Dynamic Dictionary Learning for Remote Sensing Image Segmentation

Supplementary Material

A. Dataset Details

As shown in Tab. 1, we provide comprehensive and detailed descriptions of all the datasets used in this paper.

LoveDA [12] The LoveDA dataset is a fine-resolution 0.3m dataset designed for urban and rural land cover classification. It consists of 5,987 images, each with a resolution of 512 × 512 pixels. Captured in three cities—Nanjing, Changzhou, and Wuhan, China—LoveDA spans both urban and rural environments. The metropolitan region contains dense infrastructure and complex geometries, while the rural area features natural landscapes and sparse settlements. This diversity in geographic environments provides valuable data for assessing the adaptability and generalizability of segmentation models. The dataset's varied land cover improves its versatility for real-world segmentation tasks.

UAVid [10] The UAVid dataset is a high-resolution dataset for semantic segmentation tasks in urban environments. It was captured by an unmanned aerial vehicle (UAV) flying at an altitude of 50 meters and consists of 42 video sequences and 420 images. The images are available in two spatial resolutions: 3,840 × 2,160 and 4,096 × 2,160 pixels. The dataset includes a diverse range of urban objects, such as buildings, roads, trees, vegetation, vehicles, humans, and other urban clutter. Both top-down and side views of urban scenes are provided, offering a comprehensive perspective for object recognition. In training stage, each image is divided into patches of 1,024 × 1,024 pixels.

Potsdam [2] The ISPRS Potsdam dataset consists of 38 drone images from Potsdam, Germany, each with a resolution of 6000 × 6000 pixels and a ground sampling distance (GSD) of 5 cm, designed for semantic segmentation in urban environments. This dataset is annotated into six categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter. For the experiments, the images were cropped into 1,024 × 1,024-pixel patches to ensure manageable data processing and to focus on the primary RGB images and their corresponding labels.

Vaihingen [3] The ISPRS Vaihingen dataset includes 33 high-resolution images, each with a pixel resolution of 0.5 m, covering an urban region in Vaihingen, Germany. The images are classified into six categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter. The dataset's average image size is 2,494 × 2,064 pixels. For segmentation experiments, we used only the RGB images, and each image was divided into 1,024 × 1,024-pixel patches to facilitate efficient model training and testing.

Cloud [1] The Fine-Grained Cloud Segmentation dataset consists of 96 terrain-corrected (Level-1T) scenes from Landsat 8 OLI and TIRS, covering various biomes. This diverse dataset supports cloud detection and removal tasks in complex environments, offering pixel-level annotations for cloud shadow, clear sky, thin clouds, and cloud areas. Each scene is divided into 512 × 512-pixel patches and organized into training, validation, and test sets in a 6:2:2 ratio. The dataset's wide range of cloud cover types and biomes makes it a valuable resource for training and evaluating segmentation models in cloud detection tasks.

Grass [14] This dataset was developed to overcome the limitations of existing grassland segmentation datasets, such as boundary ambiguity and misclassification in complex terrains. Created using high-resolution satellite imagery from Gaofen-2 and Gaofen-6, it was captured in 2019 over Maduo County, located in the Yellow River source area of China. The dataset includes high-resolution images (8m) and provides detailed grassland coverage classifications across five levels: low coverage, medium-low coverage, medium coverage, medium-high coverage, and high coverage. It consists of 1,500 pairs of 256 × 256-pixel patches, with manual refinement to ensure high accuracy. This dataset is especially valuable for fine-grained grassland extraction in high-altitude, ecologically sensitive regions.

B. Evaluation Metrics

Overall Accuracy (OA) calculates the ratio of correct predictions to the total number of pixels:

$$\text{OA} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i + FN_i)}, \tag{1}$$

where TP_i, FP_i, and FN_i represent true positives, false positives, and false negatives for the i-th class, respectively.

Intersection over Union (IoU) measures the overlap between predicted and ground truth regions for a class. Mean IoU (mIoU) averages the IoU scores across all classes.

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{2}$$

Table 1. Overview of remote sensing image datasets for semantic segmentation.

