

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

Hao Zhu¹, Yan Zhu¹, Jiayu Xiao¹, Tianxiang Xiao¹, Yike Ma¹, Yucheng Zhang¹, Feng Dai^{1†}

¹Institute of Computing Technology, Chinese Academy of Sciences

{zhuhao22z, fdai}@ict.ac.cn

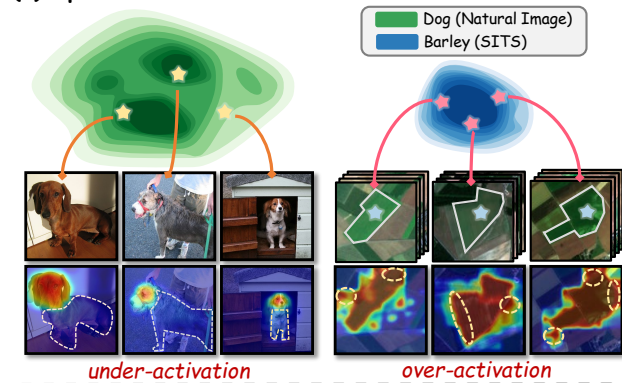
Abstract

Automated crop mapping through Satellite Image Time Series (SITS) has emerged as a crucial avenue for agricultural monitoring and management. However, due to the low resolution and unclear parcel boundaries, annotating pixel-level masks is exceptionally complex and time-consuming in SITS. This paper embraces the weakly supervised paradigm (i.e., only image-level categories available) to liberate the crop mapping task from the exhaustive annotation burden. The unique characteristics of SITS give rise to several challenges in weakly supervised learning: (1) noise perturbation from spatially neighboring regions, and (2) erroneous semantic bias from anomalous temporal periods. To address the above difficulties, we propose a novel method, termed **exploring space-time perceptive clues (Exact)**. First, we introduce a set of spatial clues to explicitly capture the representative patterns of different crops from the most class-relative regions. Besides, we leverage the temporal-to-class interaction of the model to emphasize the contributions of pivotal clips, thereby enhancing the model perception for crop regions. Building upon the space-time perceptive clues, we derive the clue-based CAMs to effectively supervise the SITS segmentation network. Our method demonstrates impressive performance on various SITS benchmarks. Remarkably, the segmentation network trained on Exact-generated masks achieves **95%** of its fully supervised performance, showing the bright promise of weakly supervised paradigm in crop mapping scenario. Our code will be publicly available [here](#).

1. Introduction

The launch of numerous public and commercial satellites provides broader opportunities to record, analyze, and predict the evolution of crop land [5, 17, 47, 50]. In this context, Satellite Image Time Series (SITS) with 10m resolution offered by the high-frequency Sentinel-2 (S2) satellites serve as a valuable data source for automated crop map-

(a) Spatial Noise Perturbation



(b) Temporal Erroneous Semantic Bias

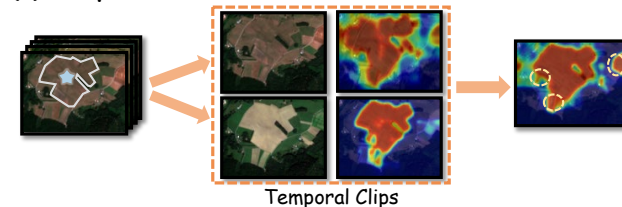


Figure 1. **Illustration of the two inherent issues arisen from spatial and temporal perspectives in SITS.** (a) shows noise perturbation from the spatial perspective. We visualize the high-level feature manifold of **Dog** (natural image) and **Barley** (SITS) to reveal the distinct spatial properties. The feature dimensions are reduced by t-SNE [48]. (b) shows the erroneous semantic bias induced by anomalous temporal clips. We denote the parcel regions with \star . The white circles refer to false positive activation regions in CAMs.

ping [6, 14, 22, 28, 35, 56]. The core of crop mapping lies in the semantic segmentation of crop parcels. Recently, many efforts are devoted to exploiting the versatile and powerful relation modeling capability of deep neural networks for this task [20, 21, 34, 46]. While such methods have shown a significant progress, they rely heavily on pixel-level manual annotation, which is notoriously complex and time-consuming [49]. The low resolution of satellite images and the indistinct boundaries between crop parcels complicate the annotation process. Even worse, the varying phenologi-

[†]Corresponding author.

cal cycles among crops require annotators to meticulously select appropriate acquisition times for annotation.

To address this, one promising solution is weakly supervised semantic segmentation (WSSS), which relies solely on time-efficient annotation form, *i.e.*, image-level categories. The image-level WSSS has commonly studied in natural image domain. Existing methods primarily extract Class Activation Map (CAM) [57] to generate pseudo labels for training semantic segmentation models. The intention of CAM is to identify the regions that highly contribute to the prediction of each class, revealing the shared patterns among the images within the same category. In context of natural images, CAM tends to highlight local discriminative parts of the target object. The reason is that the dispersed intra-class distribution enforces the model to learn a extremely sharp decision boundary, thereby the classifier weights tend to interact with feature representations within more discriminative object regions, as shown in Fig. 1a left. Thus the primary focus of researchers falls into expanding the CAM to identify the entire object semantics for natural images.

Different from WSSS for natural image, crop mapping for SITS mainly faces with two challenges: (1) from the spatial perspective, parcel objects within the same category share uniform appearance and color, exhibiting strong neighboring consistency. The intra-class compactness of local patterns looses the tolerance of the classifier to noise perturbation, resulting in pronounced over-activation phenomena in CAM, as shown in Fig. 1a right. The disparate characteristic lead to the advances in natural image domain cannot directly benefit the SITS crop segmentation. (2) From the temporal perspective, although different crops show distinct phenological cycles and varying characteristics, they may display similar appearance in some specific periods. This confusion imbues wrong semantic bias to the learning process of the model, thereby activating some undesired semantic regions in CAM. An intuitive illustration is shown in Fig. 1b, the temporal clips that deviate from the pivotal semantic affect the perceptual ability of CAM to correct crop regions.

