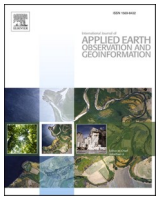
International Journal of Applied Earth Observation and Geoinformation 137 (2025) 104440

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A SAM-adapted weakly-supervised semantic segmentation method constrained by uncertainty and transformation consistency

Yinxia Cao [a,b,d], Xin Huang [c], Qihao Weng [a,b,d,*]

[a] JC STEM Lab of Earth Observations, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] Research Centre for Artificial Intelligence in Geomatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[c] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[d] Research Institute of Land and Space, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Segment anything model (SAM)
Weakly supervised learning
Uncertainty
Transformation consistency
Semantic segmentation

ABSTRACT

Semantic segmentation of remote sensing imagery is a fundamental task to generate pixel-wise category maps. Existing deep learning networks rely heavily on dense pixel-wise labels, incurring high acquisition costs. Given this challenge, this study introduces sparse point labels, a type of cost-effective weak labels, for semantic segmentation. Existing weakly-supervised methods often leverage low-level visual or high-level semantic features from networks to generate supervision information for unlabeled pixels, which can easily lead to the issue of label noises. Furthermore, these methods rarely explore the general-purpose foundation model, segment anything model (SAM), with strong zero-shot generalization capacity in image segmentation. In this paper, we proposed a SAM-adapted weakly-supervised method with three components: 1) an adapted EfficientViT-SAM network (AESAM) for semantic segmentation guided by point labels, 2) an uncertainty-based pseudo-label generation module to select reliable pseudo-labels for supervising unlabeled pixels, and 3) a transformation consistency constraint for enhancing AESAM's robustness to data perturbations. The proposed method was tested on the ISPRS Vaihingen dataset (collected from airplane), the Zurich Summer dataset (satellite), and the UAVid dataset (drone). Results demonstrated a significant improvement in mean F1 (by 5.89 %–10.56 %) and mean IoU (by 5.95 %–11.13 %) compared to the baseline method. Compared to the closest competitors, there was an increase in mean F1 (by 0.83 %–5.29 %) and mean IoU (by 1.04 %–6.54 %). Furthermore, our approach requires only fine-tuning a small number of parameters (0.9 M) using cheap point labels, making it promising for scenarios with limited labeling budgets. The code is available at https://github.com/lauraset/SAM-UTC-WSSS.

1. Introduction

Semantic segmentation of remote sensing imagery aims to assign one category to each pixel and has been widely studied in various applications, e.g., land cover classification (Chen et al., 2015; Myint et al., 2011), change detection (Bruzzone and Bovolo, 2013; Tan et al., 2016), built-up area extraction (Huang and Zhang, 2012; Pesaresi et al., 2008), tree mapping (Guan et al., 2015; Sylvain et al., 2024), and crop classification (Ajadi et al., 2021; Liu et al., 2020). High-resolution optical remote sensing imagery offers rich spatial details that can make small objects, such as cars and airplanes, visible. However, this imagery exhibits high heterogeneity and complexity, due to factors like varying illuminations, scales, and viewing angles, posing significant challenges for accurate semantic segmentation.

Semantic segmentation algorithms have evolved through three phases so far: 1) hand-crafted feature extraction; 2) specific deep learning network design; and 3) general-purpose foundation model development. The first phase focused on manually designing features with domain knowledge (Huang and Zhang, 2013; Tuia et al., 2009). Although these features have a high interpretability, they often perform poorly in complex scenarios, due to the customized perspective. This issue can be mitigated by specific deep learning networks, e.g., UNet (Ronneberger et al., 2015) and SegFormer (Xie et al., 2021). However, these networks were trained on small datasets customized for specific tasks, limiting their generalization ability. Recently, general-purpose foundation models have received great concerns, due to their strong

* Corresponding author at: JC STEM Lab of Earth Observations, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
E-mail address: qihao.weng@polyu.edu.hk (Q. Weng).

https://doi.org/10.1016/j.jag.2025.104440
Received 7 November 2024; Received in revised form 15 February 2025; Accepted 20 February 2025
Available online 25 February 2025

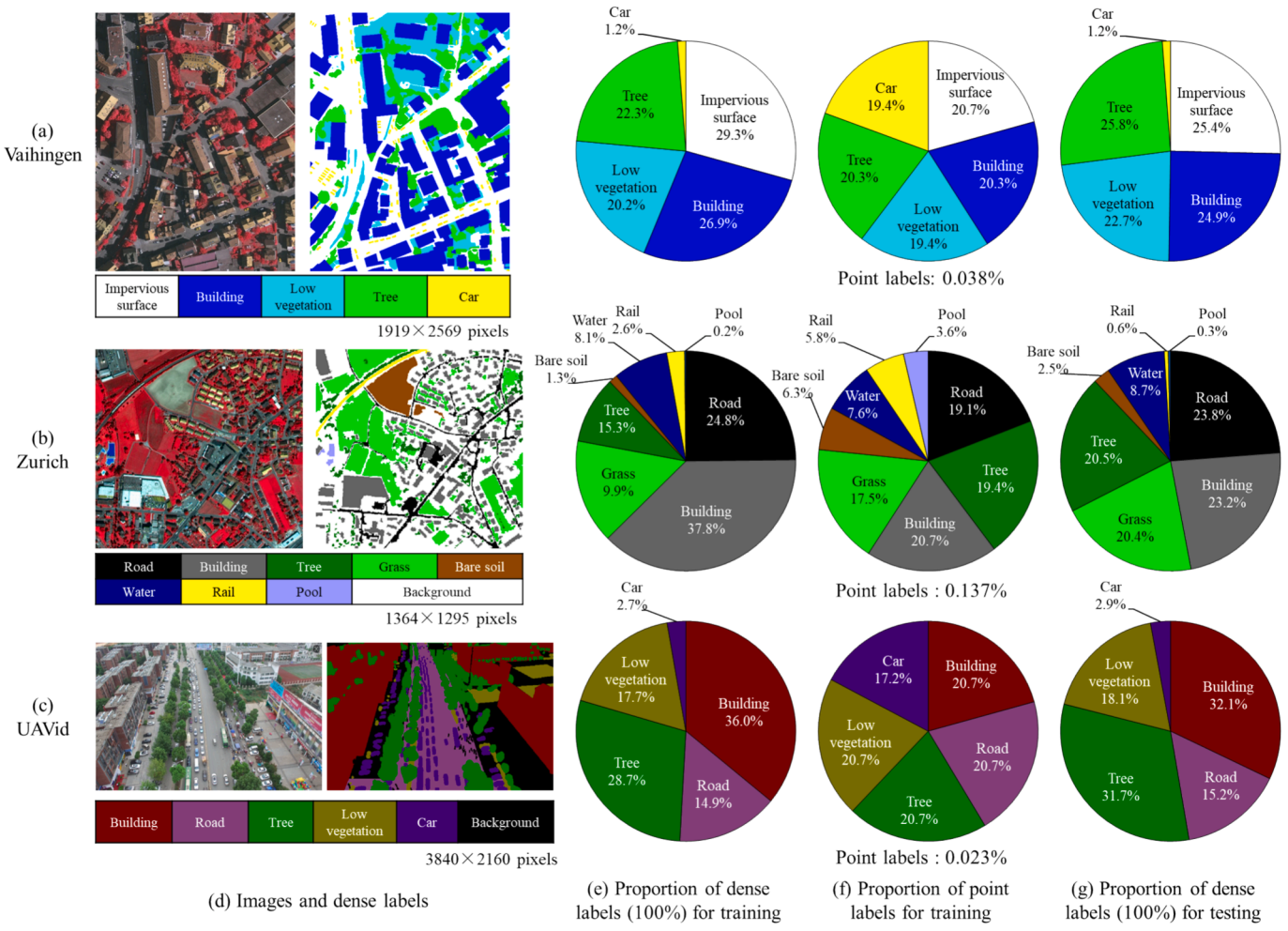


Fig. 1. Distribution of dense and sparse point labels on three datasets (a-c).

generalization ability (Hong et al., 2024). For instance, the segment anything model (SAM) (Kirillov et al., 2023), a foundation model trained with over 1 billion masks from 11 million natural images, exhibits impressive zero-shot capability in segmenting images into independent objects. However, SAM requires huge computational resources for training due to its heavy encoder. Therefore, multiple variants of SAM have been developed to improve efficiency, e.g., MobileSAM (Zhang et al., 2023), FastSAM (Zhao et al., 2023), and EfficientSAM (Xiong et al., 2024). These variants can accelerate the prediction procedure of image segmentation and can be applied to different devices with different computational abilities. In remote sensing, SAM holds great potential due to the similarity of natural and remote sensing images, e.g., similar objects, textures, and backgrounds. However, SAM cannot perform well on complex or specific remote sensing scenes, due to differences in observing angles, spatial resolutions, and spectral ranges. Moreover, to achieve semantic segmentation, they require dense pixel-level labels, which have high acquisition costs.

Weak labels, e.g., points, scribbles, and bounding boxes, require much lower labeling costs than pixel-level labels (Yue et al., 2022). This study focuses on point labels. Point labeling is simpler, less subjective, and more flexible than other forms of weak labels, since it only entails marking a single point and there is no need to strictly depict the shape of objects. However, the sparse distribution of point labels would reduce the generalization ability of deep networks, which depends heavily on the number of labels. Therefore, existing weakly-supervised studies utilize unlabeled pixels to supervise networks by three strategies: 1) regularized loss (Hua et al., 2021; Tang et al., 2018), 2) consistency learning (Xu and Ghamisi, 2022; Yang et al., 2023), and 3) pseudo-labeling (Li et al., 2024; Liang et al., 2022; Wu et al., 2023; Zhang et al., 2022). Regularized loss mostly relies on low-level vision information (e.g., color), which is often unreliable given the complexity and heterogeneity of high-resolution remote sensing images. Consistency learning usually forces high-level features to be consistent under different data disturbances, but it tends to ignore the explicit category-level information that can increase the number of labeled pixels. By contrast, pseudo-labeling can mitigate these shortcomings by combining low- and high-level features to generate category-level pseudo-labels for supervision. However, pseudo-labeling can easily suffer from confirmation bias (Arazo et al., 2020), caused by the error accumulation at each iteration of deep networks. Moreover, existing weakly-supervised methods are mostly built on task-specific deep networks, while few explore the effectiveness of foundation models.

