A NEW BENCHMARK DATASET AND AN IMPRECISE SUPERVISED LEARNING APPROACH FOR AIRPORT SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Airport semantic segmentation is fundamental to airport surveillance, yet there currently lacks a specialized benchmark and algorithms for this task. In this paper, we introduce ASS, the first large-scale Airport Semantic Segmentation dataset tailored for airport scenarios. ASS comprises 18 common semantic categories, 250 videos, and over 140,000 frames with precise manual annotations. The dataset covers a wide spectrum of challenges inherent to airport surveillance, including intra-class diversity, unusual shapes, inter-class similarity, extreme multi-scale, class imbalance, and varying weather and illumination conditions. We evaluate 20 state-of-the-art semantic segmentation methods on ASS and observe substantial performance drops, indicating that existing models remain far from practical deployment. The dataset is available at www.agvs-caac.com/ASS/ASS.html. Furthermore, we discuss future directions for airport semantic segmentation and propose a research approach based on imprecise supervised learning, with an initial implementation named SPWSeg (https://github.com/SPWSeg/SPWSeg). Additional experiments show that this research approach generalizes effectively beyond airport scenes.

1 INTRODUCTION

With the rapid growth of global air traffic, airports are becoming increasingly congested, leading to more flight delays and safety incidents like collisions. Building intelligent airports is thus crucial for improving operational efficiency and safety. Applications such as video collision warning systems typically rely on pixel-level masks from semantic segmentation for high-precision scene analysis. However, algorithms that perform well in foundational research show significant performance drops in airport scenes, as illustrated in Fig. 1.

Subsequent experiments show that the average accuracy of current semantic segmentation algorithms in airport scenes drops by up to 50%. This degradation stems from two factors: the unique scene structure and the distinctive visual appearance of airport elements. Structurally, airports are monitored from distances of several kilometers—far beyond typical urban or road scenes—posing extreme multi-scale challenges. Visually, the airplane, as the most prominent class, is non-compact and winged, while elements like runways and high-pole lights are elongated—both shapes notoriously difficult to segment completely. Additionally, inter-class similarity (e.g., runways vs. liaison roads) and intra-class variation (e.g., high-pole lights off during the day vs. on at night) further complicate segmentation. Collectively, these factors render public baselines unsuitable for airport scenes, motivating our dedicated dataset and specialized algorithms.

This paper introduces ASS, the first large-scale dataset for airport semantic segmentation. It contains 250 video sequences with over 140,000 manually annotated frames, supporting both image and video semantic segmentation research. ASS covers 18 common airport categories (Fig. 2) and captures key challenges including intra-class diversity, unusual shapes, inter-class similarity, class imbalance, extreme multi-scale, and varying weather and illumination conditions. We evaluate 20 image and video semantic segmentation algorithms on ASS and compare against five other datasets. Results confirm that airport scenes are highly challenging, justifying the need for a dedicated dataset and targeted research. Furthermore, we discuss future research directions and propose a research ap-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

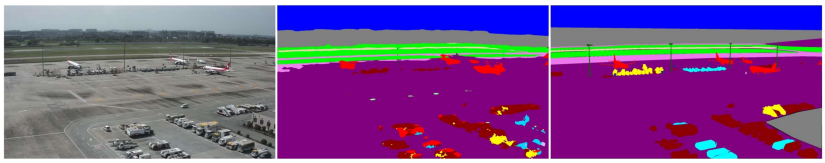

Figure 1: Illustration of airport semantic segmentation. Left to right: original frame, segmentation result by SegFormer Xie et al. (2021), and ground truth.

proach based on imprecise supervised learning, motivated by two considerations: manual annotation is prohibitively costly, and even continuous dataset expansion cannot guarantee full coverage of all scene challenges. As an initial implementation of this approach, we develop shape-Prior-Guided Weakly Supervised Semantic Segmentation (SPWSeg). Experimental results validate its feasibility and cross-scene generalizability.

The rest of the paper is organized as follows. Related work is reviewed in Section 2, the construction process of ASS is described in Section 3, comparison experiments and future research discussion are shown in Section 4, followed by the conclusion in Section 5.

2 Related Work

2.1 Semantic Segmentation Dataset

Image semantic segmentation dataset. Over the past decade, numerous image semantic segmentation datasets have been released, evolving from small-scale tasks to foundation-model-driven paradigms. Early benchmarks like Pascal VOC Everingham et al. (2010), COCO Tsung et al. (2014), and ADE20K Zhou et al. (2017) cover common object categories, such as person, bike, car, dog, sofa, tv, etc. BDD100K Yu et al. (2020) and SemanticKITTI Behley et al. (2019) extended the field to autonomous driving, while LoveDA Wang et al. (2021b) targeted remote sensing. Recently, foundation models spurred explosive data growth; SA-1B Kirillov et al. (2023) provides over a billion masks for zero-shot segmentation, and WHU-Urban3D Han et al. (2024) alongside Mosaic3D Lee et al. (2025) expanded into 3D and open-vocabulary contexts.