Type	Dataset	Source	GSD*,†	Patch Size	Category
Coarse Grained	LoveDA [12]	Satellite	0.3m	1024 * 1024	background, building, road, water, barren, forest, agriculture
	UAvid [10]	UAV	50m†	1024 * 1024	clutter, building, road, tree, low vegetation, moving car, static car, human
	Potsdam [2]	UAV	5cm	1024 * 1024	impervious surface, building, low vegetation, tree, car, clutter
	Vaihingen [3]	UAV	9cm	1024 * 1024	impervious surface, building, low vegetation, tree, car, clutter
Fine Grained	Cloud [1]	Satellite	30m	512 * 512	cloud shadow, clear sky, thin cloud, thick cloud
	Grass [14]	Satellite	8m	256 * 256	low, medium-low, medium, medium-high, high

* *GSD* (Ground Sampling Distance): The physical pixel size projected onto the ground surface (e.g., 0.3m/pixel = each pixel represents a 0.3×0.3m ground area). Smaller GSD indicates higher spatial resolution.
† For UAVid dataset: The value indicates flight altitude rather than actual GSD. True GSD can be calculated via camera parameters.

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\text{IoU}_i, \tag{3}$$

where N is the number of classes, and IoU_i is the IoU for the i-th class.

F1 balances precision and recall, providing a harmonic mean of the two. Mean F1 score (mF1) averages the F1.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

$$\text{mF1} = \frac{1}{N}\sum_{i=1}^{N}\text{F1}_i, \tag{5}$$

where $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{Recall} = \frac{TP}{TP+FN}$. F1_i balances precision and recall for the i-th class.

While OA provides a general performance measure, it is less reliable for imbalanced datasets. In contrast, **IoU** and **F1 score** (and their mean versions, **mIoU** and **mF1**) offer more robust evaluations by considering class-level performance, making them better suited for multi-class tasks.

C. Ablation Study

Dimensions of Class ID Embedding We analyzed the impact of class ID embedding dimensions on segmentation performance. As shown in Fig. 1, increasing dimensions from 64 to 256 boosts performance across all datasets. While 512 dimensions achieve peak results on UAVid and Cloud, LoveDA slightly degrades, suggesting dimension sensitivity varies with dataset complexity. Extreme dimensions (1024) caused performance collapse, indicating overfitting risks. Therefore, we select 256 as the default value for achieving optimal efficiency-discriminability trade-off.

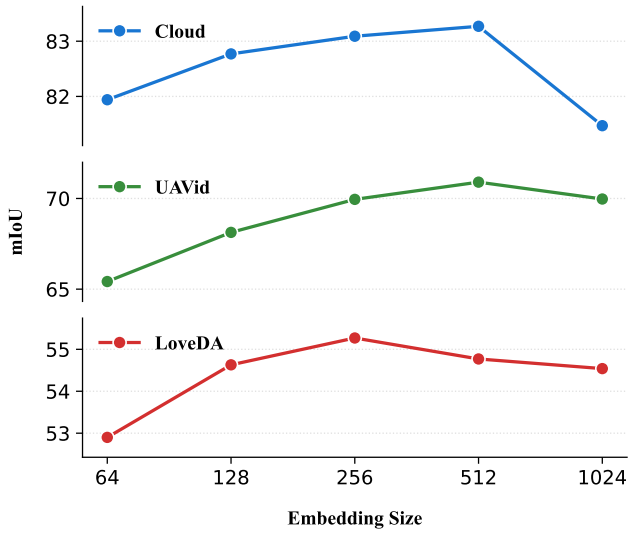


Figure 1. Performance comparison across different class ID embedding sizes for Cloud, UAVid, and LoveDA datasets. The mIoU scores are shown for embedding sizes ranging from 64 to 1024.

Backbone in Encoder We evaluate the impact of different backbone architectures on segmentation perfor-

Table 2. Performance comparison of different backbones under frozen and fully-trained settings on the LoveDA dataset.