To mitigate existing issues, this study proposes a new weakly-supervised semantic segmentation method built on one foundation model and two modules. First, considering that most weakly-supervised studies retrain task-specific networks, which would cost lots of time and effort, we design an adapted EfficientViT-SAM model (AESAM) that can reuse the weight of a foundation model, EfficientViT-SAM (Zhang et al., 2024), and only requires finetuning a few parameters (0.9 M). Second, given that pseudo-labeling can easily suffer from error accumulation of networks, we propose an uncertainty-based pseudo-label generation module. This module can leverage the model's predictions on unlabeled data as pseudo-labels (not real annotations from human interpretation), and then according to the uncertainty of predictions, select high-

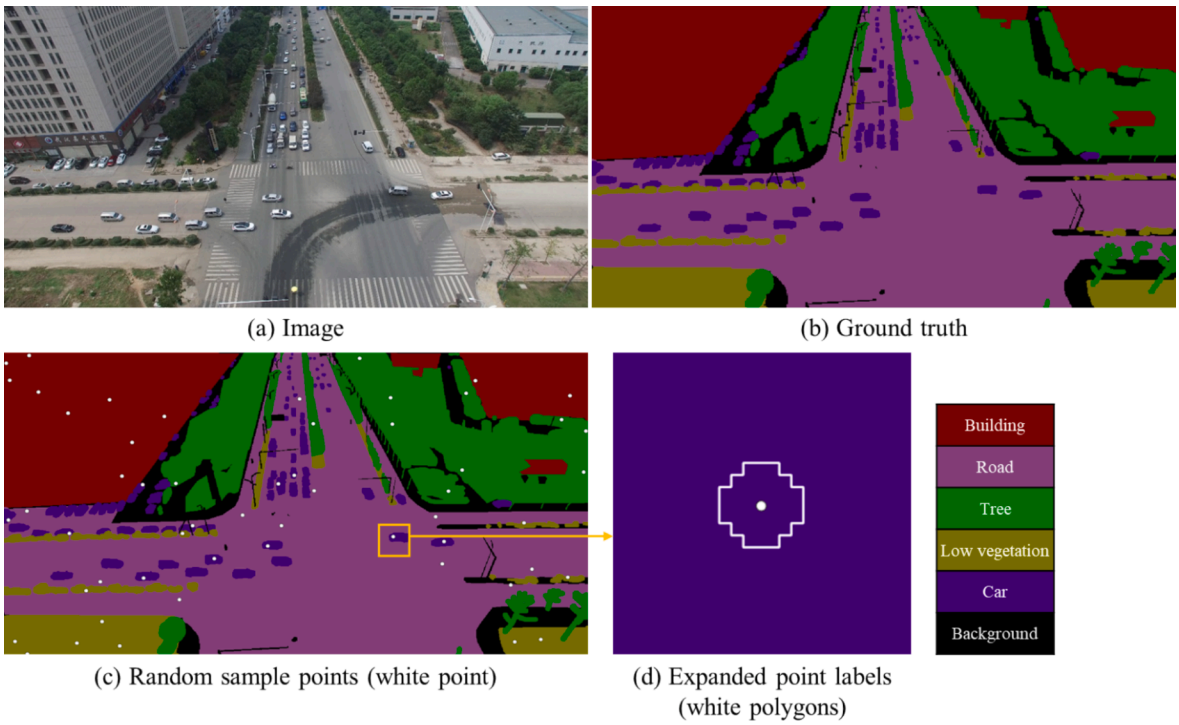


Fig. 2. Illustration of point label generation for the UAVid dataset.

confidence pseudo-labels as approximately correct labels to help train the model. Third, considering that remote sensing images can easily be affected by complex imaging conditions, a transformation consistency constraint is developed to force predictions of AESAM for the same image to be consistent under different disturbances (e.g., rotating and scaling). This module can reduce the sensitivity of AESAM to different data transformations, improving the robustness of network predictions. The effectiveness of the method was validated on three public datasets collected from airplane, satellite, and drone platforms, respectively.

2. Dataset description

Three public datasets (Fig. 1) include the ISPRS Vaihingen dataset,[1] the Zurich summer dataset (Volpi and Ferrari, 2015), and the UAVid dataset (Lyu et al., 2020), which were acquired from airplane, satellite, and drone platforms, respectively. All three datasets (Vaihingen, Zurich, UAVid) provide dense labels for both training (Fig. 1e) and test (Fig. 1g) sets. In contrast to fully-supervised methods that require complete labels, our weakly-supervised method relies solely on sparse point annotations (Fig. 1f), corresponding to 0.038 %, 0.137 %, and 0.023 % of the dense labels in each training set. The background category may contain noisy and irrelevant elements that would distract from learning meaningful features of actual target categories and therefore is excluded.

The ISPRS Vaihingen dataset. This dataset has 33 image patches, where each patch has a width and height of 2000 ~ 4000 pixels at a spatial resolution of 9 cm and includes near-infrared (NIR), red (R), and green (G) bands. It provides pixel-level labels with six categories, i.e., impervious surface, building, low vegetation, tree, car, and background. The point labels were obtained from (Hua et al., 2021) by manually selecting about 7 points for each class and expanding these points with a disk with a radius of 3 pixels for training. Following the previous work (Hua et al., 2021; Xu and Ghamisi, 2022), 11 images were used for training, and the other five images for testing.

The Zurich summer dataset. This dataset contains 20 satellite images. Each image has an extent of ~ 1000 × 1150 pixels with a spatial resolution of 0.61 m. NIR, R, and G bands were used to keep consistent with the Vaihingen dataset. Each image has been densely labeled by eight categories, i.e., road, building, tree, grass, bare soil, water, rail, and swimming pool. Following the previous work (Hua et al., 2021; Xu and Ghamisi, 2022), 15 images were selected for training and the remaining five images for testing.

The UAVid dataset. This dataset provides 42 sequences of oblique UAV videos. Each sequence consists of 10 frame images with a time interval of 5 s and 4 k resolution. Each image has a size of 4096 × 2160 pixels or 3840 × 2160 pixels and was labeled with eight classes: building, road, tree, low vegetation, static car, moving car, human, and background. We merged static and moving cars into one category, car, since its temporal variation is not the focus of this study. Besides, the human category was not considered given its rare proportion. Following the official guideline (https://uavid.nl), 20, 7, and 15 sequences were used for training, validation, and testing, respectively. One image for each sequence was randomly selected for experiments, given the high overlapping ratio of images within each sequence. To generate sparse point labels more efficiently and randomly, we revised the manually annotated method (Hua et al., 2021) to an automated version. As shown in Fig. 2, we randomly generated 10 points for each class and set the minimum distance between any two points to 100 pixels to reduce the spatial self-correlation of nearby pixels. Further, each point was expanded with a disk with a radius of 3 pixels to construct the training set.

3. Methodology

The proposed method is composed of three parts (Fig. 3): 1) an adapted EfficientViT-SAM (AESAM) trained with point labels for semantic segmentation, 2) an uncertainty-based pseudo-label generation module that fuses low- and high-level features from AESAM to generate reliable pseudo-labels for supervising unlabeled pixels, and 3) a transformation consistency constraint for keeping the consistency of the predictions of AESAM under different data transformations. The total

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx.

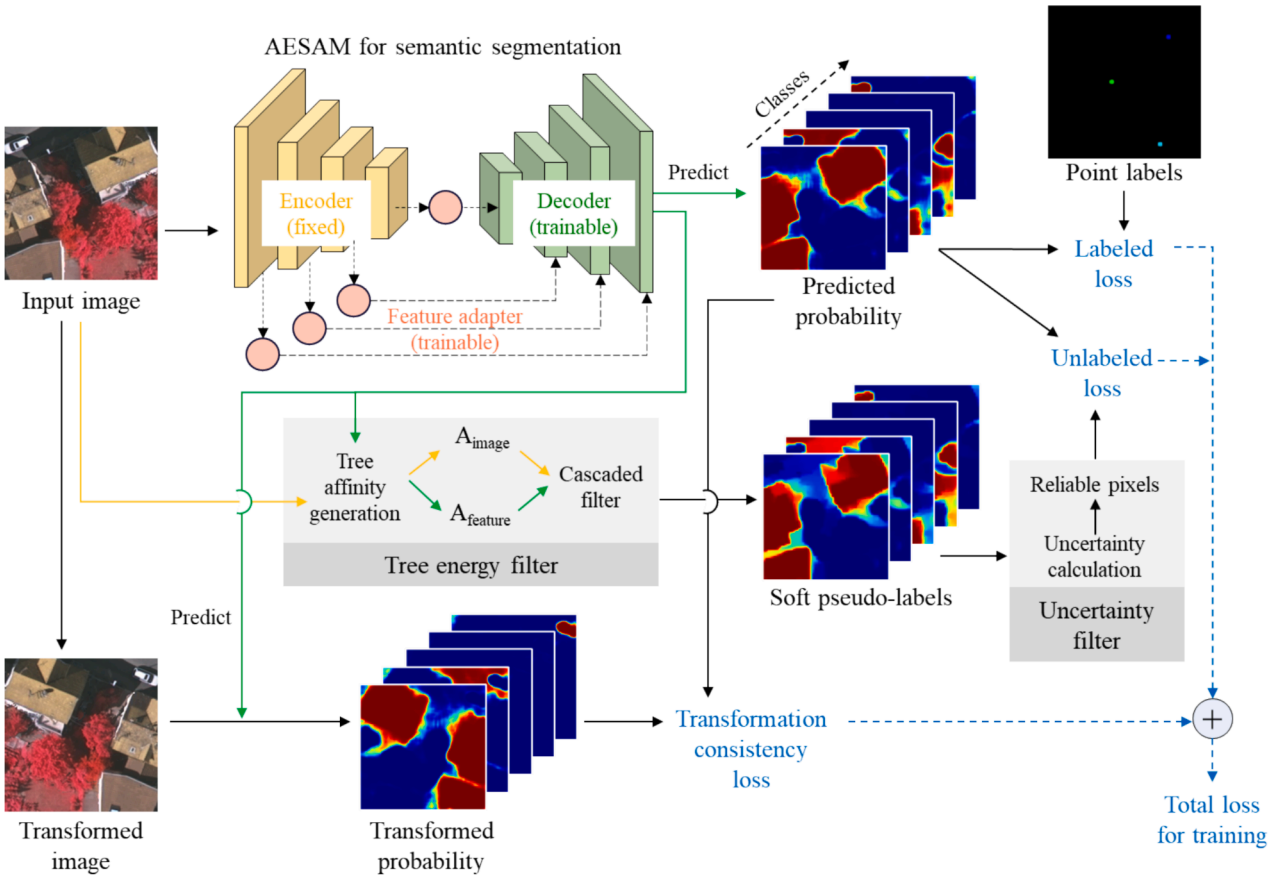


Fig. 3. Structure of the proposed SAM-adapted weakly-supervised method.

loss for training AESAM includes: 1) a labeled loss by comparing reference values of labeled pixels and predicted values, 2) an unlabeled loss by comparing pseudo-labels of unlabeled pixels and predicted values, and 3) a transform consistency loss by comparing predicted values on the input and transformed images.