Video semantic segmentation dataset. Video semantic segmentation datasets have evolved from driving-sequence annotations to dense spatio-temporal labeling of general dynamic scenes. Early benchmarks such as CamVid Brostow et al. (2009) and Cityscapes Cordts et al. (2016) relied on sparsely sampled annotations from vehicle-mounted cameras. To move beyond driving perspectives, VSPW Miao et al. (2021) and VIPSeg Miao et al. (2022) introduced large-scale continuous annotations in natural scenes, while MVSeg Ji et al. (2023) supplemented these with multispectral data for complex weather conditions. Recently, SA-V Ravi et al. (2024) has advanced the development of video foundation models by providing over 600,000 spatio-temporal masks. However, none of these datasets is specifically designed for airport scenarios, and their sparse airport samples are insufficient to cover the complex variability of real-world airport environments.

2.2 Semantic Segmentation Methods

Image semantic segmentation methods. Modern image semantic segmentation architectures are broadly categorized into CNN-based, Transformer-based, and vision-language model-based. FCN Shelhamer et al. (2017) pioneered end-to-end pixel-wise prediction. PSPNet Zhao et al. (2017) and DeepLabV3 Chen et al. (2017) enhanced context modeling via multi-scale aggregation, while the BiSeNet series Yu et al. (2018; 2021) optimized real-time performance. Later, Swin Transformer Liu et al. (2021) and SegFormer Xie et al. (2021) dominated global context understanding using self-attention, and Mask2Former Cheng et al. (2022) unified major segmentation tasks. Recently, Open-Vocabulary Semantic Segmentation (OVSS) gained prominence with CLIP Radford et al. (2021). Methods like CLIP as RNN Sun et al. (2024b) and CorrCLIP Zhang et al. (2024) align visual and language embeddings, significantly improving zero-shot generalization capability of semantic segmentation.

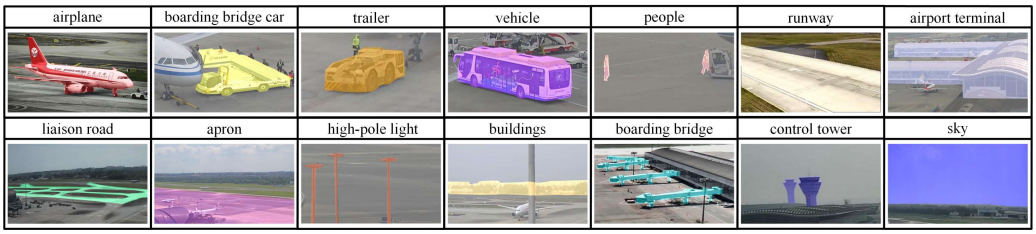

Figure 2: Illustration of some typical semantic classes in ASS.

Video semantic segmentation methods. Video semantic segmentation evolved from explicit inter-frame motion modeling to multi-modal unified architectures emphasizing long-range consistency. Early approaches like CFFM Sun et al. (2022a; 2024a) and MRCFA Sun et al. (2022b) utilized motion cues to strengthen frame-level correlations. Subsequently, frameworks like Tube-Link Li et al. (2023a) and DVIS++ Zhang et al. (2023) achieved end-to-end joint optimization of segmentation and tracking via modular designs. Latest research integrates multi-modal large models: UniVS Li et al. (2024) proposed a prompt mechanism unifying video tasks, SAM2Long Ding et al. (2025) addressed long-sequence error accumulation, and the Mamba-based TV^{3}S Hesham et al. (2025) achieved linear complexity. While yielding strong performance on general benchmarks, these architectures suffer substantial degradation when applied to specialized scenarios like airports.

3 ASS Benchmark

This paper presents the ASS benchmark, covering its construction and major challenges. The construction involves three steps: dataset design, collection, and annotation. Challenges arise from both airport-specific semantics and environmental complexities.

3.1 Dataset Construction

Dataset design. The core of a semantic segmentation dataset lies in the design of its semantic label set. We do not intend to enumerate every single semantic category in the airport, because even for a single airport scenario, listing all individual semantics would result in an excessively large label set. Our principle is to list the important individual semantics (e.g., airplane, people), while allowing certain semantics to encompass multiple subcategories (e.g., vehicle, building). The ASS dataset currently contains a total of 18 semantic categories, of some typical semantic classes are shown in Fig. 2. The definition of each semantic category and the rationale for its selection are explained below.

The airplane is a key airport semantic. Despite livery differences, civil airplanes share similar shapes, so we group them into one category. Vehicles are also common, but unifying all types (cars, trucks, deicing trucks, engineering vehicles) would create excessive intra-class variance, while separating each type would inflate the category count. Thus, functionally unique types like bridge boarding cars and trailers are kept as separate categories, while other less critical or less frequent vehicles are merged into a general vehicle category. Beyond airplanes and vehicles, people form another movable semantic class in airports.