Backbone	Status	mIoU ↑
ConvNeXt (B) [9]	Frozen	48.50
	Full	55.27
Swin (B) [7]	Frozen	49.76
	Full	53.88
Swinv2 (B) [8]	Frozen	48.43
	Full	54.41
MobileNetv3 (L) [5]	Frozen	46.34
	Full	51.17
EfficientNet (M) [11]	Frozen	44.93
	Full	53.78

Table 3. Computational efficiency comparison with input size 512×512, including * indicates parameters without backbone.

Method	Params (M)	Params* (M)	Flops (G)
AerialFormer [4]	113.8	26.9	126.8
SFA-Net [6]	**10.7**	**0.6**	**7.1**
KTDA [14]	258.3	170.7	566.4
Ours	90.9	3.3	89.2

mance under frozen and fully trained settings using the LoveDA [12] dataset, as shown in Tab. 2. Among the tested backbones, Swin (B) achieves the highest mIoU of 49.76% in the frozen setting, while ConvNeXt (B) outperforms others with 55.27% when fully trained, demonstrating its superior feature extraction capability when fine-tuned. Swinv2 (B) and Swin (B) exhibit comparable performance, with Swinv2 (B) slightly behind in both settings. MobileNetv3 (L) and EfficientNet (M), being lightweight backbones, yield lower mIoU scores, particularly in the frozen setting, indicating their limited capacity to generalize without fine-tuning. These results suggest that transformer-based backbones generally perform better than CNN-based alternatives, and that full training significantly boosts segmentation performance across all architectures.

D. Computational Efficiency

As shown in Tab. 3, our method achieves a substantial efficiency-performance tradeoff. Excluding the backbone, it has only 3.3M parameters, remaining lightweight. Though SFA-Net [6] has fewer FLOPs, its minimal parameters limit representation capacity. Remarkably, our model achieves significant performance improvements with fewer parameters than both AerialFormer [4] and KTDA [14].

E. More Visualization

Figs. 3 and 4 present comparative visualization results of our method against SFA-Net [6] and UNetFormer [13] across six challenging benchmarks. On the LoveDA (Fig. 3a) and UAVid (Fig. 3b) datasets where test set ground truth is unavailable, our segmentation masks exhibit superior alignment with visual semantics compared to baselines, particularly in preserving structural continuity of buildings and road networks. The Potsdam (Fig. 3c) and Vaihingen (Fig. 3d) results demonstrate our model's robustness against complex urban patterns, with significantly reduced segmentation artifacts in cluttered areas. Cloud (Fig. 4a) and Grass (Fig. 4b) segmentation further validate our approach's capability to handle fine-grained texture variations, achieving state-of-the-art boundary precision. A failure mode analysis (Fig. 5) reveals that tiling artifacts persist in UAVid inference due to our sliding window strategy, suggesting directions for future architectural improvements.

F. Training Dynamics

In this section, we present the curves showing how the metrics of our method and SFA-Net [6] change with the training epochs. By comparing the performance across various aspects, we highlight the advantages of our method.