In this work, we present *exploring space-time perceptive clues (Exact)*, a tailored WSSS framework for crop mapping, to cope with the challenges arised from the spatial and temporal aspects respectively. Firstly, we introduce a set of spatial clues to explicitly capture the patterns of different crops. Leveraging the filtered CAM as an indicator, we update representative clues by conducting spatial clustering from the most class-relative regions. These clues are then used to regularize the feature space via optimizing the contrastive objective, thereby sharpening the decision boundary of the model and mitigating the perturbations from illusory patterns. Secondly, to cope with the erroneous semantic bias caused by anomalous temporal periods, we propose temporal-aware affinity propagation to emphasize the contributions of pivotal clips for crop perception. In detail, we extract the temporal-

to-class attention from the model to reweight the temporal sequence embeddings. The modulated representations can be used to model temporal-aware pairwise affinity for propagation on the raw CAM, thus effectively suppressing the undesired semantic regions in a self-supervised manner.

Unlike existing WSSS methods that rely on classifier weights to generate CAMs, we exploit the well-updated space-time perceptive clues to derive the clue-based CAMs (CB-CAMs) as pseudo labels for segmentation. Compared to the raw CAM, the CB-CAMs (1) remarkably suppress the perturbations from spatial and temporal aspects, and (2) delineate the crop regions more precisely, thereby providing more reliable supervision for the subsequent segmentation.

Our main contributions can be summarized as follows:

- We introduce the WSSS paradigm to SITS crop mapping task to tackle the daunting annotation challenge. To the best of our knowledge, this is the first work that relies solely on image-level categories for crop segmentation.
- To overcome the drawbacks arised from the spatial and temporal aspects of SITS, we propose *Exact* that explores space-time perceptive clues to reduce the noise perturbation and rectify the wrong semantic bias, ultimately providing reliable supervision for SITS segmentation.
- We experimentally show that *Exact* achieves impressive performance on common benchmarks. Using *Exact*-generated labels for training, SITS segmentation model attains up to **95%** of its fully supervised performance. Our results significantly advance the upper bound of image-level WSSS technique compared to the other domains.

2. Related Work

Semantic segmentation on SITS. Automated crop monitoring through Satellite Image Time Series (SITS) has attracted great interest among researchers and demonstrated considerable social impact [14, 22, 23, 33, 36, 43, 44]. One of the challenging tasks is the semantic segmentation of agricultural parcels. The goal of the network is to learn a mapping function that assigns each pixel in the SITS to the corresponding crop type or background. Some works process SITS inputs by first extracting spatial information and then compressing the temporal dimension. For example, U-ConvLSTM [32] relied on U-Net [37] architecture to encode the spatial dimension, followed by a ConvLSTM [42] for the temporal dimension. Similarly, the FPN-ConvLSTM [34] replaced the U-Net with a Feature Pyramid Network [29] as the spatial encoder. U-TAE [20] compressed the temporal dimension through the outstanding temporal attention mechanism [21]. The other option to encode SITS is the *temporal-spatio* scheme, which first processes the temporal dimension and then extracts the spatial information. Rußwurm *et al.* [40] employed bidirectional LSTM to extract temporal features and then used CNNs to integrate spatial informa-

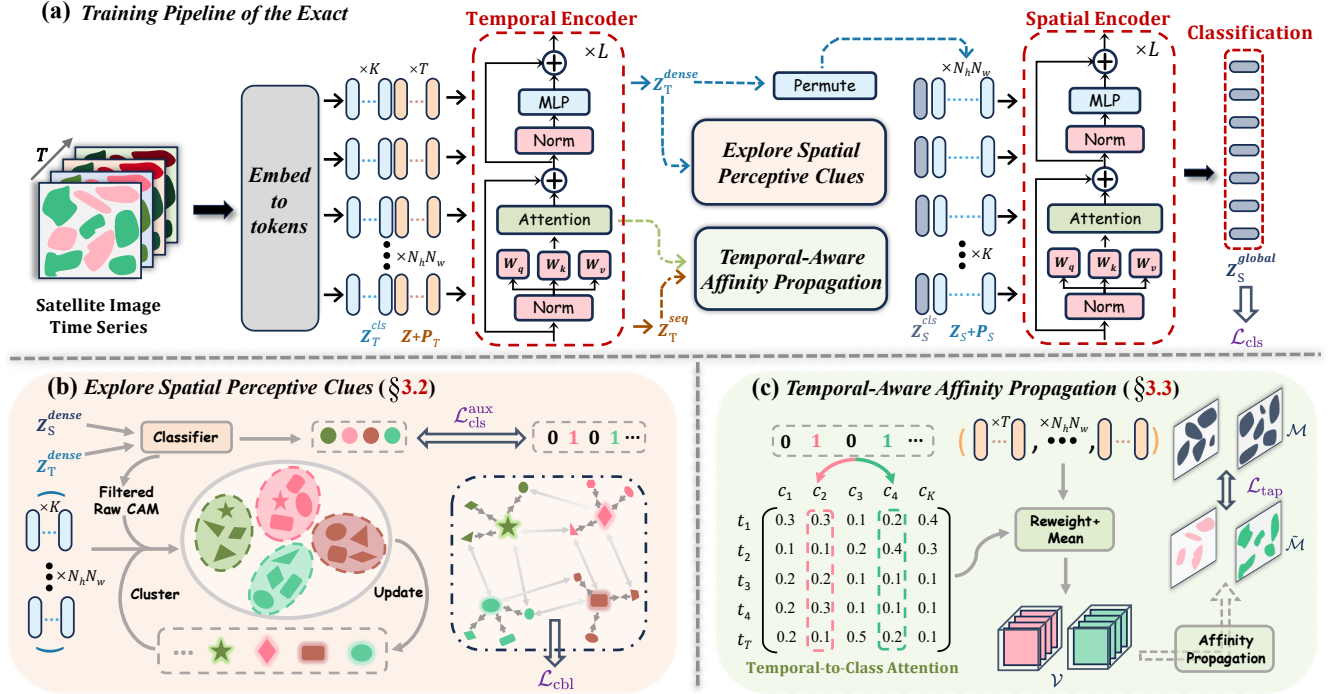


Figure 2. (a) **The training pipeline of *Exact***. We adopt the Temporal-Spatio scheme to handle the SITS input, which contains two transformer encoders. The first temporal encoder models interactions between acquisition times, then the followed spatial encoder discards the temporal dimension and models interactions between spatial positions. To overcome the difficulties arising from spatial and temporal aspects, we propose two novel technologies in temporal embedding space: (b) **Explore Spatial Perceptive Clues** (see §3.2) and (c) **Temporal-Aware Affinity Propagation** to rectify the wrong semantic bias (see §3.3).

tion. Recently, TSViT [46] proposed to borrow the powerful dependency modeling capability of the Vision Transformer (ViT) [16] to handle SITS, achieving state-of-the-art performance at lower computational cost. The TSViT also comprehensively illustrated the superiority of the temporal-spatio scheme. Taking a holistic view, we adopt this scheme to process SITS, aiming to achieve an optimal trade-off for automated crop mapping.