Airport scenes contain diverse buildings. Following the vehicle principle, we do not distinguish them individually but group them as airport terminal (if inside premises) or buildings (if outside). Small auxiliary facilities linked to the main terminal may be included in the airport terminal category, but the frequently occurring boarding bridge is treated separately. The buildings category covers surrounding structures (e.g., skyscrapers, transport buildings) and may also include natural elements like trees and mountains that mix with the built environment. Ground guidance signs are divided into sign (above ground) and landmark (on the ground), while the entire ground area is defined as the apron. Other key structural categories include liaison road, runway, lawn, and high-pole light. Some categories have limited samples, such as security post and control tower. Finally, the sky—always visible but ever-changing—forms another distinct category. Other semantics may occasionally appear in airport scenes, such as birds or foreign objects on the runway. However, these are either too small to observe or too rare to sample effectively. Therefore, we exclude them for now, but they can be added in the future if needed.

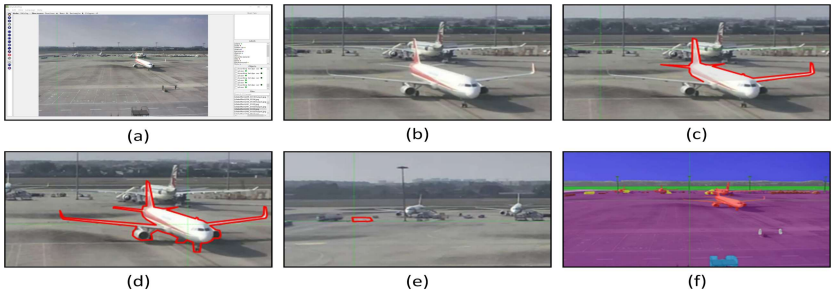

Figure 3: Illustration of data annotation process.

Data collection. To address airport-specific challenges—strict security, restricted access, and dynamic scenes—we partnered with the Civil Aviation Administration of China to gain data collection authorization at five major international airports in Southwest China. We used four fixed cameras and one PTZ camera to capture multi-perspective videos under various weather (clear, rain, fog) and lighting conditions (daylight, twilight, night). The initial raw data comprised over 5,000 short clips (averaging 15 seconds each), but with substantial redundancy due to repetitive scene layouts and motion patterns—e.g., all airports share standard components like terminals, aprons, runways, and taxiways, and aircraft movements boil down to taxiing, takeoff, landing, etc. Through multi-stage filtering based on motion and scene pattern comparisons, we removed redundancies and retained 250 diverse clips, totaling over 140,000 frames.

Data annotation. The data annotation process is shown in Fig. 3. First, open the annotation software AnyLabeling Nguyen (2023) (Fig. 3(a)). The left side shows annotation options, including commonly used polygon annotation tools; the center displays the image to be annotated; the right side lists possible semantic categories and the current annotation directory. Next, annotate objects such as airplanes (Fig. 3(b)): zoom the image to an appropriate size, then use the polygon tool to click points along the airplane's contour accurately (Fig. 3(c)–(d)). Then, annotate the next object, such as a car (Fig. 3(e)). After annotating all objects, the visualization of the annotation results is shown in Fig. 3(f). To improve annotation efficiency and maintain continuity between frames, annotations from the previous frame can be used to guide annotations in subsequent frames. Specifically, the labels from the previous frame are overlaid on the current frame, and then the polygons that make up the labels are adjusted based on visual observation to match the shape of objects in the current frame.

To ensure annotation quality despite subjectivity, we adopt a recognizability criterion: objects are annotated only if their category is discernible upon magnification. We also apply a prioritized annotation principle for semantic boundary ambiguity. Instead of discarding transitional pixels, such as the black pixels between the boarding bridge and airplane in Fig. 4, we first precisely annotate higher-priority categories and assign boundary pixels to adjacent lower-priority ones. Specifically, motion over background takes priority, and small-scale objects take priority over large-scale ones. This guarantees the highest boundary precision for critical targets like airplanes. For a more detailed introduction to the annotation rules and the ground truth format of ASS, please refer to Section A.1 in the supplementary materials.

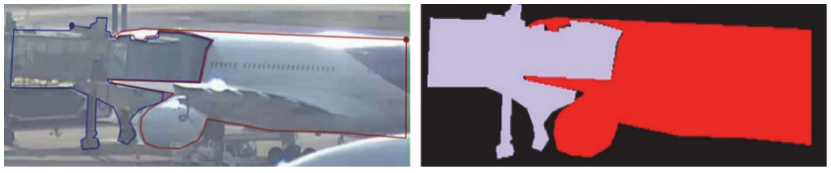

Figure 4: Illustration of the semantic boundary ambiguity problem.

3.2 Challenges

Due to the unique semantic appearances and imaging conditions of airports, semantic segmentation in this domain faces challenges. At the object level, issues include intra-class diversity, unusual shapes, inter-class similarity, extreme multi-scale, and class imbalance. At the environmental level, challenges such as occlusion, weather variations, and illumination changes complicate the task.



