Exact: Exploring Space-Time Perceptive Clues for Weakly Supervised Satellite Image Time Series Semantic Segmentation

Supplementary Material

In this supplementary material, we provide the following contents: 1) more implementation details of dataset, our *Exact* and competing methods (§A), 2) the difference with the previous prototype-based methods (§B), 3) additional experimental results and analyses (§C).

A. Additional Implementation Details

A.1. Raw CAM Generation

As discussed in Section 3.1 in the main paper, the raw fused CAM can be obtained by the dense tokens and classifier weights. Inspired by previous works in natural image[5, 13, 16], we adopt a theoretically equivalent and much straightforward way to compute the raw fused CAM:

$$\hat{y} = \mathrm{GAP}(\mathrm{Conv}(\mathbf{Z}_T^{\text{dense}})) + \mathrm{GAP}(\mathrm{Conv}(\mathbf{Z}_S^{\text{dense}})), \quad (1)$$

$$\mathcal{L}_{\text{cls}}^{\text{aux}} = \frac{1}{K}\sum_{i=1}^{K} y_i \log \sigma(\hat{y}_i) + (1 - y_i)\log(1 - \sigma(\hat{y}_i)), \quad (2)$$

$$\mathcal{M} = \mathrm{ReLU}(\mathrm{Conv}(\mathbf{Z}_T^{\text{dense}})) + \mathrm{ReLU}(\mathrm{Conv}(\mathbf{Z}_S^{\text{dense}})). \quad (3)$$

Here $\mathrm{Conv}(\cdot)$ is the convolution layer to obtain $\mathbf{Z}' \in \mathbb{R}^{N_h \cdot N_w \times K}$ and y denotes the class labels.

A.2. Dataset and Model Settings

Data Setting. The size of each sample in the PASTIS [7] is 128 × 128. We follow the original settings in TSViT [10] that *split* each sample into 24 × 24 patches to ensure that the network can be trained efficiently on the available hardware. In order to clearly visualize the activation area of CAM on the original SITS sample, we *spliced* the small patches back to 128 × 128 in Fig. 6 in main paper. Each sample in the Germany [9] has a size of 24 × 24, and no additional processing was applied. Since the multi-class labels are absent in original PASTIS and Germany datasets, we employ the following strategy to assign multi-class labels to each SITS: If the number of masks for class k constitutes at least 1% of the spatial size, the class k is deemed to be present in SITS. To accommodate a large set of experiments, we only use fold-1 among the five folds provided in PASTIS.

Model Setting. Following the settings in [10], we set the vector dimension d to 128. The temporal encoder and spatial encoder comprise 8 and 4 layers, respectively. The spatial size of each patch is set to $h = w = 2$. For the spatial clustering, we apply ℓ_2 normalization to the pixel embedding before measuring the similarity with prototypes. For the temporal-aware affinity mining, we normalize the temporal-to-class attention $\tilde{\mathcal{A}}$ to the range of [0,1] using `softmax` function and subsequently reweight the temporal sequence embeddings by the normalized attention. Besides, we iteratively propagate the temporal-aware pairwise affinity among neighboring pixels, with the iterations are set to 3. To derive the final pseudo labels, we apply a global threshold to separate the foreground and background in the CB-CAMs $\mathcal{Y}$, as following [11, 13].

Algorithm A Adaptation strategy 1.

Input: SITS $\boldsymbol{I} \in \mathbb{R}^{T \times C \times H \times W}$.
Parameter: cloud threshold thr.
Function: WSSS network f designed for natural images.
Output:single CAM $\mathcal{M} \in \mathbb{R}^{N_h \cdot N_w \times K}$ for SITS.

```
Initialize CAMs list $\mathcal{M}' \leftarrow [\,]$;
for $t \leftarrow 1$ to $T$ do
    if max($\boldsymbol{I}_t$) $< thr$ then
                                        ▷ cloud cover check
        $\boldsymbol{I}_t \leftarrow$ concat($\boldsymbol{I}_t$, time position)
                                        ▷ single temporal sequence input
        $\mathcal{M}_t \leftarrow f(\boldsymbol{I}_t)$;
        $\mathcal{M}' \leftarrow \mathcal{M}'$.append($\mathcal{M}_t$);
    end if
end for
$\mathcal{M} \leftarrow$ mean($\mathcal{M}'$)        ▷ Average over T dimension
return $\mathcal{M}$;
```