Weakly supervised semantic segmentation. Weakly supervised semantic segmentation (WSSS) with image-level labels (*i.e.* ground-truth object categories) has shown significant success in natural images [13, 18, 52, 53, 59]. Most advanced WSSS methods follow the three-steps pipeline of: 1) training a classification network with image-level labels, 2) obtaining class activation map (CAM) [9, 41, 57] from the well-optimized classification network as pixel-level coarse labels, 3) replacing ground-truth masks with the pseudo labels to train a off-the-shelf segmentation model. The objective of our work focuses on the first and second steps, *i.e.*, how to generate accurate pseudo labels. Numerous works has been proposed to address the under- and over-activation issues of the initial CAM [3, 10, 11, 26, 39, 55] in natural images. To derive the final pseudo labels, CAM requires cumbersome post-processing, including random walks [1, 2] and dense-

CRF [25] refinement. Thanks to these excellent works, the performance of image-level WSSS in natural images currently achieves 90% of pixel-level supervision. However, the WSSS networks designed for natural images require significant adaptation to be applied to SITS, and the results are still unsatisfactory due to the distinct data characteristics. In this work, we incorporate WSSS technique into the SITS and overcome the difficulties caused by inherent data properties.

3. Method

In this section, we first look more closely at the temporal-spatio scheme and class activation map technique (§3.1). Next, we introduce how to explore spatial perceptive clues (§3.2) and conduct temporal-aware affinity propagation (§3.3). Finally, we show the overall objective function and the clue-based CAM generation strategy of *Exact* (§3.4). The training pipeline of *Exact* is shown in Fig. 2.

3.1. Preliminaries

The temporal-spatio scheme that first processes the temporal dimension and then extracting the spatial information has shown significant superiority when dealing with SITS data. Following the TSViT [46], we use the variant ViT [16] as the backbone for both temporal and spatial encoders. We

consider a SITS input $\mathbf{X} \in \mathbb{R}^{T \times C \times H \times W}$ with T the length of time series, C the number of channels and $H \times W$ the spatial dimensions. The input \mathbf{X} is mapped into a sequence of patch tokens and reshaped to $\mathbf{Z} \in \mathbb{R}^{N_h \cdot N_w \times T \times d}$, where $N_h = \lfloor \frac{H}{h} \rfloor$, $N_w = \lfloor \frac{W}{w} \rfloor$ and $h \times w$ is the spatial extent for each patch. We then add the temporal position embeddings $\mathbf{P}_T \in \mathbb{R}^{T \times d}$ and concatenate the temporal multi-class tokens $\mathbf{Z}_T^{\text{cls}} \in \mathbb{R}^{K \times d}$ to obtain the input of temporal encoder:

$$\mathbf{Z}_T^{\text{in}} = [\mathbf{Z}_T^{\text{cls}}, \mathbf{Z} + \mathbf{P}_T], \mathbf{Z}_T^{\text{in}} \in \mathbb{R}^{N_h \cdot N_w \times (K+T) \times d} \quad (1)$$

here K denotes the number of categories, $\mathbf{Z}_T^{\text{cls}}$ and \mathbf{P}_T are repeated $N_h \cdot N_w$ times to match the spatial shape. With the output feature maps $\mathbf{Z}_T^{\text{out}} = [\mathbf{Z}_T^{\text{dense}} | \mathbf{Z}_T^{\text{seq}}]$ from the temporal encoder, we extract the first K tokens $\mathbf{Z}_T^{\text{dense}} \in \mathbb{R}^{N_h \cdot N_w \times K \times d}$ and permute the first and second dimensions to serve as input patch tokens $\mathbf{Z}_S \in \mathbb{R}^{K \times N_h \cdot N_w \times d}$ for spatial encoder. These inputs are then combined with spatial multi-class tokens $\mathbf{Z}_S^{\text{cls}} \in \mathbb{R}^{K \times 1 \times d}$ and spatial position embeddings $\mathbf{P}_S \in \mathbb{R}^{N_h \cdot N_w \times d}$ at all K spatial representations:

$$\mathbf{Z}_S^{\text{in}} = [\mathbf{Z}_S^{\text{cls}}, \mathbf{Z}_S + \mathbf{P}_S], \mathbf{Z}_S^{\text{in}} \in \mathbb{R}^{K \times (1+N_h \cdot N_w) \times d} \quad (2)$$

After the spatial encoder phase, we separate the output features $\mathbf{Z}_S^{\text{out}} = [\mathbf{Z}_S^{\text{global}} | \mathbf{Z}_S^{\text{dense}}]$ to align with different downstream tasks. For classification task, we feed the global tokens $\mathbf{Z}_S^{\text{global}} \in \mathbb{R}^{K \times 1 \times d}$ into the classifier to obtain classification logits. For segmentation task, to derive the dense prediction mask, the tokens $\mathbf{Z}_S^{\text{dense}} \in \mathbb{R}^{K \times N_h \cdot N_w \times d}$ are fed into the segmentation decoder.

Class activation map (CAM) [57] is widely used in WSSS to provide weak annotations that rely on image-level labels. Given a natural image, its feature maps $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D}$ are extracted by a classification backbone. To derive the classification score, the feature maps are average pooled and multiplied by the classifier weights $\mathbf{w} \in \mathbb{R}^{K \times D}$. CAM is generated by weighting and summing each channel in the feature maps with the classifier weights, as follows:

$$\mathcal{M}^k = \text{ReLU} \left(\sum_i \mathbf{w}_i^k \cdot \mathbf{F}_{:, :, i} \right), \quad \forall k \in K. \quad (3)$$

Following most WSSS methods, we normalize \mathcal{M}^k to the range $[0, 1]$ and apply a global threshold to filter out background pixels to obtain final pseudo labels. In the temporal-spatio network, we feed the dense tokens $\mathbf{Z}_T^{\text{dense}}$ and $\mathbf{Z}_S^{\text{dense}}$ into the classifier and compute the additional classification loss $\mathcal{L}_{\text{cls}}^{\text{aux}}$ to supervise the fused raw CAM for SITS. We provide more details in supplement.