Figure 5: Illustration of intra-class diversity.

Intra-class diversity. Certain semantic elements within airports that perform specific operational functions exhibit significant intra-class diversity. For example, high-pole lights differ completely when off during the day vs. on at night (Fig. 5, left); trailers vary markedly in size when loaded vs. empty (Fig. 5, middle); and aircraft, the most operationally critical element, appear substantially different under varying viewing angles (Fig. 5, right). Conventional approaches based on improved feature representations struggle to cover such variations; more powerful feature extraction or finer sub-semantic subdivision is needed.

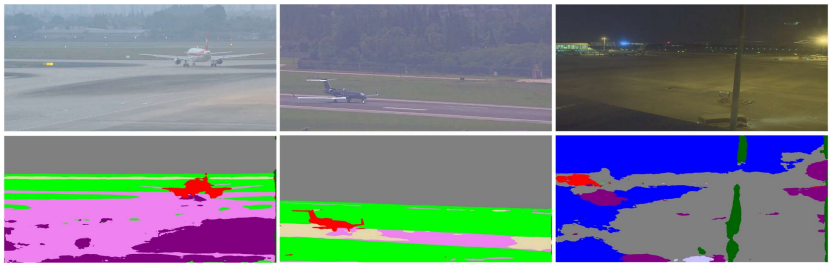

Figure 6: Illustration of unusual shapes. Top: original images. Bottom: segmentation results.

Unusual shapes. Certain semantic elements in airports exhibit unusual shapes, challenging to semantic segmentation. For example, the airplane is a typical wingspan target, and its highly non-compact wingspan shape is recognized as one of the most difficult objects to segment. As shown in the left column of Fig. 6, a significant portion of the aircraft's wingspan was misclassified as the apron category. Several other semantic elements in airports also display similarly non-compact appearances. The middle column of Fig. 6 illustrates the common runway semantic in airports, whose straight and elongated form makes complete segmentation extremely difficult. The high-pole lights in the right column of Fig. 6 also exhibit this elongated shape, and coupled with their intra-class diversity, their segmentation performance was far from satisfactory. For these unusually shaped semantic elements, incorporating shape prior knowledge may be necessary to achieve better results.

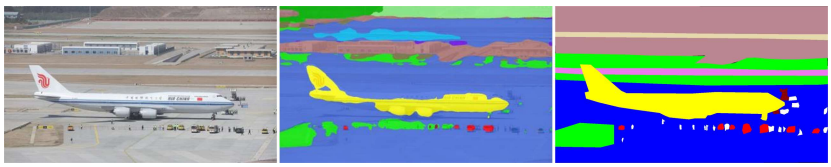

Figure 7: Illustration of inter-class similarity. Original image, segmentation result, ground truth.

Inter-class similarity. In airport scenarios, certain semantic categories resemble each other in specific appearance attributes. For instance, airplanes and the airport ground share color similarity, as shown in Fig. 7. Airplanes are predominantly white, whereas the airport ground similarly exhibits a grayish-white tone. Such color similarity, combined with the distinctive wingspan shape of airplanes, frequently causes severe segmentation defects for airplanes, as depicted in Fig. 7. Developing more expressive feature representations and more discriminative classifiers offers an effective means to address inter-class similarity.

Extreme multi-scale. Multi-scale segmentation is a common challenge, but it proves particularly acute in airport scenes. To begin with, the imaging distance in airport scenes can extend to thousands of meters, giving rise to substantial scale disparities. On one hand, even large targets like an airplane



Figure 8: Illustration of extreme multi-scale phenomenon.

typically appear small on the imaging plane. On the other hand, airport images also contain semantic elements with large imaging areas, such as the sky and apron. Even within the same semantic element, large and small scales may coexist, as shown in Fig. 8(a). This coexistence of large and small targets introduces new difficulties for multi-scale segmentation, since existing multi-scale methods often struggle to handle both extremes simultaneously. Furthermore, some inherently small semantic elements, such as the vehicle and people, are captured at a very diminutive scale in airport scenes, as shown in Fig. 8(b) and Fig. 8(c), respectively. People, in particular, appear quite blurry in the ASS dataset, where only the torso is discernible and the limbs are not. It can be argued that the semantic attributes of people have shifted in airport scenes, rendering existing pedestrian segmentation methods ill-suited for airport scenarios. For these reasons, we characterize the multi-scale phenomenon in airport scenes as extreme multi-scale. The introduction of other challenges in ASS, please refer to Section A.2 in the supplementary materials.

4 Experiments

In this section, we evaluate the ASS benchmark dataset and outline future research directions. We first assess 20 state-of-the-art image and video semantic segmentation models on ASS, comparing their performance with that on other public datasets. Next, using airport semantic segmentation as a case study, we examine the challenges of segmentation in constrained scenes and propose an imprecise supervised learning approach. Finally, we introduce a weakly supervised method for airport segmentation and use it to validate the cross-scene generalization capability of the imprecise supervision paradigm.