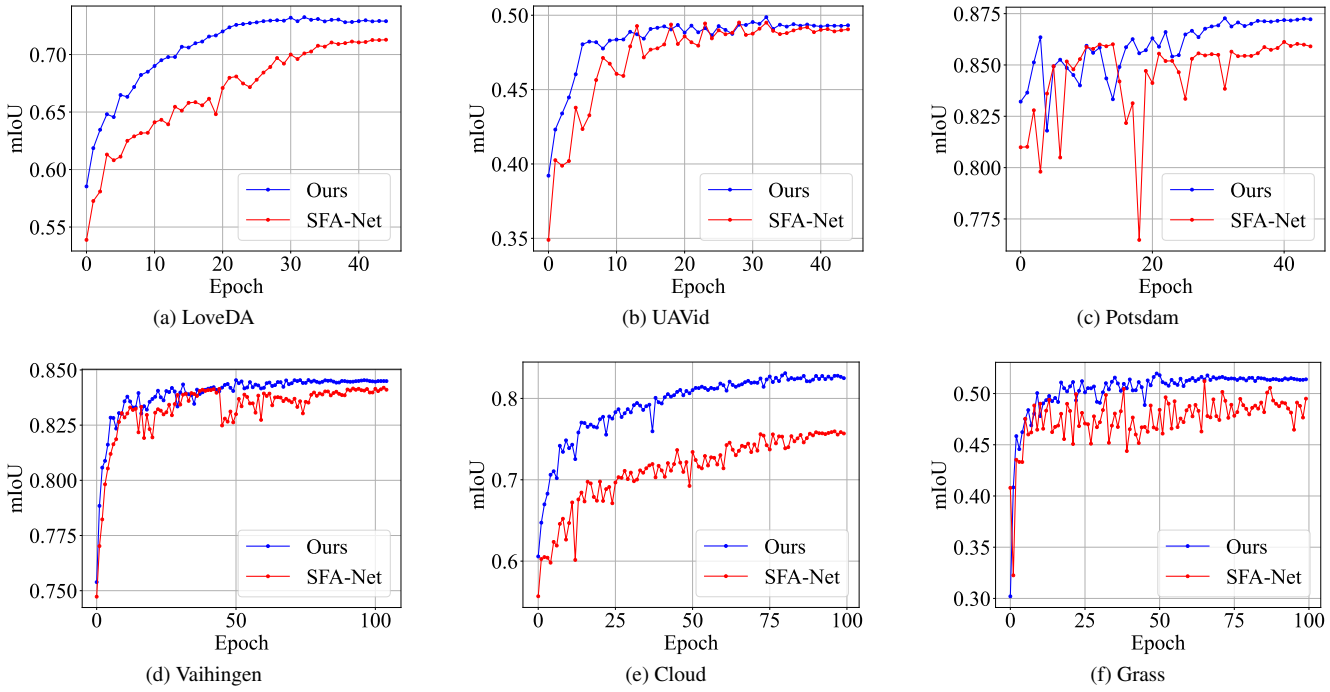
Training Stability As shown in Fig. 2b and Fig. 2f, our method demonstrates superior stability during training. Compared to SFA-Net, our approach exhibits minimal fluctuations and maintains a steady convergence throughout the training process, which reduces the need for extensive hyperparameter tuning, saving computational resources.

Fitting Speed Fig. 2a and Fig. 2e illustrate the faster fitting speed of our method. Our model achieves optimal performance in fewer epochs, while SFA-Net requires more training time to reach similar results. This faster convergence enables more efficient model training.

Final Performance Fig. 2 holistically demonstrates our method's superiority across all evaluation metrics (subfigures a.-f.). The consistent performance advantages visible throughout the training lifecycle – from initial convergence patterns to final stabilized outputs – validate our approach's end-to-end effectiveness.

References

[1] Steve Foga, Pat L Scaramuzza, and Song et al. Guo. Cloud detection algorithm comparison and validation for operational landsat data products. *Remote sensing of environment*, 194:379–390, 2017. 1, 2



(a) LoveDA (b) UAVid (c) Potsdam

(d) Vaihingen (e) Cloud (f) Grass

Figure 2. Validation set mIoU trends across all datasets over training epochs.

[2] International Society for Photogrammetry and Remote Sensing. Posdam datasets. *Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx*, 2024. 1, 2

[3] International Society for Photogrammetry and Remote Sensing. Vaihingen datasets. *Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx*, 2024. 1, 2

[4] Taisei Hanyu, Kashu Yamazaki, Minh Tran, Roy A. McCann, Haitao Liao, Chase Rainwater, Meredith Adkins, Jackson Cothren, and Ngan Le. Aerialformer: Multi-resolution transformer for aerial image segmentation. *Remote Sensing*, 16(16), 2024. 3

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *ICCV*, 2019. 3

[6] Gyutae Hwang, Jiwoo Jeong, and Sang Jun Lee. Sfa-net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sensing*, 16(17):1–18, 2024. 3

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[8] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pages 12009–12019, 2022. 3