A.3. Competing Methods and Modules

A.3.1. Adaptation of Competing Methods

For WSSS methods originally designed for natural images, we attempt several adaptation strategies to enable their application to SITS inputs, as follows:

Strategy 1. Given a SITS with dimensions $[T, C, H, W]$, we partition it into T single-temporal samples, each with dimensions $[C + 1, H, W]$. The additional channel represents the temporal position embedding, which is used to differentiate distinct temporal samples. We then check each temporal sample for cloud cover by evaluating the maximum signal intensity, and removing any samples identified as *cloudy*. Each remaining temporal sample is individually processed by the WSSS networks to generate CAM. Finally, we average the CAMs from temporal samples to obtain a single CAM for SITS. The pseudo-code is attached in Algorithm A.

Strategy 2. We reconstruct the SITS into the 3D format $[T \times C, H, W]$ by merging the first and second dimensions. Subsequently, we directly input the reconstructed data into

Method	*Strategy* 1		*Strategy* 2	
	OA	mIoU	OA	mIoU
MCTFormer$_{\text{CVPR'22}}$ [13]	63.8	41.5	66.7	49.6
ViT-PCM$_{\text{ECCV'22}}$ [8]	65.1	46.3	69.3	53.2
TSCD$_{\text{AAAI'23}}$ [14]	64.8	45.9	67.2	51.3
DuPL$_{\text{CVPR'24}}$ [12]	63.4	40.7	65.5	48.7

Table 1. **The performance of WSSS designed for natural images under different adaptation strategies.**

the WSSS model to obtain the CAM for SITS.

A.3.2. Adaptation of Competing Modules

Most of modules designed in natural image WSSS cannot be applied in SITS scenario due to the distinct data processing pipeline. We carefully select and reimplement four modules that can be adapted to temporal-spatio network.

PAMR [1]. For this method, we incorporate the nGWP and PAMR modules. We feed the score map output from temporal encoder and spatial encoder into the nGWP module to replace the convenient GAP layer. The PAMR module is widely utilized in natural image WSSS. Here, we follow the common practice that using low-level intensities as the input to PAMR to refine the fused raw CAM.

TS-CAM [6]. Since this method performs semantic reallocation and semantic aggregation from a spatial perspective, we follow the original settings in [6] that compute TS-CAM within the spatial encoder. Besides, we replace the multi-class tokens in spatial encoder with single-class token to maintain consistency with the original implementation.

SIPE [3]. This model computes the inter-pixel semantic correlations in feature space, providing additional guidance for extending the CAM. We reproduce SIPE in the temporal embedding space to align with our proposed module.

FPR [2]. We calculate the region-level contrast loss and pixel-level rectification loss proposed by FPR in the temporal embedding space. Note that both SIPE and FPR use prototypes, and we set the number of prototypes per class to 2 to better match the inherent characteristics of the SITS.

B. Remarks on difference with previous works in natural images.

Existing prototype-based WSSS works. The prototype-based methods have been explored in the WSSS for natural images. Existing prototype-based WSSS works on natural images primarily focus on the following aspects: 1) setting a large number of prototypes (approximately 30 per class) to capture diverse intra-class patterns and leveraging these prototypes to reduce intra-class variation [2, 15, 17]. 2) performing clustering on the batch-level and utilize the prototypes obtained from the current batch to expand the raw CAM [3–5].