4.1 Performance Evaluation on ASS

Before evaluation on ASS, existing segmentation models pre-trained on public datasets require fine-tuning. We adopt the following settings: initial learning rate of 1e4, ”poly” learning rate schedule, weight decay of 5e4, fine-tuning for 20k iterations, batch size of 8, and cross-entropy loss. All images are resized to 1280×960 for both training and testing. Data augmentation includes random horizontal flipping, scaling, cropping, and color jitter. Pixel-level segmentation performance is measured using class-wise Intersection over Union (IoU) and mean IoU (mIoU) across all categories.

The comprehensive evaluation of 15 image semantic segmentation models on ASS is shown in Tab. 1, which reveals significant performance disparities across different semantics. For large-scale semantics, including the sky, apron, and terminal, the average IoU exceeds 59%, while other semantics are considerably smaller in scale, and their segmentation performance is far inferior to that of large-scale semantics. In particular, some small-scale semantics, such as the people, as well as semantics with special shapes, such as the runway, have an average IoU of less than 6%. This indicates that the ASS dataset and airport scenes contain some highly challenging semantics. For the most critical semantic element, the airplane, Swin Liu et al. (2021) and GCNet Yang et al. (2025) achieved 70.20% IoU, while the average accuracy across all algorithms was only 48.31%. Although this performance is not particularly poor compared to other airport semantic categories, airplanes are non-compact wingspan targets. Consequently, even minor detection defects—such as wing damage—can result in significant appearance distortions, as illustrated in Fig. 7. For the visualization experiments of image semantic segmentation and the experiments of video semantic segmentation, please refer to Section B.1 and B.2 in the supplementary materials.

Next, we compare the ASS dataset with five other public semantic segmentation benchmarks using ten baseline models; the results are presented in Tab. 2. These five datasets are among the most widely used in the field, each comprising samples collected from diverse scenes. As shown in the

Table 1: The IoU of 15 image semantic segmentation models on 12 categories of ASS.

ModelsClasses	sky	liaison road	lawn	apron	runway	signs	landmark	terminal	security post	airplane	high-pole light	people
FCN Shelhamer et al. (2017)	76.40	33.61	44.91	67.89	5.36	29.76	18.27	75.06	33.65	63.34	43.70	12.82
PSPNet Zhao et al. (2017)	76.04	32.37	46.24	57.48	6.44	35.41	25.23	69.92	4.61	61.88	47.31	**15.38**
DeepLabV3 Chen et al. (2017)	78.09	32.60	38.38	70.90	5.92	32.66	20.49	71.68	8.00	43.09	43.09	12.83
DeepLabV3+ Chen et al. (2018)	78.10	32.24	**50.20**	55.01	**8.10**	22.12	24.60	70.24	**46.37**	58.25	49.37	5.39
OCRNet Yuan et al. (2020)	80.52	30.92	44.73	67.68	4.76	33.61	18.47	72.17	24.00	58.03	48.74	9.34
HRNet Wang et al. (2021a)	83.29	**34.47**	43.55	55.72	6.39	34.37	26.34	73.72	14.62	57.98	50.26	10.40
BisenetV1 Yu et al. (2018)	60.95	22.67	45.05	53.97	3.46	0.03	15.22	38.34	0.00	21.81	30.46	0.00
BisenetV2 Yu et al. (2021)	74.69	28.39	33.70	50.45	3.83	11.81	18.03	40.56	0.00	39.19	38.25	4.05
ICNet Zhao et al. (2018)	75.73	31.00	29.03	57.84	2.23	2.52	3.07	48.14	0.00	29.48	29.20	0.00
MobileV3 Howard et al. (2019)	73.39	30.79	30.23	62.88	4.29	9.80	18.52	72.51	0.00	52.81	54.81	0.00
SETR Si et al. (2020)	57.96	16.17	27.80	52.87	0.10	0.00	8.89	40.63	0.00	24.09	11.99	0.00
SegFormer Xie et al. (2021)	73.24	33.22	14.93	40.21	4.84	0.16	14.51	50.86	0.00	37.96	29.19	0.00
Segmentor Strudel et al. (2021)	79.09	30.28	35.95	60.66	0.32	0.00	7.74	50.11	0.00	34.77	39.64	0.00
Swin Liu et al. (2021)	77.51	34.39	42.92	58.88	6.53	36.48	28.66	**77.35**	41.84	**70.20**	62.03	15.16
GCNet Yang et al. (2025)	**89.25**	23.96	25.60	**72.45**	0.00	**35.47**	**27.05**	64.52	29.75	**70.20**	**64.45**	0.00
Average IoU	**75.68**	**29.77**	**36.83**	**59.06**	**4.14**	**19.03**	**18.39**	**61.08**	**13.60**	**48.31**	**42.94**	**5.66**

Table 2: Performance comparison on different datasets. '-' indicates model training on the corresponding dataset not reported.