3.1. AESAM for semantic segmentation

We developed AESAM to generate pixel-wise probability maps. AESAM consists of an encoder, a feature adapter, and a decoder. The encoder was set the same as the EfficientViT-SAM-L0 (Zhang et al., 2024). To exploit the prior knowledge, we fixed the weights of the encoder and applied a feature adapter, which acts as a bridge to transfer the prior knowledge learned by the encoder to remote sensing images by learning domain-specific features. The feature adapter is lightweight and capable of modeling non-linear complex relationships between input and output. It consists of one convolutional layer (Conv), batch normalization layer (BN), and ReLU layer (denoted as "Conv + BN + ReLU"). Feature maps were analyzed in Section 5.1. The decoder utilized feature maps at all scales and it contains three same blocks, where each consists of one transpose Conv and two "Conv + BN + ReLU". Point labels (denoted as y with the number of N_l) and the predicted probability (p) of AESAM were used to calculate a cross-entropy loss ($L_{labeled}$) across all classes (with the number of C):

$$L_{labeled} = -\frac{1}{N_l}\sum_{i=1}^{N_l}\sum_{c=1}^{C} y_{c,i}\log p_{c,i} \tag{1}$$

3.2. Uncertainty-based pseudo-label generation

We designed an uncertainty-based pseudo-label generation module to obtain pseudo-labels of unlabeled pixels for training AESAM. The module contains a tree energy filter (TEF) (Liang et al., 2022) and an uncertainty filter. In TEF, an image is treated as an undirected graph with vertices (V) denoting pixels and edges (E) denoting adjacencies. The image is first edge pruned to obtain a minimum spanning tree (MST) and then the distances between pixels and their nearby pixels in the MST are converted to a tree affinity map. The tree affinity map can represent the pair-wise similarity that pixels from the same objects tend to exhibit similar feature distribution. We obtained two tree affinity maps, one from the input image (with rich boundary information, denoted as A_{image}) and the other from high-level feature maps (with rich semantic information, denoted as $A_{feature}$) at the last layer of the decoder. The predicted probability by AESAM will be sequentially multiplied with A_{image} and $A_{feature}$ to generate soft pseudo-labels.

The original TEF (Liang et al., 2022) directly assigned soft pseudo labels for unlabeled pixels to optimize the network. However, noisy labels inevitably exist in pseudo-labels, which can easily misguide the network. To alleviate this issue, we introduced an uncertainty filter to gradually learn reliable parts of pseudo-labels to improve the robustness of the network to noises. First, for unlabeled pixels, we calculated the entropy (E) of the predicted probability (p) as uncertainty. The predicted probability was normalized to [0,1] for all classes (with the number of C) by the Softmax function. Therefore, the maximum value of E would achieve $\log(C)$ when pixels are assigned to each category with equal probability. Subsequently, we adopted a dynamic threshold (t) to select relatively reliable unlabeled pixels (R) with the number of N_r and used them to calculate an unlabeled loss ($L_{unlabeled}$):

$$E_i = -\sum_{c=1}^{C} p_{c,i}\log p_{c,i} \tag{2}$$

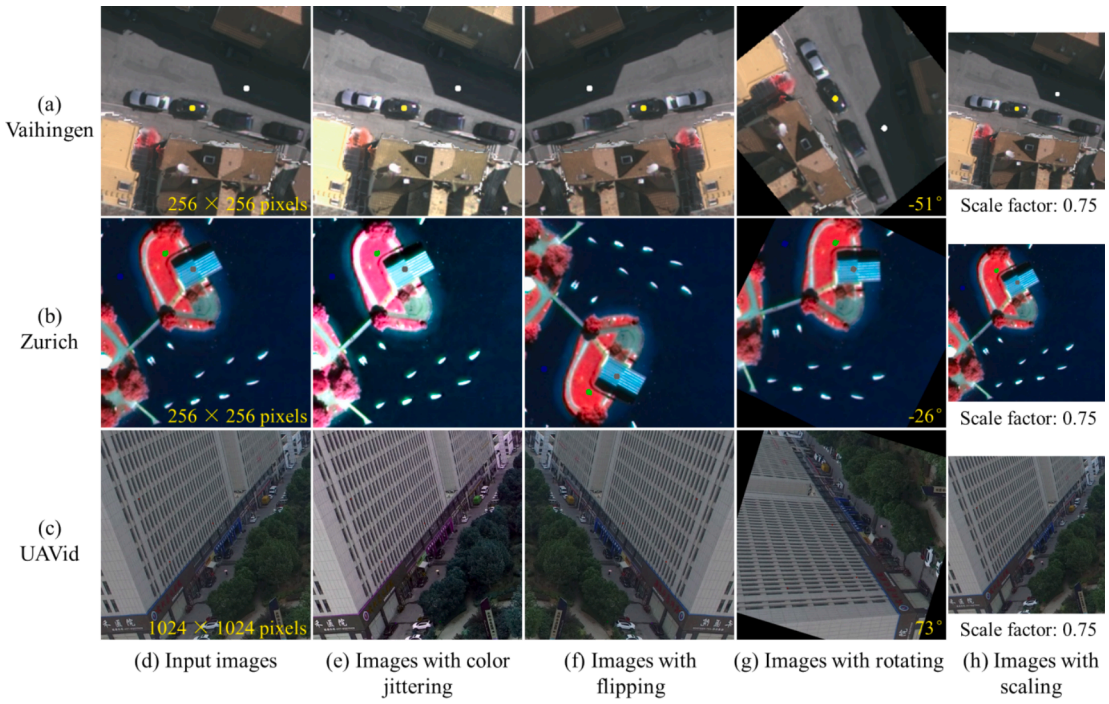


Fig. 4. Illustration of four image transformations including color jittering (e), flipping (f), rotating (g), and scaling (h). Input images (d) are overlayed with point labels.

$$R_i = (E_i < t), \text{where } t = \left[t_0 + (1 - t_0)\frac{iter}{iter\_max}\right] \bullet \log(C) \tag{3}$$

$$L_{unlabeled} = -\frac{1}{N_r}\sum_{i=1}^{N_r}\sum_{c=1}^{C} R_i \bullet \left|p_{c,i} - TEF(p_{c,i})\right| \tag{4}$$

where $TEF(p_{c,i})$ represents the soft pseudo-labels with the tree energy filter, *iter* denotes the current iteration step, and *iter_max* is the maximum iteration step. The parameter C in Eq. (3) was set to 2, considering that the network is more likely to confuse two similar classes (e.g., tree and grass). The parameter t_0 is the initial threshold in the range of [0,1] and was set to 0.9 in this study. The performance of the uncertainty filter and the sensitivity of t_0 were discussed in Section 5.2.

3.3. Transformation consistency constraint

We proposed a transformation consistency constraint module to improve the robustness of the network to noisy labels. We considered four transformations to simulate remote sensing images with different illuminations, viewing perspectives, and spatial scales. As illustrated in Fig. 4, four transformations include color jittering, flipping, rotating, and scaling. Color jittering is the random adjustment of brightness, contrast, saturation, and hue of an image, in the empirical range of [0.5, 1.5], [0.5, 1.5], [0.5, 1.5], and [-0.2, 0.2], respectively. Flipping in the horizontal and vertical directions was randomized. When rotating, the rotation angle was randomly generated in the range of [-90°, 90°]. For scaling, we set two scale factors, 0.75 and 1.5. To improve the efficiency of the network training, we transformed the input image with a random combination of four transformations, and the performance of these transformations was analyzed in Section 5.2. Note that color jittering was only performed on images, while other transformations on both images and labels. Then, we calculated a consistency loss (L_{transc}) between the prediction (p) of original images and that ($T(p)$) of transformed images and a cross-entropy loss (L_{transl}) between ground truth and predictions of transformed labeled pixels.

$$L_{transc} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(p_{c,i} - T(p_{c,i})\right)^2 \tag{5}$$

$$L_{transl} = -\frac{1}{N_l}\sum_{i=1}^{N_l}\sum_{c=1}^{C} T\left(y_{c,i}\right)\log T(p_{c,i}) \tag{6}$$

where N and N_l represent the number of all pixels and labeled pixels, respectively. The total loss (L_{total}) is the sum of labeled (Eq.(1)), unlabeled (Eq.(4)), and transformation consistency losses. Given the difficulty of determining the optimal weight for each term, these loss terms are treated equally:

$$L_{total} = L_{labeled} + L_{unlabeled} + L_{transc} + L_{transl} \tag{7}$$

3.4. Experimental settings

To train the proposed method, we adopted the stochastic gradient descent (SGD) optimizer with a weight decay of 5e-4 and a momentum of 0.9, following existing work (Xu and Ghamisi, 2022). The learning rate was updated using a polynomial function, i.e., $lr_0*(1\text{-}iter/max\_iter)^{0.9}$ at each iteration (*iter*), where the initial value (lr_0) was set to 0.1. The maximum iteration step (*max_iter*) was set to 5000. The model weights at the maximum iteration step were used for accuracy assessment, considering their relative stability. Image patches were set to the size of 256 × 256 pixels with a batch size of 32 for the Zurich and Vaihingen datasets, and 1024 × 1024 pixels with a batch size of 8 for the UAVid dataset, following the setting in (Wang et al., 2022). Data augmentation techniques were performed at each iteration to increase the image diversity, including random horizontal or vertical flipping, random rotation with the angle range of [-90°, 90°], and random shuffling of image grids (2 × 2 square grids). Experiments were conducted with a NIVIDA GeForce RTX 3080 GPU with 10 G memory. These settings remain fixed unless otherwise noted.

For accuracy assessment, we selected six measures, i.e., precision, recall, F1-score (abbrev. F1), mean F1, mean intersection over union (IoU), and overall accuracy (OA). A higher precision means a lower false

Table 1
Accuracy (%) of different methods on the Vaihingen dataset. The highest value for each column is shown in bold.

	F1 per class							
Method	Impervious surface	Building	Low vegetation	Tree	Car	mean F1	mean IoU	OA
Baseline	73.06	79.80	66.24	76.95	31.50	65.51	50.94	72.68
FESTA	73.73	77.98	64.08	76.53	33.46	65.16	50.30	72.05
CRGNet	74.22	78.78	**67.05**	79.27	35.83	67.03	52.38	73.86
TFCSD	68.94	84.90	61.01	74.60	63.62	70.61	55.28	72.41
TEL	69.09	**85.95**	60.28	73.86	64.71	70.78	55.53	72.40
DenseCRF	73.80	79.17	65.67	76.65	38.79	66.82	51.82	73.06
AGMM	72.79	78.82	66.16	72.56	42.63	66.59	51.15	71.95
Proposed	**82.18**	84.44	65.69	**81.39**	**66.66**	**76.07**	**62.07**	**79.02**
Oracle	88.13	93.29	75.62	85.05	79.76	84.37	73.46	85.76

Table 2
Accuracy (%) of different methods on the Zurich dataset. The highest value for each column is shown in bold.