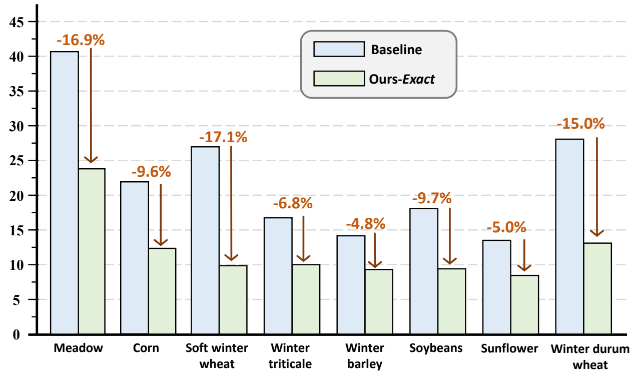


Figure 1. **False discovery rate (FDR) of baseline and *Exact*.** The results are evaluated on the PASTIS *train* set for several major crop types. *Exact* significantly reduce the FDR across different crops.

Difference with the previous works in natural images. Our approach is different from existing prototype-based WSSS methods in the following ways: 1) our objective is to explicitly capture compact intra-class patterns using a minimal number of prototypes ($N_p = 2$) and leverage the semantics to enlarge the variations across different crops. 2) we perform momentum updating at the dataset-level and conduct spatial clustering in the most class-relative regions. 3) we discard the classifier weights and directly utilize the well-updated prototypes to generate the final CAMs.

Our method thoroughly considers the unique characteristics of SITS and introduces tailored strategies, thereby achieving impressive performance. By contrast, prototype-based methods designed for natural images yield only limited improvements. As shown in the last column of Tab.1a in the main paper, SIPE [3] and FPR [2] (two prototype-based methods for natural images) bring only 0.6% and 1.5% improvements compared to baseline, while our *Exact* achieves a substantial 6.1% enhancement.

C. Additional Experiments and Analyses

Due to constraints on page space, we are unable to present all experimental results within the main paper. In this section, we provide more experimental results and analyses both quantitatively and qualitatively to support the main paper.

C.1. Different Adaptation Strategies.

We evaluate the performance of the WSSS under different adaptation strategies on PASTIS *train* set, the results are shown in Tab. 1. It can be observed that the WSSS models generally perform better under *Strategy* 2. This is because merging the temporal dimensions during input allows the model to implicitly focus on pivotal temporal clips, thereby mitigating the adverse effects of anomalous temporal periods to some extent. To eliminate the influence of irrelevant

Method	*Sup.*	**OA**	**mIoU**	O.*ratio*	m.*ratio*
TSViT [10]	$\mathcal{G}$	95.0	84.8	100%	100%
baseline	$\mathcal{P}$	84.7	73.6	89%	87%
+PAMR$_{\text{CVPR'20}}$ [1]		85.4	74.8	90%	88%
+TS-CAM$_{\text{ICCV'21}}$ [6]		82.6	71.9	87%	85%
+SIPE$_{\text{CVPR'22}}$ [3]		84.2	73.1	89%	86%
+FPR$_{\text{ICCV'23}}$ [2]		84.5	73.9	89%	87%
+ours-*Exact*		**90.1** ↑ **5.4**	**79.9** ↑ **6.3**	**95%**	**94%**

Table 2. **The TSViT segmentation network performance trained with pseudo labels on Germany *test* set.** All pseudo labels are consistent with those described in the main paper. O.*ratio* and m.*ratio* refer to the proportion of OA and mIoU between weakly supervised and fully supervised of segmentation performance.

Method	*Sup.*	**OA**	**mIoU**	O.*ratio*	m.*ratio*
U-TAE [7]	$\mathcal{G}$	82.9	62.4	100%	100%
baseline	$\mathcal{P}$	76.0	55.5	92%	89%
+PAMR$_{\text{CVPR'20}}$ [1]		77.5	56.7	93%	91%
+TS-CAM$_{\text{ICCV'21}}$ [6]		75.3	54.2	91%	87%
+SIPE$_{\text{CVPR'22}}$ [3]		76.4	56.1	92%	90%
+FPR$_{\text{ICCV'23}}$ [2]		76.8	56.4	93%	90%
+ours-*Exact*		**82.1** ↑ **6.1**	**60.7** ↑ **5.2**	**99%**	**97%**

Table 3. **The U-TAE segmentation network performance trained with pseudo labels on PASTIS *test* set.** All pseudo labels are consistent with those described in the main paper.

