2508.10104>
- [45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [46] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," 2023. [Online]. Available: <https://arxiv.org/abs/2205.08534>
- [47] T. Kerssies, N. Cavagnero, A. Hermans, N. Norouzi, G. Averta, B. Leibe, G. Dubbelman, and D. de Geus, "Your vit is secretly an image segmentation model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25303–25313.
- [48] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3041–3050.
- [49] M. Käßeler, K. Petek, N. Vödisch, W. Burgard, and A. Valada, "Few-shot panoptic segmentation with foundation models," in *2024 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 7718–7724.
- [50] Y. Wu, X. Li, J. Li, K. Yang, P. Zhu, and S. Zhang, "Dino is also a semantic guider: Exploiting class-aware affinity for weakly supervised semantic segmentation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1389–1397.
- [51] B. Zhang, S. Yu, J. Xiao, Y. Wei, and Y. Zhao, "Frozen clip-dino: A strong backbone for weakly supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [52] L. Zhang, R. Tanno, M.-C. Xu, C. Jin, J. Jacob, O. Ciccarilli, F. Barkhof, and D. Alexander, "Disentangling human error from ground truth in segmentation of medical images," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15750–15762, 2020.
- [53] S. Liu, K. Liu, W. Zhu, Y. Shen, and C. Fernandez-Granda, "Adaptive early-learning correction for segmentation from noisy annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2606–2616.
- [54] S. Girisha, U. Verma, M. M. Pai, and R. M. Pai, "Uvid-net: Enhanced semantic segmentation of uav aerial videos by embedding temporal information," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4115–4127, 2021.
- [55] A. Song, "Deep learning-based semantic segmentation of urban areas using heterogeneous unmanned aerial vehicle datasets," *Aerospace*, vol. 10, no. 10, p. 880, 2023.
- [56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [57] T. Hanyu, K. Yamazaki, M. Tran, R. A. McCann, H. Liao, C. Rainwater, M. Adkins, J. Cothren, and N. Le, "Aerialformer: Multi-resolution transformer for aerial image segmentation," *Remote Sensing*, vol. 16, no. 16, p. 2930, 2024.
- [58] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
- [59] —, "Bitsrs: A bi-decoder transformer segmentor for high-spatial-resolution remote sensing images," *Remote Sensing*, vol. 15, no. 3, p. 840, 2023.
- [60] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [61] X. Ma, R. Lian, Z. Wu, R. Guan, T. Hong, M. Zhao, M. Ma, J. Nie, Z. Du, S. Song *et al.*, "A novel scene coupling semantic mask network for remote sensing image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 221, pp. 44–63, 2025.
- [62] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.
- [63] V. Spasev, I. Dimitrovski, I. Chorbev, and I. Kitanovski, "Semantic segmentation of unmanned aerial vehicle remote sensing images using segformer," in *International Conference on Intelligent Systems and Pattern Recognition*. Springer, 2024, pp. 108–122.
- [64] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [65] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [66] A. M. Mansourian, R. Ahmadi, M. Ghafouri, A. M. Babaei, E. B. Golezani, Z. Y. Ghamchi, V. Ramezani, A. Taherian, K. Dinashi, A. Miri, and S. Kasaei, "A comprehensive survey on knowledge distillation," 2025. [Online]. Available: <https://arxiv.org/abs/2503.12067>
- [67] The GIMP Development Team, "Gnu image manipulation program (gimp), version 3.0.6," <https://gimp.org/>, 2025, free software, GPLv3 license.
- [68] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [69] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, 2019, pp. 9268–9277.

- [70] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.
- [71] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, “Boundary iou: Improving object-centric image segmentation evaluation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 334–15 342.
- [72] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, “Openearthmap: A benchmark dataset for global high-resolution land cover mapping,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6254–6264.
- [73] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>
- [74] —, “Sgdr: Stochastic gradient descent with warm restarts,” 2017. [Online]. Available: <https://arxiv.org/abs/1608.03983>

VII. BIOGRAPHY SECTION



Vivian Chiciudean received the engineer and M.S. degrees in the field of Computer Science from the Technical University of Cluj-Napoca in 2022 and 2024, respectively. He is currently pursuing a Ph.D. degree in Computer Science at the same university, specializing in Computer Vision approaches for environment perception. His research interests include data augmentation methods, semantic segmentation and UAV applications.



Sergiu Nedevschi (Member, IEEE) received his M.S. and Ph.D. in electrical engineering from the Technical University of Cluj-Napoca, Romania, in 1975 and 1993, respectively. He has been with TUCN since 1983, serving as Full Professor of Computer Science, founder and head of the Image Processing and Pattern Recognition Research Center, Head of the Computer Science Department (2000-2004), Dean of the Faculty of Automation and Computer Science (2004-2012), and Vice-Rector for Scientific Research (2012-2020). He has coordinated over 85 research projects, collaborating with automotive companies and research institutes. He has published more than 400 articles and edited over 20 volumes. His research focuses on image processing, pattern recognition, computer vision, machine learning, and intelligent and autonomous vehicles.



vehicles.

Florin Oniga received the Diploma Engineer degree in Computer Science in 2002 from the Technical University of Cluj-Napoca, Romania, followed by the M.S. degree in 2003 and the Ph.D. (Computer Science) degree in 2011, from the same university. He is a full professor at the Computer Science Department, Technical University of Cluj-Napoca, teaching Image Processing, Pattern Recognition, and Computer Architecture. His research interests include stereovision, digital elevation maps processing, and environment perception systems for intelligent