DatasetsModels	FCN	PSPNet	DeepLabV3	DeepLabV3+	OCRNet	HRNet	BisenetV1	BisenetV2	SETR	SegFormer	Average mIoU
Cityscapes Cordts et al. (2016)	72.25	77.85	79.09	79.61	77.72	77.19	74.44	73.21	78.10	76.54	76.60
ADE20K Zhou et al. (2017)	**35.94**	41.13	42.42	42.72	**37.79**	36.27	-	-	48.28	37.41	40.24
Pascal VOC Everingham et al. (2010)	67.08	76.78	76.17	75.93	74.75	72.30	-	-	-	-	73.84
Pascal Context Mottaghi et al. (2014)	44.43	46.60	46.55	47.30	-	45.14	-	-	-	-	46.00
COCO-Stuff 164k Tsung et al. (2014)	-	38.80	39.38	-	-	-	**25.45**	-	-	-	34.54
ASS	37.62	**37.05**	**37.32**	**37.62**	38.93	**20.27**	26.44	**15.52**	**22.79**	**20.39**	**29.40**

table, baseline performance on ASS is substantially lower than on the other datasets. For instance, compared with Cityscapes and Pascal VOC, the average mIoU on ASS drops by over 40 percentage points. These results clearly indicate that algorithms excelling in general research settings fall short of the demands of real-world applications such as airport surveillance. Hence, there is a pressing need to construct dedicated airport segmentation datasets and develop algorithms tailored to such scenarios.

4.2 Future Research Discussion

Previous experiments show that standard semantic segmentation performs poorly in airport scenes. Expanding ASS with more training samples is a direct solution, but manual annotation is costly—labeling ASS required over ten students and several months. We propose an imprecise supervision approach for airport semantic segmentation, using samples with coarse or no annotations. Such labels are easy to acquire; e.g., unlabeled airport imagery can be collected via fixed cameras. Imprecise supervised segmentation includes unsupervised, semi-supervised, and weakly supervised learning Shen et al. (2023).

Unsupervised semantic segmentation. Unsupervised semantic segmentation relies solely on unlabeled images Zhang et al.. Supervisory signals are derived via strategies like contrastive learning, where augmented views of the same image are positive pairs and different images are negative. In airports, intra-class similarity can provide additional cues. For example, airplanes with identical motion poses across images serve as positive samples, while differing poses serve as negative ones, enhancing contrastive learning (Fig. 9(a)).

Semi-supervised semantic segmentation. Semi-supervised semantic segmentation relies on a small subset of pixel-wise labeled data, while most remain unlabeled Li et al. (2026). A teacher model trained on labeled data generates pseudo-labels for unlabeled samples to train a student model. Pseudo-label quality is critical. For airports, domain-specific priors can enhance accuracy. For instance, airports have consistent layouts—airplanes appear only on runways, taxiways, and aprons—reducing label noise. Moreover, airplane motion follows three modes: straight at constant speed, straight with acceleration, and turning at constant speed, with transitions. A Markov state model captures these dynamics, and the resulting motion model can produce more precise pseudo-labels in video semantic segmentation (Fig. 9(b)).

Weakly supervised semantic segmentation. Weakly supervised semantic segmentation relies on non-pixel-level training labels Hanna & Borth (2025). Since image-level labels (which classes are present) are easiest to obtain, we focus on this setting. The process typically has three stages: generating seed regions via Class Activation Maps (CAM), propagating seeds to produce pseudo-labels across the image, and training a segmentation model on these pseudo-labels. As in semi-

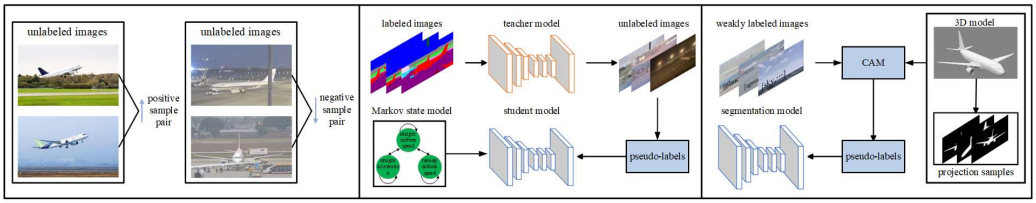

Figure 9: Illustration of three types of imprecise supervised airport semantic segmentation.

supervised learning, the main challenge is pseudo-label noise. Thus, prior knowledge can also benefit weakly supervised airport segmentation. For instance, shape priors for movable objects like airplanes are available online, and incorporating them into CAM and pseudo-label generation can effectively mitigate noise, as shown in Fig. 9(c).