Method	$\mathcal{L}_{\text{cbl}}$	$\mathcal{L}_{\text{tap}}$	*CB-CAM*	**OA**	**mIoU**
baseline				81.2	69.5
	✓			81.9	71.6
		✓		82.4	72.3
			✓	82.2	72.5
	✓	✓		83.0	73.4
	✓		✓	82.8	73.3
		✓	✓	83.2	73.8
ours-*Exact*	✓	✓	✓	**84.1**	**75.6**

Table 4. **Additional ablation results of different components.**

factors, we choose the *Strategy* 2 in the main paper to reimplement the WSSS method designed for natural images.

C.2. More Comparisons

Effect of our method on correcting false positives. We evaluate the false discovery rate (FDR) for each category of pseudo-labels to quantitatively analyze the effectiveness of our method in mitigating the over-activation regions. The FDR can be computed as follows:

μ_l	μ_h	**OA**	**mIoU**
0.15	0.40	83.8	75.2
0.20	0.35	83.2	75.0
0.20	**0.40**	**84.1**	**75.6**
0.20	0.45	83.3	75.1
0.25	0.40	83.0	75.3

(a) Filtering thresholds μ.

λ_1	λ_2	**OA**	**mIoU**
0.005	0.015	83.5	75.0
0.01	**0.015**	**84.1**	**75.6**
0.01	0.010	83.7	75.3
0.01	0.02	83.9	74.9
0.015	0.015	84.0	75.2

(b) Loss coefficients λ.

Table 5. **Effect of the filtering thresholds and loss coefficients.**

$$\mathtt{FDR} = \frac{\mathtt{FP}}{\mathtt{FP} + \mathtt{TP}} \tag{4}$$

here $\mathtt{FP}$ and $\mathtt{TP}$ denote the number of false positive and true positive pixel pseudo labels for each class. We compare the FDR of pseudo labels generated by TSViT-CAMs (baseline) and ours *Exact*, the results are shown in Fig. 1. It can be seen that due to the noise perturbations, the baseline exhibits more false positive pixels, leading to inferior perceive ability. Our method significantly suppresses erroneous activation regions and reduces the FDR across different crops, thereby delineating the crop regions more precisely.

Segmentation results on Germany dataset. We further validate the performance of segmentation network trained by different pseudo labels on the Germany [9] dataset. As in the main paper, we employ the original TSViT [10] with a segmentation decoder as our SITS semantic segmentation network. As shown in Tab. 2, training the segmentation network with *Exact*-generated labels achieves the best results, improving the baseline by 5.4% in OA and 6.3% in mIoU, respectively. This indicates that our method can show consistently superior performance across various SITS crop mapping benchmarks.

Segmentation results for other SITS segmentation network. To further demonstrate the superiority of our method, we replace the TSViT with the U-TAE [7] as our semantic segmentation network and evaluate its performance under various pseudo labels generated by different methods. The results are shown in Tab. 3. Notably, using the pseudo labels generated by *Exact*, U-TAE can achieve 99% and 97% of the fully supervised OA and mIoU respectively, showcasing the impressive performance of our method. These findings indicate that training lightweight network with the pseudo labels generated by our method has the potential to achieve performance comparable to its fully supervised paradigm.

C.3. More Ablation Studies

We provide more ablation experiments in this section, and all results are reported on the PASTIS *train* set.

Comprehensive ablation results on proposed modules. In Tab. 4, we present additional ablation results of our proposed modules. It can be observed that our proposed mod-