	F1 per class										
Method	Road	Building	Tree	Grass	Bare Soil	Water	Rail	Pool	mean F1	mean IoU	OA
Baseline	81.31	88.03	87.2	81.08	54.04	95.24	10.12	62.75	69.97	58.95	83.29
FESTA	81.54	88.21	87.73	81.99	54.50	94.93	7.50	74.85	71.41	60.86	83.84
CRGNet	81.43	87.99	87.27	81.34	53.74	95.06	9.10	62.04	69.75	58.78	83.38
TFCSD	79.48	84.02	89.41	83.05	**65.89**	95.48	5.25	73.97	72.07	61.51	83.07
TEL	82.47	89.13	**90.64**	**88.53**	59.64	96.24	7.60	74.10	73.54	63.86	**86.37**
DenseCRF	81.84	87.91	88.47	83.08	53.4	95.75	12.96	62.16	70.69	59.79	84.09
AGMM	79.35	86.63	87.23	82.55	56.08	93.58	9.27	65.92	70.08	58.84	82.71
Proposed	**83.08**	**89.49**	87.58	82.19	55.58	**96.37**	**37.82**	**78.54**	**76.33**	**64.90**	85.17
Oracle	91.93	93.45	94.56	90.72	74.16	98.37	44.73	95.54	85.43	77.68	92.52

alarm rate, while a higher recall a lower missed alarm rate. F1 balances the precision and recall. IoU measures how well the detected object matches the real one. Mean F1 or IoU denotes the arithmetic mean values over all classes. OA represents the ratio of correctly detected pixels to all pixels over all classes. Compared to OA, F1 and IoU are more robust in case of class imbalance.

4. Results

To demonstrate the effectiveness of the proposed method, we compared it with two supervised methods (oracle and baseline) and six advanced weakly-supervised ones. Oracle uses all pixel-wise labeled labels for training, while baseline and weakly-supervised methods only point labels. Details of these methods are outlined below:

1) FESTA (Hua et al., 2021) designed an unsupervised loss to encode spatial- and feature-level relations for supervising unlabeled pixels;
2) CRGNet (Xu and Ghamisi, 2022) developed a dual-branch network, where one branch applied a region-growing algorithm to generate labels of unlabeled pixels for guiding another branch;
3) TFCSD (He et al., 2024) introduced a dual-branch network, where one branch was guided by the original sparse labels while the other by the soft pseudo-labels generated from TEL (Liang et al., 2022);
4) TEL (Liang et al., 2022) designed a cascade tree energy filter (see details in Section 3.2) to generate soft pseudo-labels for supervising unlabeled pixels;
5) DenseCRF (Tang et al., 2018) developed a denseCRF loss (Krähenbühl and Koltun, 2011) to introduce the similarity of low-level features of unlabeled pixels;
6) AGMM (Wu et al., 2023) applied the Gaussian Mixture Model to measure the similarity between labeled and unlabeled pixels for generating soft pseudo-labels as supervision signals.

These weakly-supervised methods were chosen because they utilize optical images and weak labels similar to our task. Methods like SQN (Hu et al., 2022), OCOC (Wang et al., 2023), or DAAL-WS (Lei et al., 2024), were excluded as they focus on point cloud data, requiring domain-specific priors beyond our scope. For a fair comparison, we use the same network (i.e., AESAM) as the backbone of these methods. FESTA and DenseCRF introduced regularized losses, which can easily suffer from the heterogeneity of images (e.g., spectral confusion). By contrast, the other four methods utilized consistency learning or pseudo-labeling techniques to generate pseudo-labels for unlabeled pixels, which can be easily affected by the error accumulation of deep networks. In this regard, the proposed method introduced uncertainty and transformation consistency constraints to mitigate the error

Table 3
Accuracy (%) of different methods on the UAVid dataset. The highest value for each column is shown in bold.

	F1 per class							
Method	Building	Road	Tree	Low vegetation	Car	mean F1	mean IoU	OA
Baseline	84.40	79.43	74.13	63.99	52.33	70.86	56.05	75.36
FESTA	83.64	79.84	74.41	63.80	53.35	71.01	56.16	75.21
CRGNet	84.86	81.38	75.61	64.35	54.11	72.06	57.52	76.42
TFCSD	86.53	79.44	75.20	62.96	51.98	71.22	56.69	76.12
TEL	88.32	**82.32**	76.91	**68.08**	63.96	75.92	62.03	79.38
DenseCRF	87.32	78.71	**78.20**	64.87	57.79	73.38	59.05	78.19
AGMM	84.38	81.29	73.99	64.37	49.07	70.62	56.03	75.39
Proposed	**90.23**	84.48	74.71	66.06	**68.29**	**76.75**	**63.23**	**79.48**
Oracle	92.94	89.23	83.66	72.89	76.64	83.07	71.75	85.48

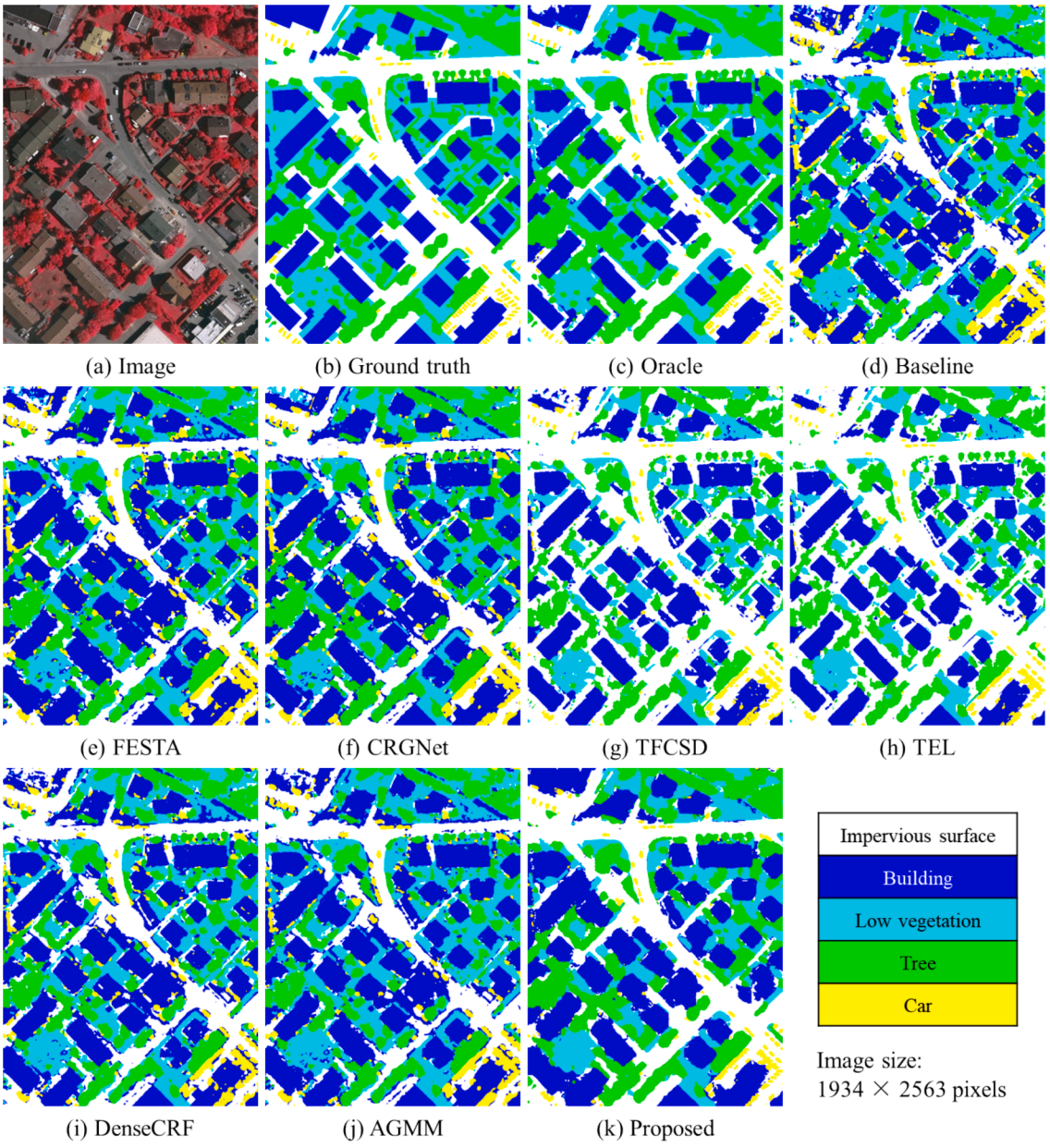


Fig. 5. Examples of predicted results by different methods on the Vaihingen dataset.

accumulation.

Tables 1–3 record the accuracy of different methods on three datasets. It can be observed that the proposed method significantly improved mean F1 (by 5.89 %-10.56 %), mean IoU (by 5.95 %-11.13 %), and OA (by 1.88 %-6.34 %) compared to the baseline method, and reduced the performance gap with the oracle method on three datasets. Compared to six weakly-supervised methods, the proposed one achieved the highest mean F1, mean IoU, and OA on the Vaihingen and UAVid datasets. On the Zurich dataset, the proposed method obtained slightly lower OA values but higher mean F1 and mean IoU, compared to the TEL. This phenomenon is mostly due to the extremely imbalanced distribution of classes (e.g., rails with low proportion, see Fig. 1), since mean F1 and mean IoU treat rare or major classes equally, whereas OA highlights major classes. This finding indicates that the proposed method can mitigate the imbalanced class distribution and can effectively utilize a large number of unlabeled pixels to improve the performance of land cover classification.

Figs. 5–7 show examples of predicted results with different methods on three datasets. Predicted results by the oracle method had the highest similarity to ground truth, while those by the baseline method exhibited lots of false alarms and omissions due to the low availability of labeled pixels. Among weakly-supervised methods, the proposed one achieved relatively good results for each category, while the other ones suffered from overestimation and underestimation to some extent. This phenomenon is mostly due to the fact that the proposed method introduced the uncertainty of pseudo-labels to retain reliable pixels (Section 3.2) and suppressed the confirmation bias of deep networks by the transformation consistency constraint (Section 3.3), improving the robustness of networks to different data disturbances.