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 3

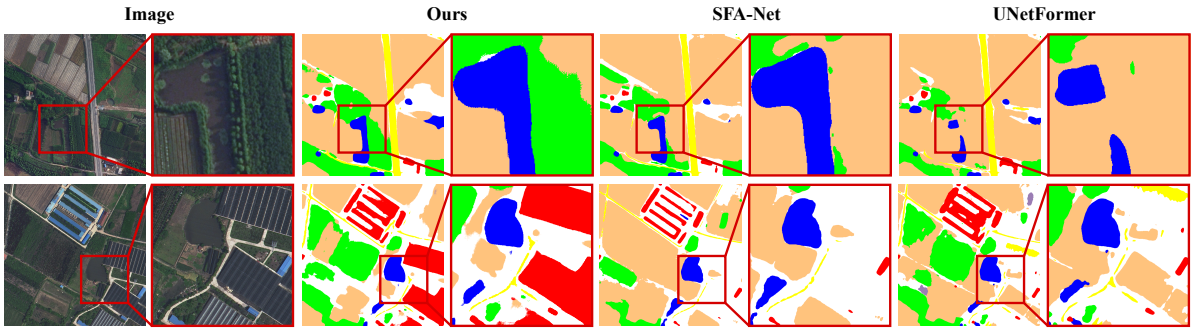
[10] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS J. Photogramm. Remote Sens.*, 165:108–119, 2020. 1, 2

[11] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *ICML*, pages 10096–10106. PMLR, 2021. 3

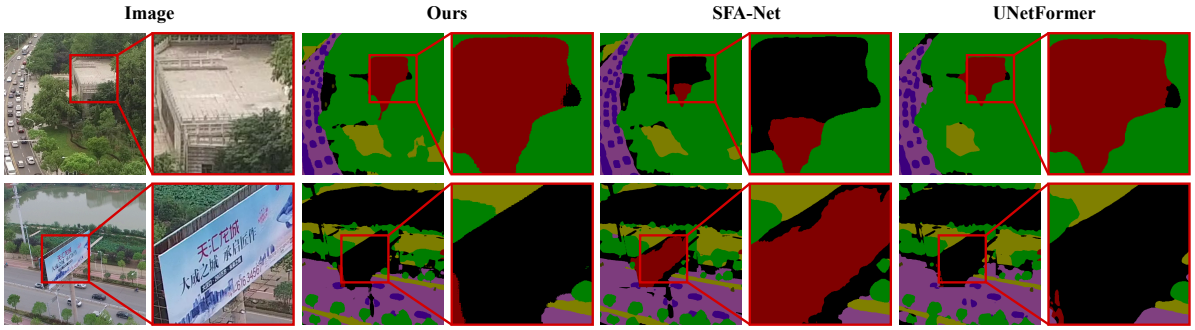
[12] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *NeurIPS*, pages 1–17, 2021. 1, 2, 3

[13] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.*, 190:196–214, 2022. 3

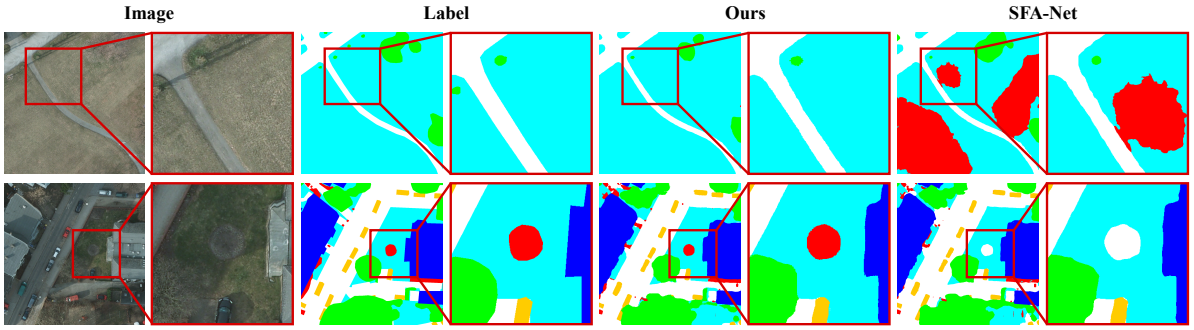
[14] Shun Zhang, Xuechao Zou, Kai Li, Congyan Lang, Shiying Wang, Pin Tao, and Tengfei Cao. Knowledge transfer and domain adaptation for fine-grained remote sensing image segmentation, 2025. 1, 2, 3