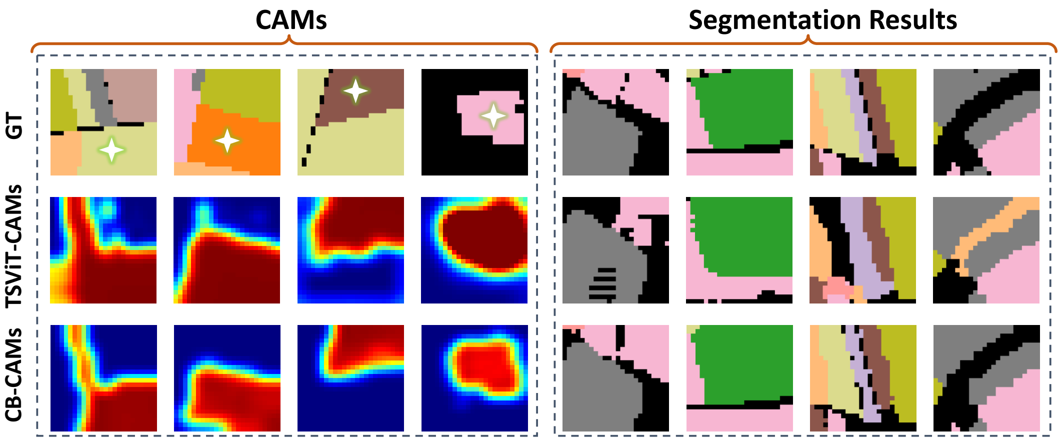


Figure 2. **Qualitative results between baseline TSViT-CAM and CB-CAMs derived by *Exact* on Germany dataset. Left**: CAMs comparisons. **Right**: Semantic segmentation comparison results. The stars represent the corresponding activation category.

	2000	**4000**	5000	6000
OA	82.0	**84.1**	83.4	83.0
mIoU	72.1	**75.6**	75.2	74.9

(a) Warm up stages.

	w/o Neg	Neg
OA	83.3	**84.1**
mIoU	74.6	**75.6**

(b) Negative set.

Table 6. **Effect of the warm-up stages and negative prototype set.** w/o refers to without.

ules synergize effectively, as mentioned in the main paper. $\mathcal{L}_{\text{cbl}}$ regularizes the embedding space, facilitating the global perception of the space-time clues to crop regions. Simultaneously, $\mathcal{L}_{\text{tap}}$ mitigates anomalous semantics while indirectly reinforcing the stability of spatial clustering process.

Effect of filtering thresholds μ and loss coefficients λ_i. In Sec. 3.2 of the main paper, we employ two thresholds (μ_l, μ_h) to filter out the most class-relative regions, both positively and negatively, as follows:

$$\hat{\mathcal{M}} = \begin{cases} 0, & \text{if } \mathcal{M} \leq \mu_l, \\ 1, & \text{if } \mathcal{M} \geq \mu_h, \\ \text{ignore}, & \text{otherwise.} \end{cases} \tag{5}$$

Tab. 5a shows the performance variations under different filtering thresholds. As we can see, an excessively stringent threshold may impede the ability to capture the patterns of crops, whereas a lenient threshold may introduce undesired noise to the prototypes. In addition, we report the impact of different loss coefficients on accuracy, the results are presented in Tab. 5b.

Effect of the warm up stage. The prototype learning relies on the raw CAM's accurate perception of parcel objects. Since the network lacks the capability to perceive parcel objects at the early training stages, prematurely introducing prototype learning and feature space shaping may result in gradient explosion. Tab. 6a shows the impact on the warm up stages for the model performance. We can see that starting the prototype learning and clue-based contrastive learning at 4000 iterations can achieve the best performance. While an excessively prolonged warm-up stage may cause the model to memorize noise perturbations, thereby hindering the shaping of the feature space.

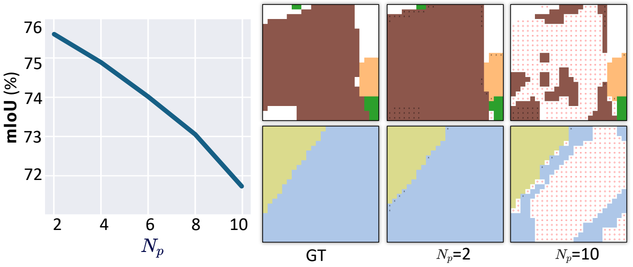


Figure 3. **Quantitative and qualitative results of pseudo labels with different N_p.** The red dot ○ and black cross × in qualitative results denote the false negative and the false positive activations, respectively.