3.2. Explore Spatial Perceptive Clues

CAM filtering. Given an input SITS \mathbf{X} and its image-level label $y \in [0, 1]^K$, we first compute its normalized fused

CAM $\mathcal{M} \in \mathbb{R}^{N_h \cdot N_w \times K}$ by the classifier weights and the output dense tokens of temporal and spatial encoder. We use two threshold scores μ_l and μ_h to filter out the reliable foreground, background and uncertain regions:

$$\hat{\mathcal{M}} = \begin{cases} 0, & \text{if } \mathcal{M} \leq \mu_l, \\ 1, & \text{if } \mathcal{M} \geq \mu_h, \\ \text{ignore}, & \text{otherwise.} \end{cases} \quad (4)$$

Clues clustering. To capture the compact patterns of different crops, we establish a group of class-wise representative prototypes, which later serve as perceptive clues to generate high-quality pseudo labels. In crop segmentation task, temporal features often provide more information than spatial context [46], so we choose to perform spatial clustering on temporal dense embeddings $\mathbf{Z}_T^{\text{dense}}$ over the whole dataset. Specifically, we build the class-wise positive prototype set $\mathcal{P}_{\text{pos}} = \{\mathbf{p}_{\text{pos}}^k \in \mathbb{R}^{N_p \times d}\}_{k=1}^K$ and negative prototype set $\mathcal{P}_{\text{neg}} = \{\mathbf{p}_{\text{neg}}^k \in \mathbb{R}^{N_p \times d}\}_{k=1}^K$, where N_p denotes the number of prototypes. If k -th class appears in a training batch, we update the prototypes \mathbf{p}^k via solving the optimal transport problem [8, 58]. Given a mapping matrix \mathbf{C}^k that represents the assignment between each pixel and its prototype, it can be referred as an element of the transportation polytope [4]:

$$\mathbb{C} = \{\mathbf{C}^k \in \mathbb{R}_+^{N_p \times N_k} | \mathbf{C}^k \mathbf{1} = \mathbf{u}, \mathbf{C}^{k\top} \mathbf{1} = \mathbf{r}\}, \quad (5)$$

where N_k denotes the number of pixels belonging to class k . $\mathbf{1}$ represents the vectors of ones in appropriate dimensions, \mathbf{u} and \mathbf{r} are the marginal projections onto the rows and columns of \mathbf{C}^k . We can maximize the following objective function to optimize the mapping matrix:

$$\underset{\mathbf{C}^k \in \mathbb{C}}{\text{maximize}} \text{Tr}(\mathbf{C}^{k\top} \mathbf{p}^k \mathbf{Z}^{k\top}) - \eta \sum_{n_p, n_k \in N_p, N_k} \mathbf{C}_{n_p n_k}^k \log \mathbf{C}_{n_p n_k}^k, \quad (6)$$

here $\mathbf{Z}^k \in \mathbb{R}^{N_k \times d}$ is the corresponding temporal dense embeddings belonging to class k , η controls the smoothness of entropy regularization term. The continuous approximate solution of Eq. (6) can be obtained through iteratively applying the Sinkhorn-Knopp algorithm [15]. Subsequently, we momentum update the n_p -th prototype of class k according to the assignment matrix and the embeddings:

$$\mathbf{p}_{n_p}^k = \alpha \mathbf{p}_{n_p}^k + \frac{1 - \alpha}{\|\mathbf{C}_{n_p}^k\|_1} \cdot (\mathbf{C}_{n_p}^k \mathbf{Z}^k), \quad (7)$$

here $\alpha \in [0, 1]$ is the momentum coefficient. We leverage the fused CAM $\hat{\mathcal{M}}$ as the pseudo labels to update the class-wise positive prototype set \mathcal{P}_{pos} and $1 - \hat{\mathcal{M}}$ for the negative set \mathcal{P}_{neg} . Notably, each prototype is not involved in gradient backpropagation to avoid the noise from the classifier.

Clue-based contrastive learning. Based on the spatial perceptive clues, we introduce the clue-based contrastive learning [7, 12] to regularize the embedding space. More specifically, for each temporal dense embedding $\mathbf{z}_{n_k}^k$ and its most

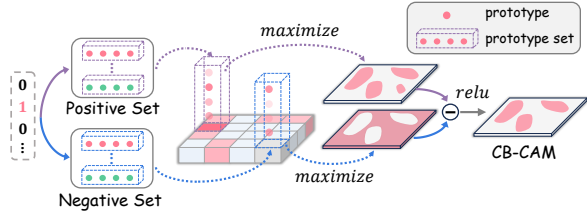


Figure 3. **Visualization of the clue-based CAM generation.** This process is performed after training the classification network.

relative prototype $\mathbf{p}_{n_p}^k$, we adopt the cosine distance to measure their similarity [24]:

$$S(\mathbf{z}_{n_k}^k, \mathbf{p}_{n_p}^k) = \frac{\mathbf{z}_{n_k}^k \mathbf{p}_{n_p}^{k\top}}{\|\mathbf{z}_{n_k}^k\| \cdot \|\mathbf{p}_{n_p}^k\| / \tau}, \quad (8)$$

where τ indicates the temperature parameter. Subsequently, we enforce the pixel embedding to closely match its prototype and be distinct from other prototypes:

$$\mathcal{L}_{cbl} = \sum_k \sum_{n_k} \mathbb{1}(\log(\sum_{\mathbf{p} \in \mathcal{P}^-} \exp S(\mathbf{z}_{n_k}^k, \mathbf{p})) - S(\mathbf{z}_{n_k}^k, \mathbf{p}_{n_p}^k)), \quad (9)$$

where $\mathcal{P}^- = \{\mathcal{P}_{\text{pos}} \cup \mathcal{P}_{\text{neg}} \setminus \mathbf{p}_{n_p}^k\}$ and $\mathbb{1}(\cdot)$ is an indicator function, being 1 if class k appears in image-level labels and 0 otherwise. Minimizing the above objective can pull pixel embedding closely to the semantic center and push it away from other illusory patterns, thereby sharpening the model decision boundary and mitigating the noise perturbation.

3.3. Temporal-Aware Affinity Propagation

In this section, we propose the temporal-aware affinity propagation to cope with the wrong semantic bias arised from anomalous temporal clips.