Implementation. In this section, we adopt the weakly supervised paradigm to illustrate algorithm design for airport semantic segmentation under imprecise supervision. We propose SPWSeg, which uses image-level labels to provide supplementary supervisory signals. For such labels to be reliable, the labeled category must dominate each training image, reducing irrelevant content. However, airport scenes typically contain multiple semantic categories, making it difficult to collect sufficient real images meeting this criterion. To address this, we generate training samples with image-level labels. Specifically, we employ publicly available 3D shape models to generate such samples for categories with stable shapes across scenes—such as people and airplanes—whose geometry and appearance are well captured by these models. In contrast, categories like sky, ground, and lawn lack consistent shapes and are therefore unsuitable for this generation strategy.

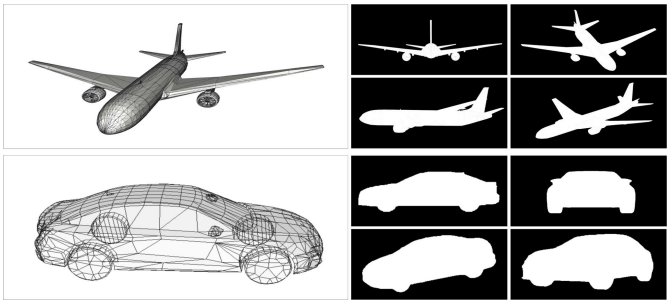

Figure 10: 3D shape models and projection samples for the airplane and vehicle.

The generation of training samples and the network design are described as follows. First, multi-view projections (e.g., varying degrees of top-down, side, and front views) are performed on the 3D shape models to obtain a large number of binary 2D projection samples. These projection samples serve directly as the image-level labeled training data. The top row of Fig. 10 illustrates the 3D shape model of the airplane and its corresponding projection samples. Second, we extract shape features from these projection samples to construct a shape prototype library $\{P_i^j\}$, where P_i^j denotes the i-th shape prototype for class j. Based on a shape feature extraction function $\phi(\cdot)$ implemented via a pre-trained shape autoencoder Li et al. (2023b), each binary projection sample is mapped into a shape feature space. K-means clustering is then applied to these projected features, and the features at the cluster centers are defined as the shape prototypes. Then, the loss item E_2 is formulated based on these shape prototypes:

$$E_2 = \sum_j \min_i ||\phi(S^j) - p_i^j||^2, \tag{1}$$

where S^j represents the semantic segmentation mask for the j-th category. By minimizing the loss item E_2, the segmentation results are enforced to align with the shape prototypes, thereby effectively leveraging the additional supervision provided by the image-level labeled samples. The total loss function of the network is defined as:

$$E_{\text{total}} = E_1 + \lambda E_2, \tag{2}$$

where E_1 denotes the conventional loss item based on CAM Zhou et al. (2016), and λ is a weighting factor. Finally, we employ the Expectation-Maximization algorithm to optimize



Figure 11: From left to right, original images from ASS and Cityscapes, by SPWSeg, and by Ahn et al. (2019).

the objective function E_{total}. The hyperparameter setting and pseudo-code of SPWSeg are presented in Section B.3 of the supplementary materials, and the source code is available at https://github.com/SPWSeg/SPWSeg.

The experimental results of SPWSeg on ASS are presented in Fig. 11 and Tab. 3. The comparative method Ahn et al. (2019) is a weakly supervised semantic segmentation approach. Its difference from SPWSeg lies in the absence of the E_2 item in the loss function, making its image-level labeled samples imprecise. It can be observed that, compared to Ahn et al. (2019), SPWSeg brings a significant performance improvement of about 6.3 percentage points. This experiment clearly confirms the effectiveness of the research approach of imprecise supervision for airport semantic segmentation. By comparing Tab. 1 and Tab. 3, we observe that for airport semantic segmentation, the imprecise supervision method has lower accuracy than the fully supervised method.Therefore, rather than replacing the latter with the former, we should combine both to further improve performance—this will be our future work.

The SPWSeg method is initially applicable only to airport semantic segmentation, as it relies on shape models specific to airport-related semantics. However, given that similar shape models exist for other semantic categories across different scenes, SPWSeg can be generalized to other domains or datasets. To verify this, we selected the vehicle category from the Cityscapes dataset, constructed a vehicle-specific shape prototype library, and adapted SPWSeg to Cityscapes accordingly. The 3D vehicle model and its corresponding 2D projection samples are shown in the bottom row of Fig. 10. Both quantitative and qualitative results on Cityscapes are presented in Fig. 11 and Tab. 3. These results clearly demonstrate that the proposed imprecise supervision approach for semantic segmentation is also effective in other scenes.

Table 3: Semantic segmentation performance of the SPWSeg method on ASS and Cityscapes.

Models	**IRNet**		**SPWSeg**	
Datasets	**Cityscapes**	**ASS**	**Cityscapes**	**ASS**
Results	14.22	11.33	22.06	17.62

5 Conclusion

We introduce ASS, the first large-scale dataset for airport semantic segmentation. It contains over 250 videos, 18 categories, and over 140,000 pixel-level annotated frames, posing challenges such as extreme scale variation, intra-class diversity, inter-class similarity, and complex environments. We evaluate 20 state-of-the-art image/video semantic segmentation models, and the performance drops significantly, indicating that general-purpose models still fall short of practical deployment. We then discuss future directions, propose an imprecise supervision-based approach for airport semantic segmentation, and verify its effectiveness and cross-scene generalization; combining it with other learning paradigms is our future direction.