5. Discussion

5.1. Sensitivity of network selection

To evaluate the sensitivity of the proposed method to different networks, we replaced AESAM with two advanced networks, UNetFormer

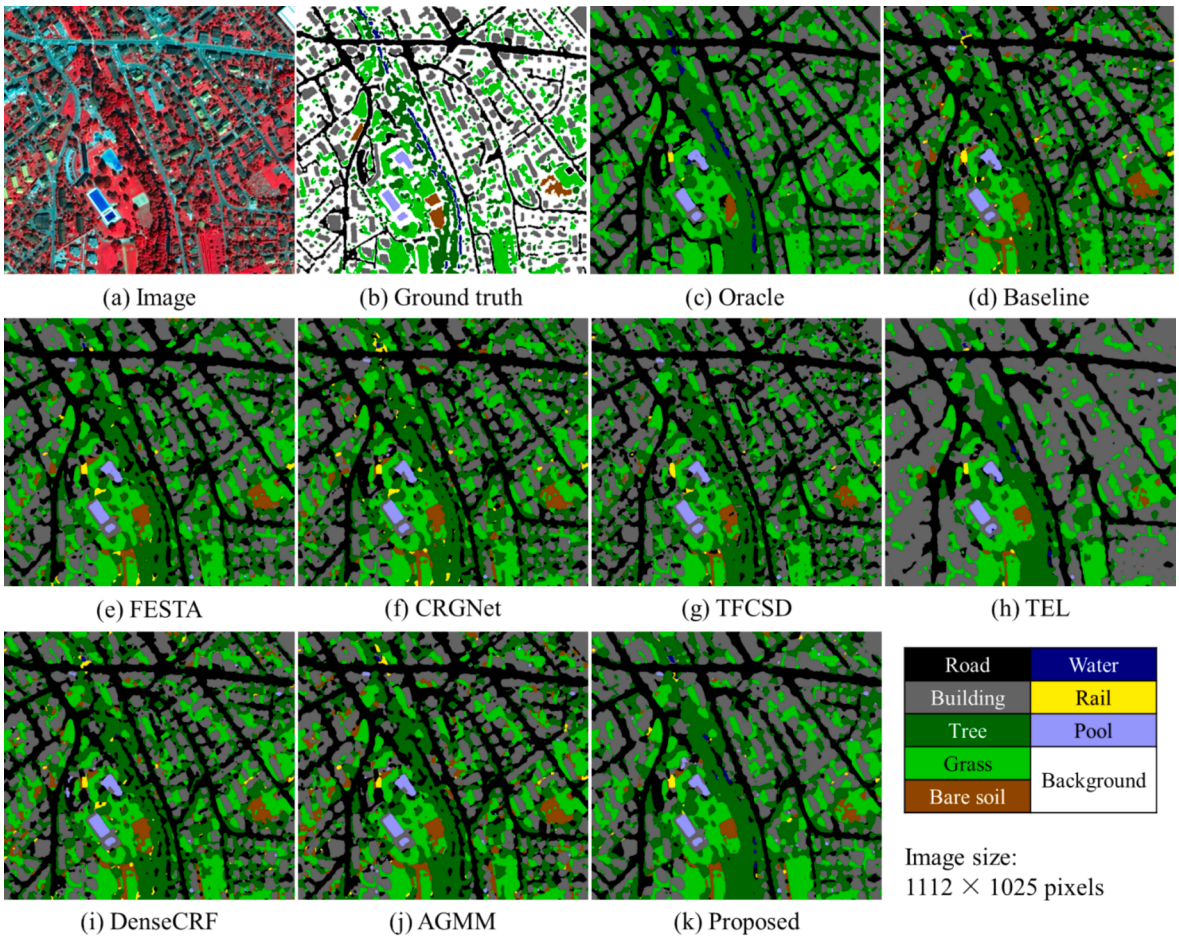


Fig. 6. Examples of predicted results by different methods on the Zurich dataset.

(Wang et al., 2022) and DeepLabv2-VGG16 (Chen et al., 2017). Unetformer consists of a CNN-based encoder (i.e., ResNet-18 (He et al., 2016)) and a transformer-based decoder with a global–local attention module. DeepLabv2-VGG16 consists of a CNN-based encoder (i.e., VGG-16 (Szegedy et al., 2015)) and a decoder with an atrous spatial pyramid pooling module. As shown in Table 4, we selected six metrics to assess the network complexity, including the number of multiply-accumulate operations (MACs), the number of floating-point operations (FLOPs), the number of total parameters, the number of trainable parameters, peak memory usage, and the inference speed measured by frames per second (FPS). It can be seen that UnetFormer achieves the fewest MACs, FLOPs, total parameters, and peak memory usage, DeepLabv2-VGG16 exhibits the highest inference speed, and AESAM has the fewest trainable parameters (0.9 M).

Tables 5-7 report the accuracies of two networks (UNetFormer and DeepLabv2-VGG16) on three datasets. For both networks, the proposed method exhibited advanced performances, especially in terms of mean F1 and mean IoU, demonstrating its robustness to the network structure and class imbalance. In addition, we found that AESAM (Tables 1-3) has the highest mean F1 and mean IoU, followed by UnetFormer, and DeepLabv2-VGG16 the least, on the Vaihingen and Zurich datasets, while UnetFormer obtained the highest mean F1 and mean IoU, followed by DeepLabv2-VGG16, and AESAM the least on the UAVid dataset. This phenomenon indicates that AESAM may not always perform better than task-specific networks. In comparison with the Vaihingen (from airplane platform) and Zurich (satellite) datasets, the UAVid dataset, which was obtained from drones, has higher spatial resolution and heterogeneity. As indicated by existing studies (Wang et al., 2022), this requires global context information to better distinguish different categories. UnetFormer introduces global–local attention modules to model global and local contexts, and DeepLabv2-VGG16 designs an atrous spatial pyramid pooling with different sampling rates to capture multi-scale contextual information. However, in AESAM, the feature adaptor and decoder mainly comprise convolution layers that are proficient in local feature extraction but cannot capture global contexts, resulting in a lower performance on the UAVid dataset. Therefore, there is still substantial room to improve the accuracy of AESAM, which merits further research. Moreover, AESAM can reuse most of pretrained parameters and only requires fine-tuning a few parameters (0.9 M in this study) using customized labels, making it promising for a range of downstream tasks.

To elaborate on the effect of fine-tuning AESAM, we compared original and adapted features (Fig. 8), which were obtained from AESAM before and after fine-tuning, respectively. The original feature maps did not provide any category information, and therefore applying them to downstream tasks requires semantic labels. The number of adapted feature maps was reduced by half to make them more compact, compared to original features. As displayed in Fig. 8, the original feature maps can capture some general information (e.g., boundaries of objects in graph (b)), without any fine-tuning, demonstrating the strong feature representation ability. After fine-tuning with semantic labels, adapted feature maps focused more on task-specific regions and tried to distinguish between different classes, significantly reducing the proportion of feature values (about 0.5) with high heterogeneity.

5.2. *Effectiveness of the proposed method*

To evaluate the effectiveness of the proposed method, we analyzed

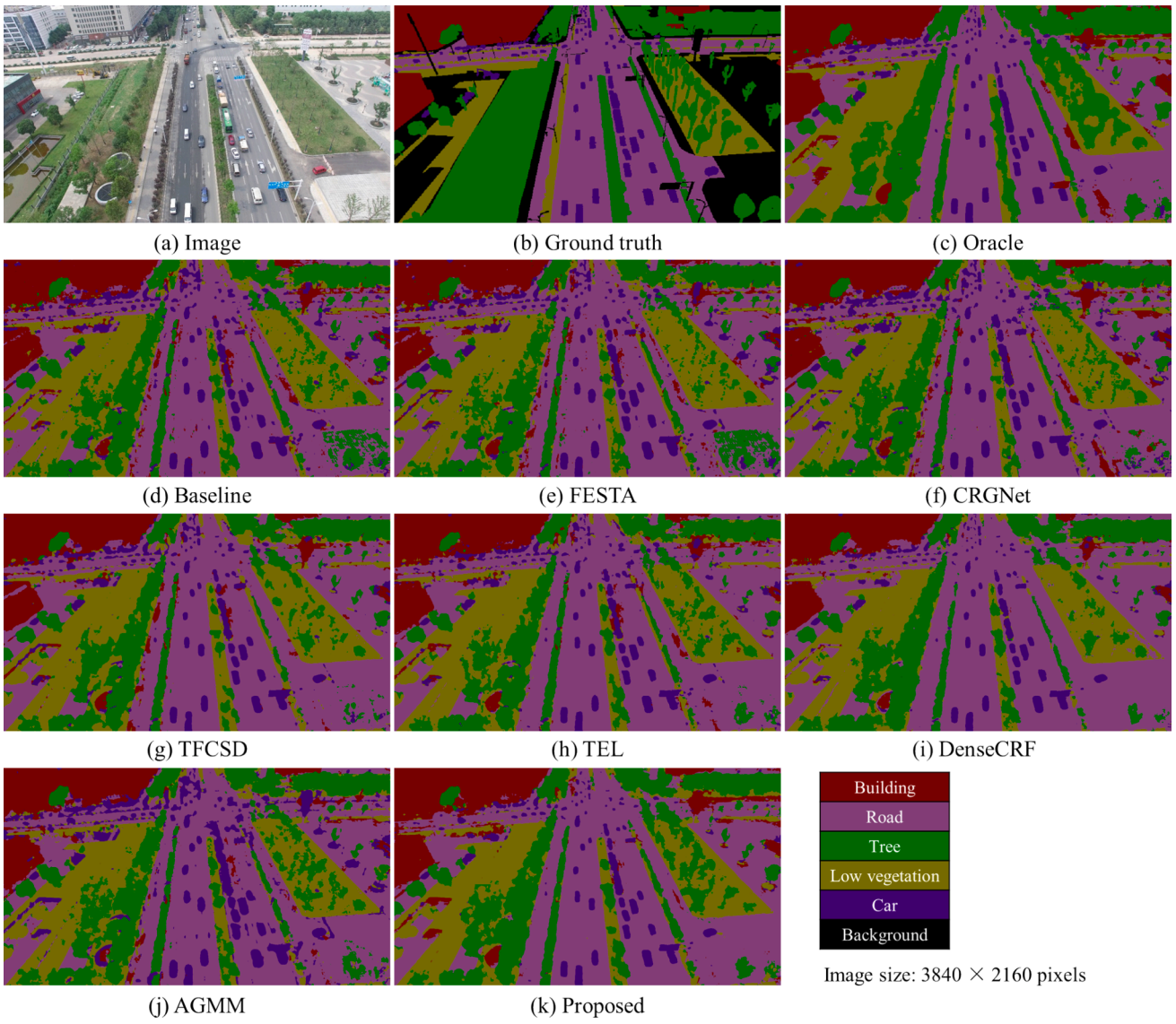


Fig. 7. Examples of predicted results by different methods on the UAVid dataset.