Temporal-aware affinity mining. In the temporal encoder, the input tokens are normalized and projected into query matrix $\mathbf{Q} \in \mathbb{R}^{(K+T) \times d}$ and key matrix $\mathbf{K} \in \mathbb{R}^{(K+T) \times d}$ respectively. Then the self-attention $\mathcal{A} \in \mathbb{R}^{(K+T) \times (K+T)}$ with respect to $N_h \cdot N_w$ sequences is computed as below:

$$\mathcal{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right). \quad (10)$$

Without additional computations nor supervision, we can explicitly extract the temporal-to-class attention $\tilde{\mathcal{A}} \in \mathbb{R}^{T \times K}$ from the self-attention \mathcal{A} , which represents the contributions of high-level representations to crop recognition at different temporal clips. We then reweight the temporal sequence embeddings $\mathbf{Z}_T^{\text{seq}}$ based on the normalized attention $\tilde{\mathcal{A}}$:

$$\mathcal{V}^k = \sum_{t=1}^T \tilde{\mathcal{A}}_t^k \cdot (\mathbf{Z}_T^{\text{seq}})_{:,t,:}. \quad (11)$$

Intuitively, the modulated high-level representation $\mathcal{V}^k \in \mathbb{R}^{N_h \cdot N_w \times d}$ enlarges the variation among different semantic crops, thereby modeling more precise pixel relations.

Affinity propagation and guidance. After obtaining the temporal-aware representation \mathcal{V} , we conduct affinity propagation to suppress erroneous semantic regions on the raw CAM. Particularly, the temporal-aware pairwise affinity of the pixel at position i can be estimated as follows:

$$\text{Aff}(\mathbf{v}_i^k, \mathbf{v}_j^k) = \exp\left(\frac{S(\mathbf{v}_i^k, \mathbf{v}_j^k)}{\sigma_i^k}\right), \quad (12)$$

here σ_i^k is the standard deviation of \mathbf{v}_i^k and \mathcal{N}_i indicates the local receptive fields (e.g., 8-way local neighbors [27]). We denoise the raw CAM via iteratively propagating the pairwise affinity from the temporal-aware representation \mathcal{V} :

$$\begin{aligned} \tilde{\mathcal{M}}_i^k &= \delta_i^k \sum_{j \in \mathcal{N}_i} \text{Aff}(\mathbf{v}_i^k, \mathbf{v}_j^k) \cdot \mathcal{M}_j^k, \quad \text{where} \\ \delta_i^k &= \frac{1}{\sum_{j \in \mathcal{N}_i} \text{Aff}(\mathbf{v}_i^k, \mathbf{v}_j^k)}. \end{aligned} \quad (13)$$

After obtaining the improved activation map, we align the raw CAM with it to effectively guide the learning process:

$$\mathcal{L}_{\text{tap}} = \sum_k \mathbb{1}(|\tilde{\mathcal{M}}^k - \mathcal{M}^k|). \quad (14)$$

By minimizing \mathcal{L}_{tap} , we can rectify erroneous activations in the raw CAM and incorporate temporal-aware priors into the embedding space, ultimately benefiting the perceptive clues.

3.4. Network Optimization and Clue-based CAM Generation Strategy

As shown in Fig. 2, the whole objective function for training the classification network consists of four components:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cls}}^{\text{aux}} + \lambda_1 \mathcal{L}_{\text{cbl}} + \lambda_2 \mathcal{L}_{\text{tap}}, \quad (15)$$

here \mathcal{L}_{cls} and $\mathcal{L}_{\text{cls}}^{\text{aux}}$ are the conventional binary cross entropy loss and λ_i denotes the weight to rescale the loss terms.

After training process, we perform per-pixel perception based on the well-updated space-time clues in temporal dense embedding space to generate clue-based CAMs (CB-CAMs). For each embedding \mathbf{z}_i in temporal dense embeddings $\mathbf{Z}_T^{\text{dense}}$, we measure the maximum similarity with the positive prototypes and minus the misguide activations with the negative prototypes:

$$\mathcal{Y}_i^k = \text{ReLU}\left(\max_{\mathbf{p}^+ \in \mathcal{P}_k^{\text{pos}}} S(\mathbf{z}_i, \mathbf{p}^+) - \max_{\mathbf{p}^- \in \mathcal{P}_k^{\text{neg}}} S(\mathbf{z}_i, \mathbf{p}^-)\right). \quad (16)$$

If the class k does not appear in a training image, we set \mathcal{Y}^k to all zeros. Fig. 3 gives a visualization of this generation process. Finally, the CB-CAMs \mathcal{Y} are filtered using a global background score to generate the pseudo labels, which are then used to train the SITS semantic segmentation network.

Method	Type	Germany [40]		PASTIS [20]		Method	Sup.	OA	mIoU	ratio
		OA	mIoU	OA	mIoU					
MCTFormer _{CVPR'22} [52]	<i>RGB</i>	70.6	56.3	66.7	49.6	ConvLSTM [42]	\mathcal{G}	78.2	50.1	
ViT-PCM _{ECCV'22} [38]		66.8	52.2	69.3	53.2	BiCGRU [40]		80.5	56.2	
LP-CAM _{CVPR'23} [13]		64.7	49.1	67.0	50.1	FPN-ConvLSTM [34]		81.9	59.5	
TSCD _{AAAF'23} [54]		68.4	52.7	67.2	51.3	Unet-3D [32]		82.3	60.4	-
DuPL _{CVPR'24} [51]		66.2	50.6	65.5	48.7	Unet-3Df [45]		82.1	60.2	
SeCOC _{CVPR'24} [55]		64.5	48.3	63.6	46.1	U-TAE [20]		82.9	62.4	
baseline		82.5	74.1	81.2	69.5	TSViT [46]		83.4	65.1	100%
+PAMR _{CVPR'20} [3]	<i>SITS</i>	83.9	76.3	82.0	71.2	baseline	\mathcal{P}	77.1	57.7	88%
+TS-CAM _{ICCV'21} [19]		81.1	70.5	80.4	67.6	+PAMR _{CVPR'20} [3]		78.5	58.7	90%
+SIPE _{CVPR'22} [11]		82.0	73.2	81.5	70.1	+TS-CAM _{ICCV'21} [19]		76.8	56.5	87%
+FPR _{ICCV'23} [10]		82.4	75.6	81.7	71.0	+SIPE _{CVPR'22} [11]		77.1	58.0	89%
+ours- <i>Exact</i>		88.3 \uparrow 5.8 80.6 \uparrow 6.5	84.1 \uparrow 2.9 75.6 \uparrow 6.1	+FPR _{ICCV'23} [10]	78.2	58.4		90%		
					+ours- <i>Exact</i>	80.3 \uparrow 3.2 61.8 \uparrow 4.1 95%				

(a) Pseudo labels performance of different methods on Germany and PASTIS *train* set.(b) Segmentation performance trained with pseudo labels generated by different methods on PASTIS *test* set.