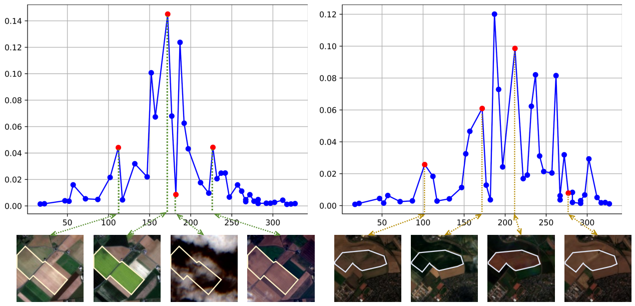


Figure 4. **Single satellite image extracted by temporal-to-class attention.** The line chart represents the temporal-to-class attention, and the red dots correspond to the satellite images shown below.

Effect of the negative set. In the main paper, we introduce a positive prototype set and a negative prototype set to capture class-relevant positive and negative patterns, respectively. We present additional quantitative results in Tab. 6b to demonstrate the effectiveness of the negative set. The results indicate that the negative class-relevant semantics can complement with positive patterns, thereby assisting the model in eliminating erroneous crop regions.

Adverse impact of increasing the number of prototypes. As discussed in the main paper, an excessive number of prototypes may impair the model's ability to perceive the global unified semantics of the crop parcel. In Fig. 3, we present both quantitative and qualitative experimental results to illustrate the impact of increasing the number of prototypes. As we can see, the prototypes' ability to perceive crop parcels declines sharply as N_p increases to 10 (75.6% vs. 71.8% mIoU). This is mainly due to the large N_p compels the prototypes to capture local discriminative patterns, thereby resulting in a severe under-activation issue.

C.4. Additional Qualitative Results

Low-level mapping of the temporal-to-class attention. In order to intuitively demonstrate the effect of temporal-to-class attention on the perception of temporal sequences, we list the satellite images under different attention scores. As shown in Fig. 4, temporal clips with high attention scores contain pivotal information for crop recognition, whereas those low scores are associated with anomalous temporal periods (*e.g*., cloud cover, barren land). Therefore, explicitly emphasizing the contributions of different temporal clips to crop recognition can mitigate the confusion arising from anomalous semantics.

Visual comparison of CAMs and segmentation results. We additionally provide visual comparison between the TSViT-CAMs (baseline) and the proposed CB-CAMs on Germany dataset, as shown in Fig. 2. The first four columns show the visualization of the CAMs. One can observe that the CB-CAMs generated by *Exact* are capable of accurately delineating the crop regions. Therefore, the semantic segmentation model tends to show more powerful perceptual capability under *Exact*-generated pseudo labels' supervision.

Visual comparison of pseudo labels. In Fig. 5, we show the pseudo labels derived by TSViT-CAMs and CB-CAMs on both PASTIS and Germany *train* set. Consistent with the main paper, we observe that *Exact* remarkably addresses both under- and over-activation issues in the baseline, thereby providing more reliable supervision for SITS semantic segmentation network.

References

[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020. 2, 3

[2] Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, and Lei Zhang. Fpr: False positive rectification for weakly supervised semantic segmentation. In *ICCV*, 2023. 2, 3

[3] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *CVPR*, 2022. 2, 3

[4] Zhaozheng Chen and Qianru Sun. Extracting class activation maps from non-discriminative features as well. In *CVPR*, 2023.

[5] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *CVPR*, 2022. 1, 2

[6] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *CVPR*, 2021. 2, 3

[7] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *ICCV*, 2021. 1, 3

[8] Simone Rossetti, Damiano Zappia, Marta Sanzari, Marco Schaerf, and Fiora Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *ECCV*, 2022. 2