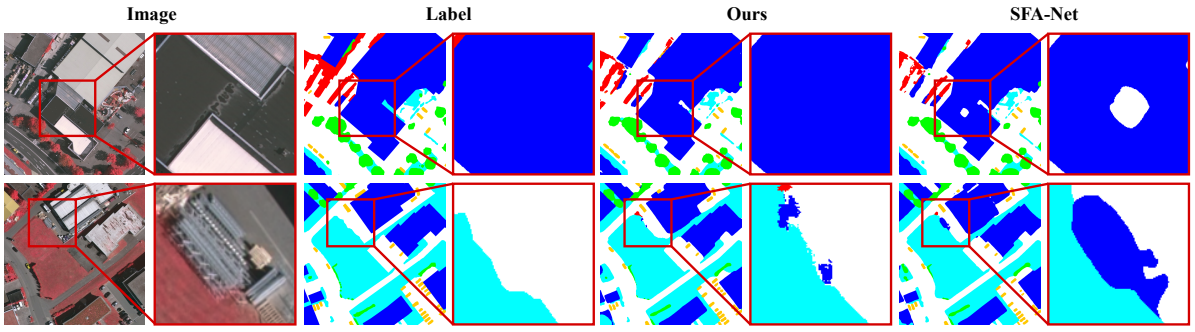
(a) LoveDA



(b) UAVid

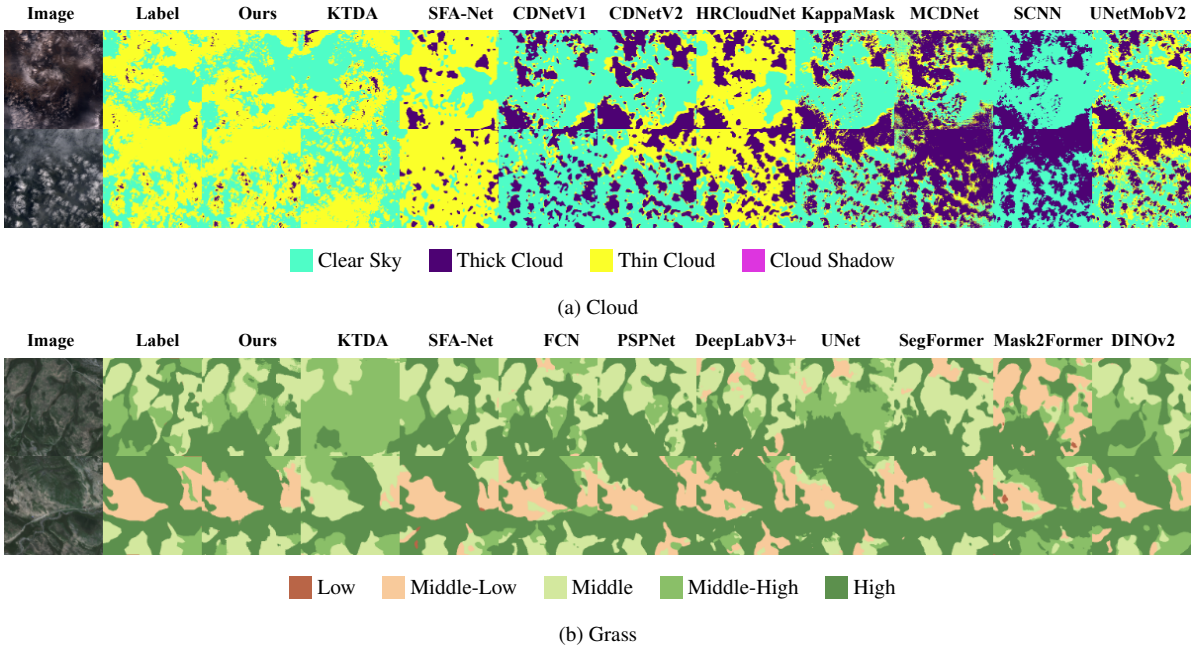


(c) Potsdam



(d) Vaihingen

Figure 3. More visualization results of coarse-grained remote sensing images.



(a) Cloud

(b) Grass

Figure 4. More visualization results of fine-grained remote sensing images.

Figure 5. Visualization examples of bad cases.