Table 1. Comparisons with existing WSSS methods in OA(%) and mIoU(%). *RGB*: the networks designed for processing natural images. *SITS*: the networks designed for processing SITS (include four adapted modules from natural image domain). \mathcal{G} : various SITS segmentation networks supervised by ground-truth labels. \mathcal{P} : the TSViT segmentation model supervised by different pseudo labels from the methods in *SITS*. The *ratio* refers to the proportion of mIoU between fully supervised and weakly supervised of TSViT segmentation.

4. Experiments

4.1. Experimental Setup

Datasets & evaluation metric. We conducted comprehensive experiments on two widely used Satellite-2 time series crop recognition datasets to validate the effectiveness of our approach quantitatively. The *PASTIS* dataset [20] contains 2433 multi-spectral time series of size 128×128 , each series including 10 bands and 33 to 61 temporal observations. It consists of 18 crop types and a background category. Following the settings of [46], we used the fold-1 among the five folds provided in PASTIS. We partitioned each sample into multiple patches and assigned category labels according to the mask annotations (see the supplement for details). The *Germany* dataset [40] comprises 137k field parcels of time series imagery with 13 spectral bands. Each series contains 36 observations and is labeled with 17 crop types. We employed the mean Intersection over Union (mIoU) and pixel-wise overall accuracy (OA) as evaluation metrics, both widely used to measure segmentation performance.

Implementation details. The classification network was trained on 8 NVIDIA RTX 3090 GPUs with batch size 8 for 15k iterations. During the training stage, we used the AdamW optimizer [30] with an initial learning rate of $1e-3$ and cosine weight decay [31]. As the raw CAM were unreliable in early iterations, the gradients of \mathcal{L}_{cbl} were backpropagated after the 4k iteration. For the raw CAM filtering, the threshold score (μ_l, μ_h) was set to $(0.2, 0.4)$. Due to the compact intra-class patterns, we set the number of class-specific prototypes N_p to 2, and the momentum used to update the prototypes was set to 0.999. Moreover, the other hyper-parameters η , λ_1 , λ_2 and τ were empirically set to 0.05, 0.01, 0.015 and

0.1, respectively. We employed the original TSViT with a segmentation decoder as the semantic segmentation model for SITS. The training details of the segmentation model exactly followed the settings in [46] without any modifications. Please refer to supplementary materials for more details.

4.2. Experimental Results

Base Architectures. In the evaluation stage, we fuse the raw CAMs generated by the spatial dense embeddings $\mathbf{Z}_S^{\text{dense}}$ and temporal dense embeddings $\mathbf{Z}_T^{\text{dense}}$ as our baseline.

4.2.1. Comparison with other WSSS Methods on CAMs.

The key of the weakly supervised learning is to provide reliable training pseudo labels for segmentation network. Thus, we conducted comparative experiments on the *train* splits to validate the quality of pseudo labels.

Competing methods adaptation. We reimplemented six well-performing transformer-based WSSS methods designed for natural images as competing methods. To accommodate the temporal dimension, we transformed the input dimensions of SITS into a 3D format $(T \times C, H, W)$ for these methods. In addition, to enable a fairer comparison between *Exact* and existing WSSS works, we carefully selected four off-the-shelf modules from other WSSS methods in the natural image domain and adapted them to the temporal-spatio framework. These modules were also reimplemented within the temporal dense embedding space to align with our approach. We attempted various adaptation strategies, more details and analysis available in supplementary materials.

Quantitative results. Tab. 1a presents the mIoU and OA of the pseudo labels generated by other WSSS methods and our *Exact*. As can be seen, the WSSS methods designed for

Method	\mathcal{L}_{cbl}	\mathcal{L}_{tap}	CB-CAM	Recall	OA	mIoU
baseline				81.1	81.2	69.5
		✓		81.8	82.4	72.3
	✓	✓		82.5	83.0	73.4
ours-Exact	✓	✓	✓	84.8	84.1	75.6

Table 2. **Ablation study results of different components.** The baseline is a setting with only image-level classification loss.

Method	Spa.	Tem.	Recall	Precision	OA	mIoU
	✓		81.9	78.4	80.5	66.7
		✓	80.8	81.6	81.0	68.3
baseline CAM	✓	✓	81.1	83.5	81.2	69.5

Table 3. **Baseline CAMs from different embedding spaces.** Overall, the fused CAM shows the best performance.

Method	Low.	Ta.	Recall	Precision	OA	mIoU
baseline			81.1	83.5	81.2	69.5
	✓		80.9	84.0	81.1	69.9
ours-Exact		✓	81.8	85.6	82.4	72.3
	✓	✓	81.7	85.2	82.0	71.8

Table 4. **Impact of modeling affinity from different sources.** *Low.* and *Ta.* refer to low-level cues and temporal-aware affinity. The temporal-aware affinity remarkably reduces noise.

natural images only achieve a maximum 53.2% mIoU (refer to PASTIS) with the limited encoding capability for SITS. Furthermore, due to the distinct data characteristics, the off-the-shelf modules struggle to perform well on SITS, and may even disturb the learning process. These phenomena indicate that the advancements in natural image cannot be directly translated to benefit the SITS domain. By contrast, our *Exact* overcome the intrinsic challenges from SITS, delivering the best performance compared to others.

4.2.2. Segmentation results trained by pseudo labels.

After obtaining the pseudo labels, we use them as replacement for ground truth to train the SITS semantic segmentation network. We employ TSViT followed by a segmentation head as our segmentation model, which is the top-performing approach in fully supervised oracles.

Quantitative results. In Tab. 1b, we show that the semantic segmentation model can achieve 95% to its fully supervised performance (refer to mIoU) using pseudo labels generated by *Exact*. Our best results are 80.3% OA and 61.8% mIoU on PASTIS *test* set, which surpasses the baseline by 3.2% and 4.1%, respectively. The experiment results quantitatively imply that the labels quality outperforms other methods by a

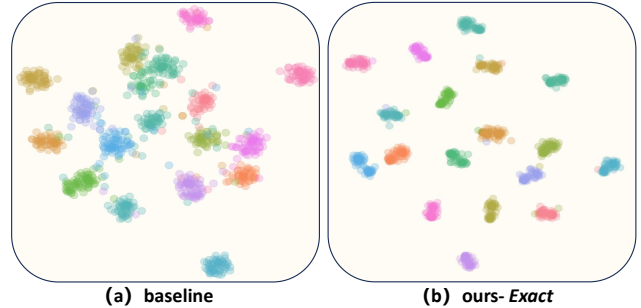


Figure 4. **Visualization of temporal feature spaces on PASTIS train set.** The feature dimensions are reduced by t-SNE [48].