REFERENCES

Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2209–2218, 2019.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. A dataset for semantic segmentation of point cloud sequences. *arXiv preprint arXiv:1904.01416*, 2(3):11, 2019.

G.J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Lett.*, 30(2):88–97, 2009.

L.C. Chen, G. Papandreou, and F. Schroff. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, pp. 801–818, 2018.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

M. Cordts, M. Omran, S. Ramos, M. Enzweiler T. Rehfeld, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3213–3223, 2016.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13614–13624, 2025.

M. Everingham, L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88:303–338, 2010.

Xu Han, Chong Liu, Yuzhou Zhou, Kai Tan, Zhen Dong, and Bisheng Yang. Whu-urban3d: An urban scene lidar point cloud dataset for semantic instance segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209:500–513, 2024.

Joëlle Hanna and Damian Borth. Know your attention maps: Class-specific token masking for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23763–23772, 2025.

Syed Ariff Syed Hesham, Yun Liu, Guolei Sun, Henghui Ding, Jing Yang, Ender Konukoglu, Xue Geng, and Xudong Jiang. Exploiting temporal state space sharing for video semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24211–24221, 2025.

A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *IEEE Int. Conf. Comput. Vis.*, pp. 1314–1324, 2019.

Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1094–1104, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, and Chris Choy. Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14089–14101, 2025.

Minghan Li, Shuai Li, Xindong Zhang, and Lei Zhang. Univs: Unified and universal video segmentation with prompts as queries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3227–3238, 2024.

Wangkai Li, Rui Sun, Zhaoyang Li, and Tianzhu Zhang. Towards robust pseudo-label learning in semantic segmentation: An encoding perspective. *Advances in Neural Information Processing Systems*, 38:110068–110107, 2026.

Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube framework for universal video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13923–13933, 2023a.

Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. Gin: Generative invariant shape prior for amodal instance segmentation. *IEEE Transactions on Multimedia*, 26:3924–3936, 2023b.

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 *IEEE Int. Conf. Comput. Vis.*, pp. 10012–10022, 2021.

J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4133–4143, 2021.

Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21033–21043, 2022.

R. Mottaghi, X. Chen, X. Liu, N.G. Cho, S.W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 891–898, 2014.

Viet-Anh Nguyen. Anylabeling: Auto labeling tool for computer vision. https://github.com/vietanhdev/anylabeling, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 640–651, 2017. doi: 10.1109/TPAMI.2016.2572683.

Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Qi Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):9284–9305, 2023.

X. Si, C. Jia, S. Heng, T. Xia, K. Ze, B. Ya, W. Yan, F. Jian, X. Tao, Philip H. S. T., and Z. Li. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 6877–6886, 2020. doi: 10.1109/CVPR46437.2021.00681.

R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE Int. Conf. Comput. Vis.*, pp. 7262–7272, 2021.

G. Sun, Y. Liu, H. Ding, T. Probst, and L. Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3126–3137, 2022a.

Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. In *European Conference on Computer Vision*, pp. 522–539. Springer, 2022b.

Guolei Sun, Yun Liu, Henghui Ding, Min Wu, and Luc Van Gool. Learning local and global temporal contexts for video semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13171–13182, 2024b.

Y. L. Tsung, J. B. Michael, M.and Serge, D. B. Lubomir, B. G. Ross, H. James, P. Pietro, R. Deva, D.ár Piotr, and C. Lawrence Z. Microsoft COCO: common objects in context. *Comput. Vis. – ECCV 2014*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 3349–3364, 2021a. doi: 10.1109/TPAMI.2020.2983686.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021b.

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.*, pp. 12077–12090, 2021. doi: 10.48550/arXiv.2105.15203.

Guoyu Yang, Yuan Wang, Daming Shi, and Yanzhong Wang. Golden cudgel network for real-time semantic segmentation. *arXiv preprint arXiv:2503.03325*, 2025.

C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Eur. Conf. Comput. Vis.*, pp. 325–341, 2018.

C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.*, 129:3051–3068, 2021.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. In *Eur. Conf. Comput. Vis.*, pp. 173–190. Springer, 2020.

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing correlations in clip with off-the-shelf foundation models for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.10086*, 2024.

Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023.

Xiao Zhang, Xiangyu Han, Xiwen Lai, Yao Sun, Pei Zhang, Xia Liu, and Konrad Kording. Falcon: Fast proximal linearization of normalized cuts for unsupervised image segmentation. In *The Fourteenth International Conference on Learning Representations*.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2881–2890, 2017.

H. Zhao, X. Qi, X. Shen, X. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Comput. Vis. – ECCV 2018*, pp. 418–434, 2018. doi: 10.1007/978-3-030-01219-9_25.

B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 633–641, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.