Table 4
Complexity comparison of the proposed method with UNetFormer and DeepLabv2-VGG16. The input image was set to 256 × 256 × 3 pixels and the number of predicted classes was set to five for complexity calculation. The best values for each column are in bold.

	MACs (G)	FLOPs (G)	Total parameters (M)	Trainable parameters (M)	Peak memory (M)	Speed (FPS)
Adapted SAM	6.4	12.8	35.6	**0.9**	279.4	257.1
UNetformer	**2.9**	**5.8**	**11.7**	11.7	**212.2**	216.9
DeepLabv2-VGG16	40.1	80.2	29.3	29.3	354.0	**515.5**

the contribution of two modules: 1) uncertainty-based pseudo-label generation (U), and 2) transformation consistency constraint (T). The proposed method without the two modules is the baseline method. As depicted in Fig. 9, the combination of the two modules achieved the highest accuracies (i.e., mean F1, mean IoU, and OA) across all categories in most cases. Besides, we can find that for rare classes (e.g., rail and car), the inclusion of U or T can significantly improve the F1 value, indicating that they can mitigate the class imbalance issue and the overconfidence of networks.

Furthermore, we assessed the sensitivity of the proposed method to different transformation types. We compared the single-use and the random combination of four transformation types: flipping, scaling, color jittering, and rotating. The specific parameters of these types were set the same as in Section 3.3. As shown in Fig. 10, the difference in accuracy of different transformation types varied across categories and datasets, but it was not very significant in most cases. In addition, the random combination did not always yield higher performance than the single-use on all the datasets (e.g., the Zurich dataset) and the optimal type was different for different datasets. Automatically selecting the optimal type requires more in-depth exploration and is beyond the scope of this study.

For the uncertainty filter, we analyzed the sensitivity of initial thresholds (i.e., t_0 in Eq.(3)). Since the number of categories with maximum uncertainty was set to 2 in this study, the maximum

Table 5
Accuracy (%) of different networks (UNetFormer and DeepLabv2-VGG16) on the Vaihingen dataset. The best values for each column are in bold.

Model	Method	F1 per class					mean F1	mean IoU	OA
		Impervious surface	Building	Low vegetation	Tree	Car			
UNetformer	Baseline	77.40	83.96	64.81	78.35	29.45	66.79	53.02	74.84
	FESTA	77.28	83.10	63.62	77.46	29.55	66.20	52.25	74.07
	CRGNet	78.86	86.23	66.53	79.13	33.28	68.81	55.24	76.51
	TFCSD	74.29	80.26	61.38	78.65	34.86	65.89	51.27	73.17
	TEL	80.04	**86.96**	**66.89**	74.11	54.60	72.52	58.07	76.87
	DenseCRF	79.05	84.73	65.40	78.57	35.99	68.75	54.82	76.16
	AGMM	76.47	81.71	65.50	78.67	39.98	68.47	53.90	75.06
	Proposed	**81.86**	85.07	66.36	**79.44**	**59.16**	**74.38**	**60.17**	**78.35**
	Oracle	87.31	91.84	73.67	84.58	71.29	81.74	69.87	84.61
Deeplabv2-VGG16	Baseline	76.50	**85.60**	66.17	78.88	27.38	66.91	53.44	75.16
	FESTA	75.56	83.85	65.91	77.50	33.73	67.31	53.12	74.76
	CRGNet	76.61	85.15	66.66	**79.22**	27.85	67.10	53.60	75.31
	TFCSD	68.21	84.43	62.83	72.26	38.12	65.17	50.15	71.07
	TEL	76.94	84.66	66.14	71.43	47.53	69.34	54.41	74.62
	DenseCRF	77.30	85.14	**67.30**	78.23	29.91	67.58	53.93	75.46
	AGMM	77.26	83.92	67.11	77.59	30.55	67.29	53.43	75.06
	Proposed	**80.23**	83.93	65.72	77.81	**54.96**	**72.53**	**57.96**	**77.18**
	Oracle	86.76	91.59	74.73	84.45	64.2	80.35	68.22	84.47

Table 6
Accuracy (%) of different networks (UNetFormer and DeepLabv2-VGG16) on the Zurich dataset. The best values for each column are in bold.

Model	Method	F1 per class								mean F1	mean IoU	OA
		Road	Building	Tree	Grass	Bare Soil	Water	Rail	Pool			
UNetformer	Baseline	78.21	83.51	91.01	86.69	50.49	95.44	7.39	66.75	69.94	59.36	83.27
	FESTA	77.30	83.74	**91.24**	**87.34**	52.52	95.16	8.40	69.58	70.66	60.07	83.70
	CRGNet	79.75	84.75	90.80	86.55	51.98	96.45	8.53	59.03	69.73	59.24	84.29
	TFCSD	78.13	82.68	90.66	86.82	**55.73**	95.34	8.83	67.94	70.77	60.00	83.56
	TEL	79.56	83.82	90.96	87.15	51.66	**97.15**	13.54	71.02	71.86	61.31	84.35
	DenseCRF	78.63	83.46	90.37	86.33	47.47	96.64	12.77	67.57	70.40	59.65	83.17
	AGMM	77.32	83.24	89.31	84.36	55.45	95.79	7.78	72.19	70.68	59.84	82.76
	Proposed	**81.10**	**87.92**	90.31	85.22	54.42	96.35	**29.17**	**73.01**	**74.69**	**63.52**	**85.62**
	Oracle	89.28	92.09	94.79	92.08	71.00	97.37	14.63	0.00	68.90	62.40	91.48
Deeplabv2-VGG16	Baseline	75.98	83.01	82.96	73.86	43.36	92.26	10.31	52.61	64.29	52.01	77.99
	FESTA	71.53	76.54	78.17	72.28	51.28	92.43	9.11	47.53	62.36	49.35	74.86
	CRGNet	**76.89**	**83.61**	84.37	78.66	**56.46**	94.02	5.29	54.55	66.73	55.04	80.44
	TFCSD	74.65	81.09	83.55	78.74	56.26	93.65	7.90	55.02	66.36	54.21	79.10
	TEL	76.21	82.34	85.41	**81.12**	51.24	92.93	16.10	55.94	67.66	55.39	80.34
	DenseCRF	76.28	83.08	84.32	78.17	50.39	92.91	9.19	52.90	65.91	53.87	79.64
	AGMM	73.57	80.39	83.09	74.93	38.57	92.35	13.51	57.39	64.23	51.69	76.85
	Proposed	75.48	82.76	**86.79**	80.81	48.23	**94.85**	**35.67**	**61.79**	**70.80**	**58.01**	**80.95**
	Oracle	88.07	92.33	92.31	89.69	70.37	96.04	10.09	92.4	78.91	71.17	90.27

Table 7
Accuracy (%) of different networks (UNetFormer and DeepLabv2-VGG16) on the UAVid dataset. The best values for each column are in bold.

Model	Method	F1 per class					mean F1	mean IoU	OA
		Building	Road	Tree	Low vegetation	Car			
UNetformer	Baseline	89.57	85.53	79.14	68.73	67.16	78.03	64.85	81.11
	FESTA	88.95	86.94	78.44	68.65	67.98	78.19	65.06	80.94
	CRGNet	89.42	87.02	78.56	69.76	70.96	79.14	66.23	81.49
	TFCSD	89.90	83.94	77.34	68.69	72.12	78.40	65.15	80.65
	TEL	90.63	87.48	78.96	70.16	69.97	79.44	66.74	82.10
	DenseCRF	**92.36**	88.88	**81.59**	70.15	**74.31**	81.46	69.57	**84.13**
	AGMM	89.03	85.97	75.27	65.56	71.12	77.39	63.99	79.44
	Proposed	92.05	**89.40**	81.29	**70.36**	74.29	**81.48**	**69.59**	83.99
	Oracle	95.39	93.10	86.75	78.88	84.16	87.65	78.53	89.08
Deeplabv2-VGG16	Baseline	90.92	85.87	77.43	65.79	65.42	77.08	63.87	80.53
	FESTA	89.76	84.76	76.50	64.35	66.37	76.35	62.80	79.57
	CRGNet	89.75	85.84	77.22	64.73	64.72	76.45	63.04	79.87
	TFCSD	91.38	86.26	76.89	**66.31**	70.31	78.23	65.25	80.89
	TEL	91.29	85.01	76.41	66.05	69.14	77.58	64.38	80.42
	DenseCRF	91.99	86.33	**79.05**	63.51	70.66	78.31	65.53	81.76
	AGMM	91.15	85.70	77.68	64.85	68.71	77.62	64.51	80.85
	Proposed	**92.55**	**88.61**	77.74	65.70	**73.60**	**79.64**	**67.28**	**81.92**
	Oracle	94.65	91.06	84.43	72.49	76.82	83.89	73.14	86.55