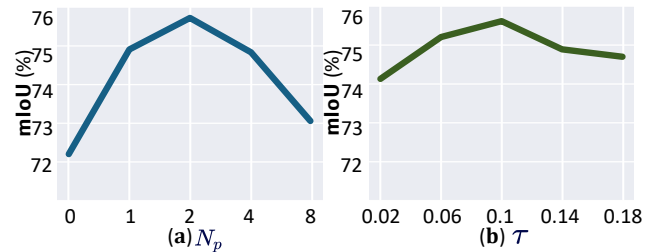


Figure 5. **Effect of the hyper-parameters.** (a) the number of class-specific prototypes N_p . (b) the temperature of similarity τ .

large margin. In contrast, WSSS networks in natural image have undergone a series of evolutions to barely achieve 90% of fully supervised performance. This suggests that SITS WSSS techniques offers greater potential in SITS scenario.

4.3. Ablation Studies

In this section, our primary purpose is to demonstrate the collective effectiveness of all components within our approach. We also choose the fused raw CAM as the baseline, and the CAM comparison results reported on PASTIS *train* set.

All the components matter. Our *Exact* consists of several components, including spatial perceptive clues exploration, temporal-aware affinity propagation and clue-based CAM generation strategy. We validate the contribution of each module, the results are presented in Tab. 2. We can see that the clue-based contrastive learning \mathcal{L}_{cbl} and the temporal-aware affinity propagation \mathcal{L}_{tap} improves the performance by 1.8% OA and 3.9% mIoU, respectively. This suggests that these two modules complement each other, regularizing the entire embedding space and ultimately sharpening the decision boundary. We provide an intuitive comparison in Fig. 4, after introducing both objectives, the intra-class features become more compact while the inter-class features are more separated. Besides, using the well-updated space-time clues to generate CAM (*i.e.*, CB-CAM) within the temporal dense embedding space brings a further improvement of 2.3% recall, 1.1% OA and 2.2% mIoU.

Baseline CAM from different embedding spaces. In Tab. 3, we investigate the performance of the raw CAMs derived

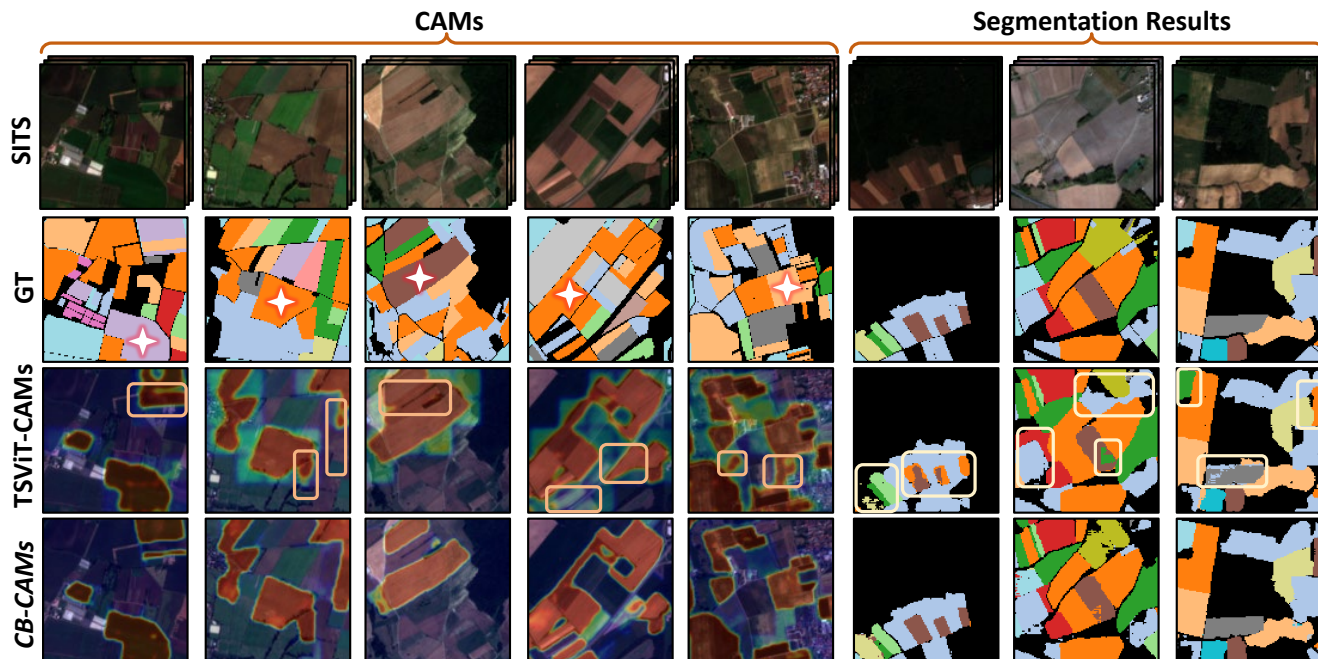


Figure 6. **Qualitative comparisons.** **Left:** CAMs comparisons on PASTIS *train* set. The stars represent the corresponding activation category. **Right:** Outputs of segmentation models trained with different CAMs on the PASTIS *test* set.

from different dense embedding spaces. The raw CAM obtained from the spatial dense embeddings $\mathbf{Z}_S^{\text{dense}}$ falls significantly behind that obtained from the temporal dense embeddings $\mathbf{Z}_T^{\text{dense}}$ in terms of precision, OA and mIoU. It demonstrates that the temporal correlations contain more critical information compared to the spatial context in this task. We combine the CAMs from both embedding spaces as our baseline (the last row in Tab. 3) and utilize the resulting pseudo labels to guide spatial clustering.