[9] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 2018. 1, 3

[10] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *CVPR*, 2023. 1, 3

[11] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020. 1

[12] Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *CVPR*, 2024. 2

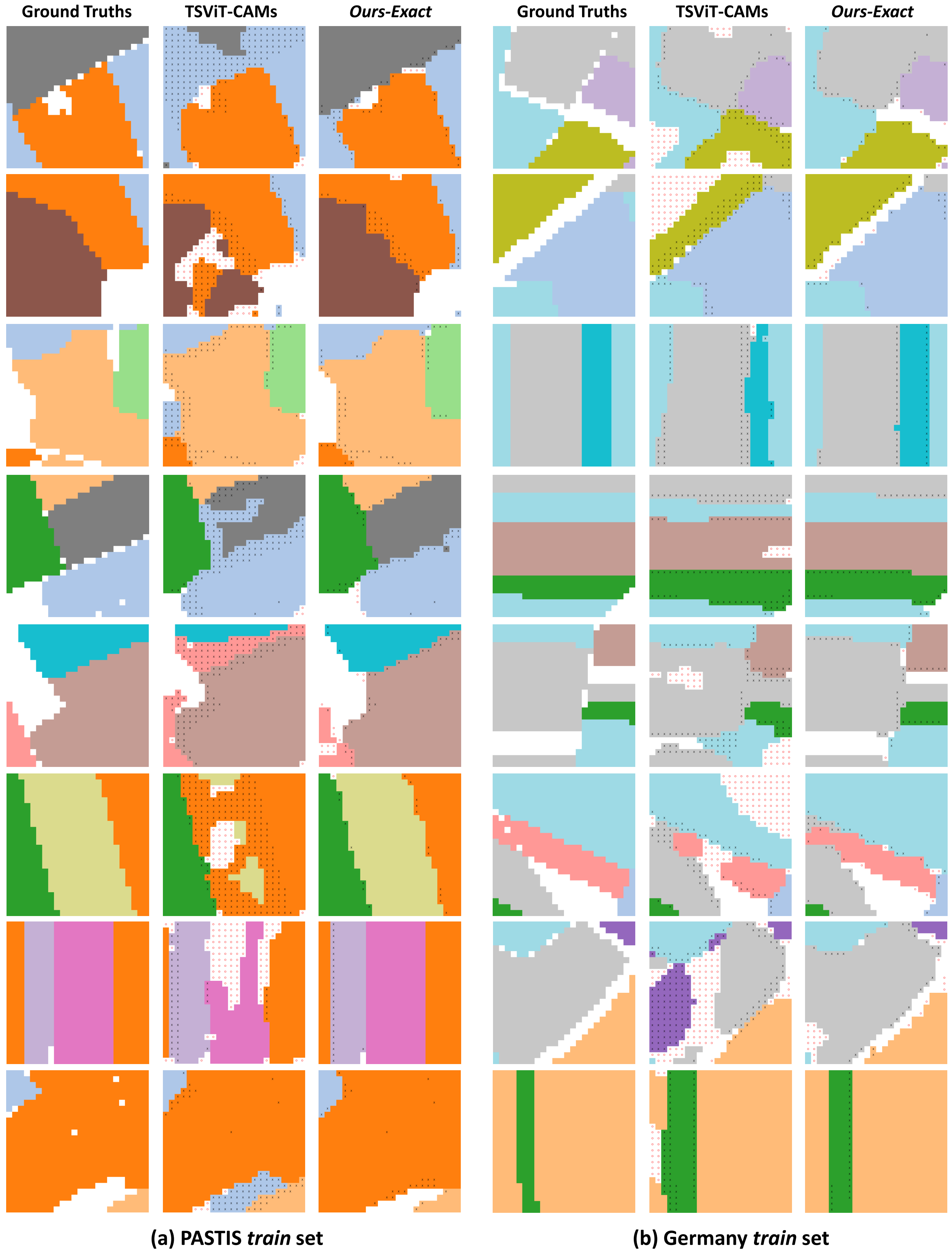
[13] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 1, 2

[14] Rongtao Xu, Changwei Wang, Jiaxi Sun, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. In *AAAI*, 2023. 2

[15] Zhiwei Yang, Kexue Fu, Minghong Duan, Linhao Qu, Shuo Wang, and Zhijian Song. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *CVPR*, 2024. 2

[16] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 1

[17] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*, 2022. 2



(a) PASTIS *train* set

(b) Germany *train* set

Figure 5. **Qualitative comparison of pseudo labels among baseline TSViT-CAMs and ours *Exact* on PASTIS and Germany *train* set.** The red dot ○ and black cross × in qualitative results denote the false negative and the false positive activations, respectively.