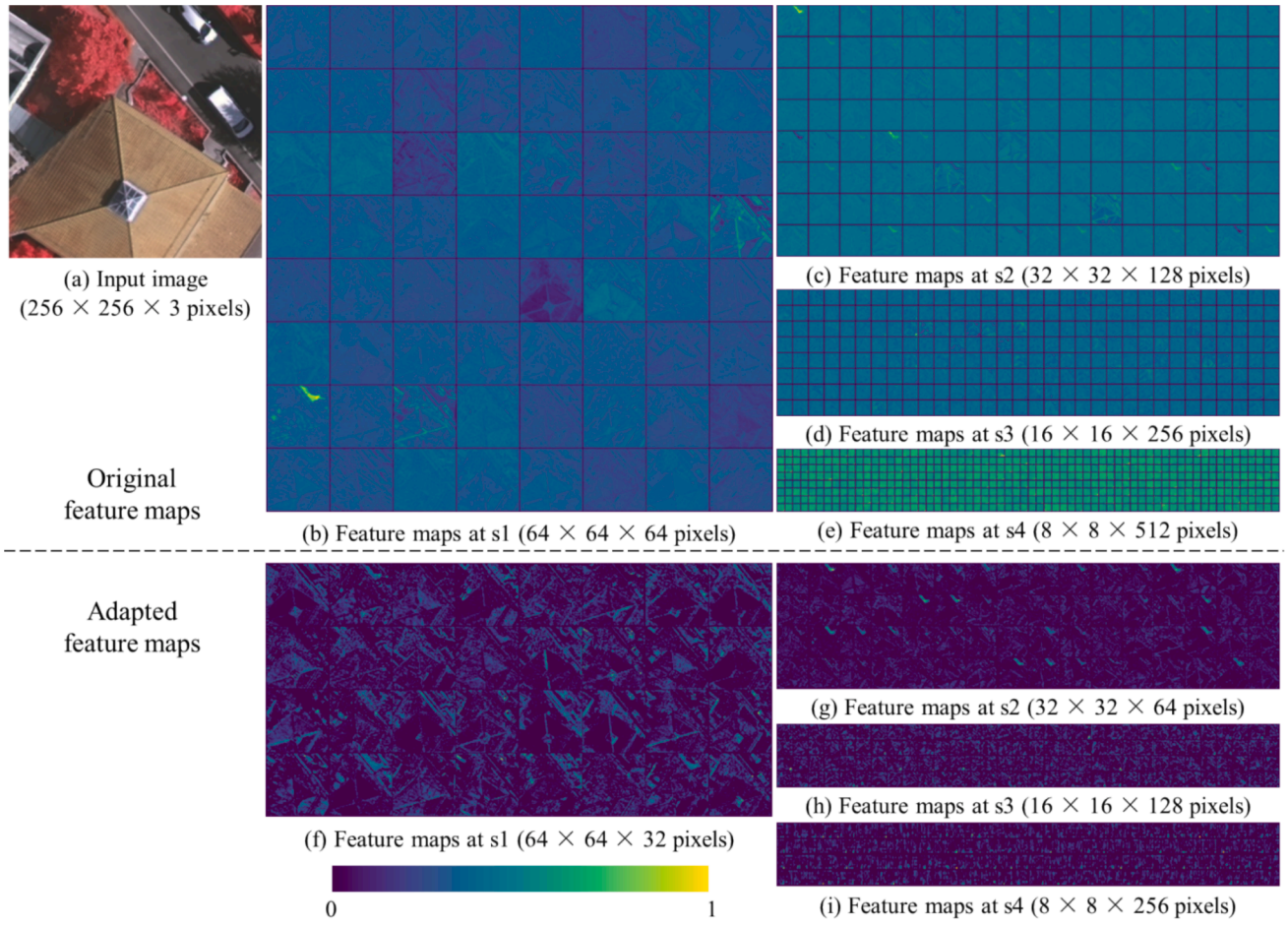


Fig. 8. Examples of original and adapted features of AESAM at four scales (s1-s4).

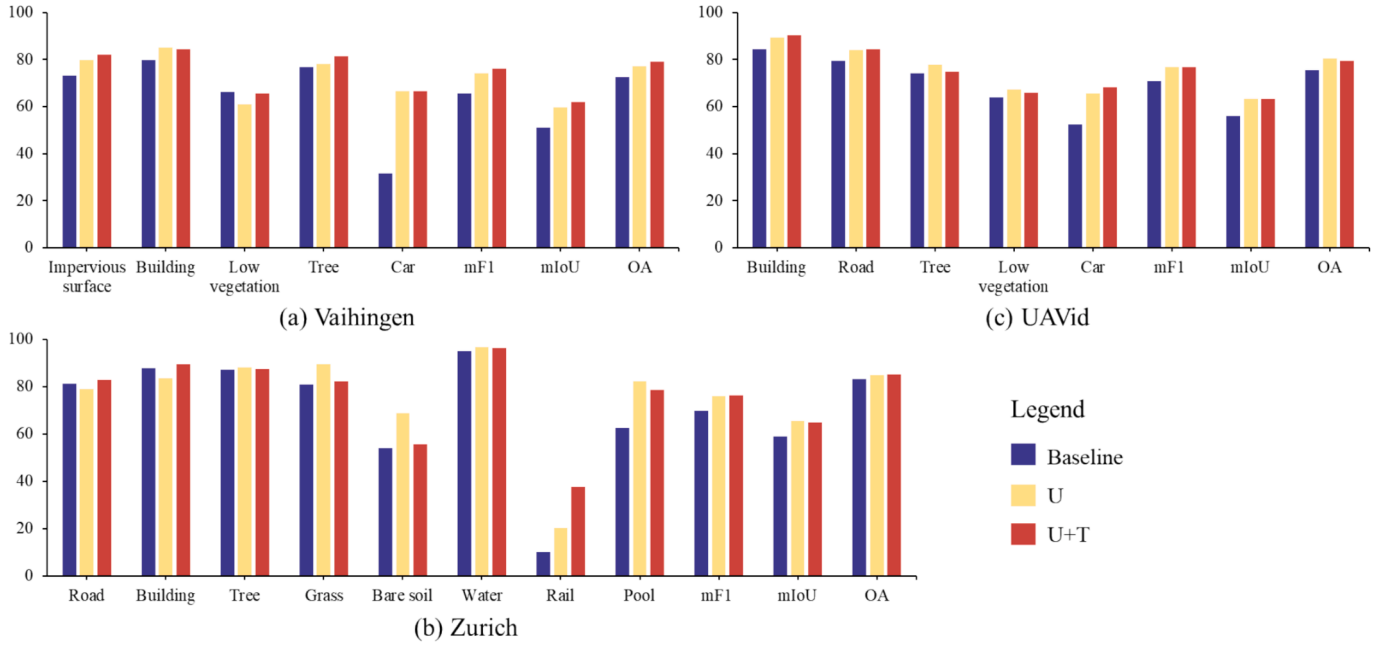


Fig. 9. Accuracy (%) of the proposed two modules, i.e., uncertainty-based pseudo-label generation (U) and transformation consistency constraint (T), on three datasets (a-c).

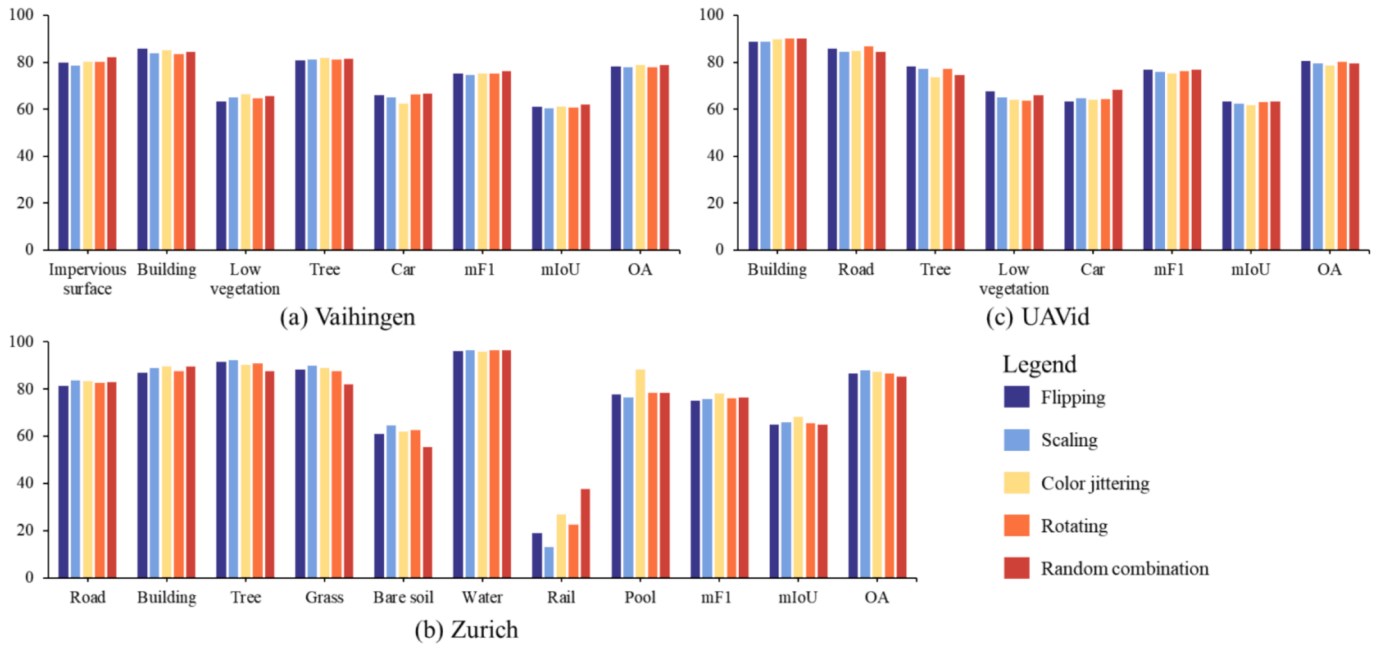


Fig. 10. Accuracy (%) of the proposed method with different transformation types on three datasets (a-c).

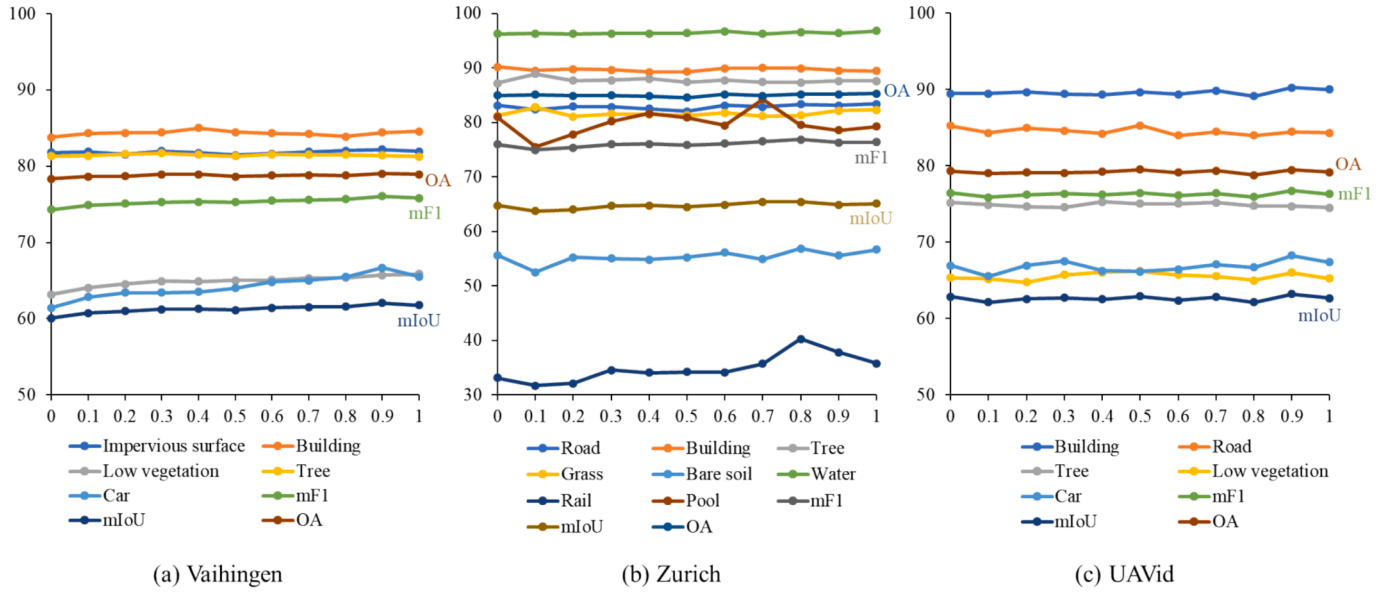


Fig. 11. Accuracy (%) of the proposed method with different thresholds on three datasets (a-c).

uncertainty value is about 0.7 (i.e., log(2)). The range of the initial threshold was set to [0, 1] with an interval of 0.1. The higher the initial threshold, the lower the effect of the iteration steps on the uncertainty filter, according to Eq.(3). Fig. 11 showed that optimal initial thresholds changed across datasets and categories, but the difference between different thresholds was little in most cases, concerning mean F1, mean IoU, and OA. On the Vaihingen dataset, the F1 values of the pool and rail significantly varied with the change of thresholds. This phenomenon can be explained by the dependence of F1 values on both precision and recall makes it highly sensitive to threshold tuning in imbalanced scenarios. The model's uncertainty on minority classes (pool and rail) exacerbates this effect, as small threshold shifts disproportionately alter the ratios of false positives and false negatives. To mitigate this, we can use techniques like class reweighting, resampling, or optimizing thresholds based on the precision-recall curve. This study set the initial threshold to 0.9, considering that this value has relatively stable performance on three datasets. Other values can also be considered, while selecting optimal ones still deserves further work.