Effect of the temporal-aware affinity. Mining low-level cues affinity (*e.g.*, color and intensity) as additional guidance is prevalent in natural images. We also attempted to incorporate the low-level cues to model pairwise affinity in SITS, the results can be found in Tab. 4. Compared to our proposed temporal-aware affinity, modeling low-level cues offers only limited improvement. This is due to the presence of substantial noise interference within the low-level cues of SITS (*e.g.*, cloud cover and shadow). In contrast, our approach highlights the contributions of pivotal temporal clips within a high-level embedding space, effectively mitigating the wrong semantic bias.

Effect of the hyper-parameters. Fig. 5 shows the effects of the prototype number N_p and the temperature τ . Notably, utilizing only a minimal number of prototypes (*i.e.*, $N_p = 2$) yields optimal result, whereas increasing the number of prototypes beyond this point brings negative impact on the model. This indicates that the excessive number may enforce prototypes to focus on local discriminative patterns

within the class, leading to under-activation issue (the major difficulty in natural image). More analysis of parameters is specifically discussed in supplementary materials.

4.4. Qualitative Results

Fig. 6 presents the visual comparison between the baseline TSViT-CAMs and the proposed CB-CAMs generated by *Exact* on PASTIS dataset. The first five columns show the visualization of the CAMs. As we can see, our method remarkably suppress the erroneous regions, whose shape is closer to ground truth masks than baseline. Our high-quality pseudo labels subsequently enhance the segmentation performance compared to the baseline, as shown in last three columns. However, *Exact* still exhibits limitations in edge processing, which will be the focus in our future work.

5. Conclusion and Broader Impact

In this paper, we propose a tailored WSSS approach *Exact* to alleviate the daunting annotation challenge in SITS crop mapping task. *Exact* explores space-time perceptive clues to capture the essential patterns of different crop types, thereby overcoming the issues of spatial noise perturbation and wrong temporal semantic bias. Extensive experiments show the superiority of *Exact*-generated masks both quantitatively and qualitatively. We believe that our method marks a pioneering step in effectively applying WSSS technologies to SITS. If follow-up work can find and resolve the limitations under our framework, there is great potential that only image-level labels are needed for crop mapping in the future.

Acknowledgements

This work is supported by National Key R&D Program of China (2022YFD2001601) and National Natural Science Foundation of China (62372433).

References

- [1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 3
- [2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 3
- [3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020. 3, 6
- [4] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. 4
- [5] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *ECCV*, 2024. 1
- [6] Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. In *CVPR*, 2024. 1
- [7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 4
- [8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 4
- [9] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018. 3
- [10] 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. 3, 6
- [11] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *CVPR*, 2022. 3, 6
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4
- [13] Zhaozheng Chen and Qianru Sun. Extracting class activation maps from non-discriminative features as well. In *CVPR*, 2023. 3, 6
- [14] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, 2022. 1, 2
- [15] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 4
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [17] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa’s optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 2012. 1
- [18] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *CVPR*, 2022. 3
- [19] 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. 6
- [20] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *ICCV*, 2021. 1, 2, 6
- [21] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *CVPR*, 2020. 1, 2
- [22] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *CVPR*, 2024. 1, 2
- [23] Pengyu Hao, Yulin Zhan, Li Wang, Zheng Niu, and Muhammad Shakir. Feature selection of time series modis data for early crop classification using random forest: A case study in kansas, usa. *Remote Sensing*, 2015. 2
- [24] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Re-thinking federated learning with domain shift: A prototype view. In *CVPR*, 2023. 5
- [25] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 3
- [26] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *CVPR*, 2021. 3
- [27] Wentong Li, Yuqian Yuan, Song Wang, Wenyu Liu, Dongqi Tang, Jianke Zhu, Lei Zhang, et al. Label-efficient segmentation via affinity propagation. In *NeurIPS*, 2023. 5
- [28] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *CVPR*, 2024. 1
- [29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2
- [30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2022. 6

- [32] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *CVPR Workshops*, 2019. 2, 6
- [33] Yuchi Ma, Shuo Chen, Stefano Ermon, and David B Lobell. Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, 2024. 2
- [34] Jorge Andres Chamorro Martinez, Laura Elena Cué La Rosa, Raul Queiroz Feitosa, Ieda Del’ Arco Sanches, and Patrick Nigri Happ. Fully convolutional recurrent networks for multi-date crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021. 1, 2, 6
- [35] Sina Mohammadi, Mariana Belgiu, and Alfred Stein. Improvement in crop mapping from satellite image time series by effectively supervising deep neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023. 1
- [36] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, and Gérard Dedieu. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, 2016. 2
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2
- [38] 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. 6
- [39] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *CVPR*, 2023. 3
- [40] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 2018. 2, 6
- [41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3
- [42] Xingjian Shi, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 2, 6
- [43] Sofia Siachalou, Giorgos Mallinis, and Maria Tsakiri-Strati. A hidden markov models approach for crop classification: Linking crop phenology to time series of multi-sensor remote sensing data. *Remote Sensing*, 2015. 2
- [44] Haopeng Sun, Yingwei Zhang, Lumin Xu, Sheng Jin, and Yiqiang Chen. Ultra-high resolution segmentation via boundary-enhanced patch-merging transformer. *arXiv preprint arXiv:2412.10181*, 2024. 2
- [45] Michail Tarasiou, Riza Alp Güler, and Stefanos Zafeiriou. Context-self contrastive pretraining for crop type semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 6
- [46] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *CVPR*, 2023. 1, 3, 4, 6
- [47] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote Sensing of Environment*, 2012. 1
- [48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 1, 7
- [49] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 2020. 1
- [50] Curtis E Woodcock, Richard Allen, Martha Anderson, Alan Belward, Robert Bindschadler, Warren Cohen, Feng Gao, Samuel N Goward, Dennis Helder, Eileen Helmer, et al. Free access to landsat imagery. *Science*, 2008. 1
- [51] 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. 6
- [52] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 3, 6
- [53] Lian Xu, Mohammed Bennamoun, Farid Boussaid, Hamid Laga, Wanli Ouyang, and Dan Xu. Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3
- [54] Rongtao Xu, Changwei Wang, Jiayi Sun, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. In *AAAI*, 2023. 6
- [55] 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. 3, 6
- [56] Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, 2019. 1
- [57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2, 3, 4
- [58] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022. 4
- [59] Hao Zhu, Yan Zhu, Jiayu Xiao, Yike Ma, Yucheng Zhang, Jintao Li, and Feng Dai. Misa: mining saliency-aware semantic prior for box supervised instance segmentation. In *IJCAI*, 2024. 3