5.3. Implications and limitations

This study designed a SAM-adapted weakly-supervised method with point labels to achieve semantic segmentation, alleviating the heavy reliance of deep networks on dense labels. Different from existing studies (Hua et al., 2021; Xu and Ghamisi, 2022), the proposed method was built on an efficient general-purpose foundation model (EfficientViT-SAM), and it introduced only a small number of trainable parameters, which can be readily fine-tuned for other applications (e.g., building detection). This study has some limitations too. In terms of data, the number and location of point labels were not explored, but they

were found important for foundation models (Chen et al., 2024). If point labels contain noise, the generalization ability of deep networks would be easily reduced. In this regard, noise label learning techniques (Song et al., 2023) can be introduced to refine point labels. In addition, this study did not explore the class imbalance problem, and it would lead to the omission of minor classes. This issue can be mitigated by weighted loss functions (Lin et al., 2017). Besides, more complicated transformation types can be used, e.g., CutMix (Yun et al., 2019) and ClassMix (Olsson et al., 2021).

To improve the performance of the proposed method, several strategies can be employed, including: 1) applying active learning to select informative labels to annotate instead of random selection; 2) developing transfer learning to utilize the learned knowledge of relevant tasks; and 3) introducing diverse information, e.g., object boundaries and height. With slight modifications, the proposed method can be adapted to other datasets such as SAR and multispectral imagery. Specifically, the foundation model, SAM, can be replaced by more domain-specific alternatives like SpectralGPT (Hong et al., 2024) and SARATR-X (Yang et al., 2024). In future work, we will explore how to extend the proposed method to other types of weak labels, e.g., image tags, scribbles, and bounding boxes, and how to combine them according to their properties. Moreover, we will consider how to fuse multi-modal data, e.g., texts, SAR, and nighttime light images, by transferring relevant models like optical-SAR models (Guo et al., 2024).

6. Conclusions

This study proposed a SAM-adapted weakly-supervised method using low-cost point labels for semantic segmentation of high-resolution remote sensing imagery. The method consists of three components: 1) an adapted EfficientViT-SAM, 2) uncertainty-based pseudo-label generation, and 3) transformation consistency constraint. Experimental results showed that the proposed method exhibited competitive results, especially in identifying rare classes, compared to existing weakly-supervised methods. Furthermore, we analyzed the sensitivity of the proposed method to network selection. We found that the proposed method still obtained advanced performance when using task-specific networks, but AESAM may not always perform better than task-specific networks. However, AESAM requires fine-tuning only a few parameters (0.9 M), which makes it promising for a range of downstream tasks. These findings demonstrated the promising potential of the proposed method in scenarios with limited labeling budgets.

CRediT authorship contribution statement

Yinxia Cao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Xin Huang:** Writing – review & editing, Writing – original draft, Supervision, Resources, Investigation, Formal analysis, Conceptualization. **Qihao Weng:** Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has received funding from Global STEM Professorship, Hong Kong SAR Government (P0039329), and Hong Kong Polytechnic University (P0046482 and P0038446). The authors express their gratitude to the editors and reviewers for their constructive comments and suggestions, which helped improve the manuscript.

Data availability

Data will be made available on request.

References

Ajadi, O.A., Barr, J., Liang, S.-Z., Ferreira, R., Kumpatla, S.P., Patel, R., Swatantran, A., 2021. Large-scale crop type and crop area mapping across Brazil using synthetic aperture radar and optical imagery. Int. J. Appl. Earth Obs. Geoinf. 97, 102294.

Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K., 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Bruzzone, L., Bovolo, F., 2013. A Novel Framework for the design of change-detection systems for very-high-resolution remote sensing images. Proc. IEEE 101, 609–630.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30m resolution: A POK-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z., 2024. RSPrompter: learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Trans. Geosci. Remote Sens. 62, 1–17.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Guan, H., Yu, Y., Ji, Z., Li, J., Zhang, Q., 2015. Deep learning-based tree classification using mobile LiDAR data. Remote Sens. Lett. 6, 864–873.

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., 2024. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27672–27683.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, Y., Wang, J., Zhang, Y., Liao, C., 2024. An efficient urban flood mapping framework towards disaster response driven by weakly supervised semantic segmentation with decoupled training samples. ISPRS J. Photogramm. Remote Sens. 207, 338–358.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., 2024. SpectralGPT: Spectral remote sensing foundation model. IEEE Trans. Pattern Anal. Mach. Intell.

Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A., 2022. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. European Conference on Computer Vision. Springer 600–619.

Hua, Y., Marcos, D., Mou, L., Zhu, X.X., Tuia, D., 2021. Semantic segmentation of remote sensing images with sparse annotations. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Huang, X., Zhang, L., 2013. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 51, 257–272.

Huang, X., Zhang, L., 2012. Morphological Building/Shadow index for building extraction from high-resolution imagery over urban Areas. IEEE J. Sel. Top. Appl. EARTH Obs. Remote Sens. 5, 161–172.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., 2023. Segment anything. In: In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with Gaussian edge potentials. Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011 24, 109–117.

Lei, X., Guan, H., Ma, L., Liu, J., Yu, Y., Wang, L., Dong, Z., Ni, H., Li, J., 2024. DAAL-WS: A weakly-supervised method integrated with data augmentation and active learning strategies for MLS point cloud semantic segmentation. Int. J. Appl. Earth Obs. Geoinf. 131, 103970.

Li, B., Gong, A., Zhang, J., Fu, Z., 2024. From image-level to pixel-level labeling: a weakly-supervised learning method for identifying aquaculture ponds using iterative anti-adversarial attacks guided by aquaculture features. Int. J. Appl. Earth Obs. Geoinf. 132, 104023.

Liang, Z., Wang, T., Zhang, X., Sun, J., Shen, J., 2022. Tree energy loss: Towards sparsely annotated semantic segmentation. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16907–16916.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Liu, L., Xiao, X., Qin, Y., Wang, J., Xu, X., Hu, Y., Qiao, Z., 2020. Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine. Remote Sens. Environ. 239, 111624.

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M.Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. ISPRS J. Photogramm. Remote Sens. 165, 108–119.

Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sens. Environ. 115, 1145–1161.

Olsson, V., Tranheden, W., Pinto, J., Svensson, L., 2021. ClassMix: Segmentation-Based data augmentation for semi-supervised learning, in. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1369–1378.

Pesaresi, M., Gerhardinger, A., Kayitakire, F., 2008. A Robust Built-Up area presence index by anisotropic rotation-invariant textural measure. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1, 180–192.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G., 2023. Learning From noisy labels with deep neural networks: a survey. IEEE Trans. Neural Networks Learn. Syst. 34, 8135–8153.
Sylvain, J.-D., Drolet, G., Thiffault, É., Anctil, F., 2024. High-resolution mapping of tree species and associated uncertainty by combining aerial remote sensing data and convolutional neural networks ensemble. Int. J. Appl. Earth Obs. Geoinf. 131, 103960.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
Tan, K., Jin, X., Plaza, A., Wang, X., Xiao, L., Du, P., 2016. Automatic Change detection in high-resolution remote sensing images by using a multiple classifier system and spectral–spatial features. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 3439–3451.
Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y., 2018. On regularized losses for weakly-supervised cnn segmentation. In: In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 507–522.
Tuia, D., Pacifici, F., Kanevski, M., Emery, W.J., 2009. Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. IEEE Trans. Geosci. Remote Sens. 47, 3866–3879.
Volpi, M., Ferrari, V., 2015. Semantic segmentation of urban scenes by learning local class interactions, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9.
Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.
Wang, P., Yao, W., Shao, J., 2023. One Class One Click: quasi scene-level weakly supervised point cloud semantic segmentation with active learning. ISPRS J. Photogramm. Remote Sens. 204, 89–104.
Wu, L., Zhong, Z., Fang, L., He, X., Liu, Q., Ma, J., Chen, H., 2023. Sparsely annotated semantic segmentation with adaptive Gaussian mixtures. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15454–15464.
Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.
Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., Krishnamoorthi, R., Chandra, V., 2024. In: EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything, in, pp. 16111–16121.
Xu, Y., Ghamisi, P., 2022. Consistency-regularized region-growing network for semantic segmentation of urban scenes with point-level annotations. IEEE Trans. Image Process. 31, 5038–5051.
Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y., 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7236–7246.
Yang, W., Hou, Y., Liu, L., Liu, Y., Li, X., 2024. SARATR-X: A foundation model for synthetic aperture radar images target recognition. arXiv Prepr. arXiv2405.09365.
Yue, J., Fang, L., Ghamisi, P., Xie, W., Li, J., Chanussot, J., Plaza, A., 2022. Optical remote sensing image understanding with weak supervision: concepts, methods, and perspectives. IEEE Geosci. Remote Sens. Mag. 10, 250–269.
Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.-H., Lee, S., Hong, C.-S., 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. ArXiv abs/2306.1.
Zhang, J., Jia, X., Hu, J., 2022. SP-RAN: Self-Paced Residual aggregated network for solar panel mapping in weakly labeled aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1.
Zhang, Z., Cai, H., Han, S., 2024. EfficientViT-SAM: accelerated segment anything model without performance loss. arXiv Prepr.
Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J., 2023. Fast Segment Anything. arXiv Prepr.