Elevation Guided Global and Local Smoothness for Unsupervised Semantic Segmentation in Remote Sensing Imagery

Kevin Qiu[1], Isabella Mebus Kishi de Oliveira[1], Dimitri Bulatov[1], Dorota Iwaszczuk[2]

[1] Fraunhofer IOSB Ettlingen, Germany - (kevin.qiu@iosb.fraunhofer.de)
[2] Technical University of Darmstadt, Civil and Environmental Engineering Sciences, Germany

Keywords: Multimodal Training, Self-Supervision, NDSM, Energy Minimization, Conditional Random Fields

Abstract

Unsupervised and self-supervised deep learning networks for semantic segmentation of images have made impressive progress in the last years. They can be trained without any labelled data and yet are able to effectively segment RGB images into meaningful semantic groups. In remote sensing, supplementary information, such as elevation, improves class separation by differentiating classes based to their height above ground. We take SmooSeg, a recently developed, state-of-the-art unsupervised network for semantic segmentation, and guide its training process by infusing elevation information into its projector and smoothness prior. This ensures global label consistency across the entire dataset and improves the segmentation performance, since patches of the same semantic group often exhibit similar elevation characteristics. We also extend the Conditional Random Field (CRF) to refine the low-resolution segmentation results in a post-processing step with elevation information. We introduce a second pairwise potential that encourages neighboring pixels with similar elevation to have the same label, ensuring local label consistency. Our multi-modal training strategy remains unsupervised and improves the segmentation performance on the ISPRS Potsdam-3 dataset by $+4.0\%$ in mIoU over the RGB-only SmooSeg baseline and by 4.4% when also using the multi-modal CRF post-processing. Collectively, our approach surpasses all state-of-the-art unsupervised segmentation networks that rely solely on RGB data for the Potsdam-3 dataset, highlighting the important role of elevation data in label-free segmentation for remote sensing applications.

1. Introduction

Semantic segmentation of urban scenes using aerial images is important for various applications, such as assessing land sealing, creating digital twins, conducting thermal simulations, and urban planning (Tuia and Camps-Valls, 2011; Marmanis et al., 2016; Bulatov et al., 2020). Traditional supervised semantic segmentation methods depend heavily on large and diverse labeled datasets, which are often limited in the field of remote sensing (Yuan et al., 2021). To overcome this, self-supervised and unsupervised networks present an interesting alternative. While these networks may produce inferior segmentation maps compared to fully supervised models, they can still be useful for providing initial segmentation estimates, conducting anomaly detection, or serving as a feature extractor for downstream tasks, all without the need for costly labeled data.

Self-supervised learning is usually defined as a subset of unsupervised learning and generates its own supervisory signals from the data itself, without explicit labels. A fundamental principle is semantic consistency, which asserts that an object's semantic label must remain invariant regardless of any photometric or geometric transformations it undergoes. Here, DINO (self-Distillation with No labels) (Caron et al., 2021) is a model that utilizes a teacher-student framework to improve the quality of learned representations by encouraging the student model to mimic the teacher model's outputs on unlabeled data. SmooSeg (Lan et al., 2024) employs a frozen DINO backbone for feature extraction for semantic segmentation. SmooSeg leverages the smoothness prior, positing that similar features within image patches should share semantic labels. By treating this segmentation task as an energy minimization problem, SmooSeg introduces a pairwise smoothness loss that encourages coherence within segments while preserving discontinuities between them. This is achieved through the modeling of relationships among observations, within and across images, using high-level features extracted from DINO. While DINO is a self-supervised network, SmooSeg is considered unsupervised since it essentially performs clustering on the DINO features by minimizing an energy function, even though it uses a teacher-student architecture that is often found in self-supervised networks.

As SmooSeg relies on low resolution features, a Conditional Random Field (CRF) refines the segmentation result. It smooths the labels using the unary potential from initial segmentation probabilities and using the pairwise potentials from neighboring pixels' labels and features. Here, unary potentials measure label fit, while pairwise potentials usually use RGB information to encourage similarly colored neighboring pixels to share the same label. Note that CRFs are class-agnostic, which is necessary because the order and semantics of semantic groups from unsupervised networks are random and undefined. Our contributions are as follows:

- We extend the Potsdam-3 dataset, adding elevation information in the form of the Normalized Digital Surface Model (NDSM).
- We develop a method to inject elevation information into the predictor and smoothness prior of SmooSeg, to help the network better separate classes (global smoothness).
- We explore parameter sensitivity through an ablation study focusing on the fusion parameter as well as other design decisions.
- Utilizing the framework established in Qiu et al. (2025), we integrate elevation information into the pairwise potential of the CRF post-processing step to further enhance the segmentation quality during inference (local smoothness).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-177-2025 | © Author(s) 2025. CC BY 4.0 License.

Our approach remains fully unsupervised, solely relying on the additional information and patterns contained in the elevation data with no prior knowledge or handcrafted rules. Thus, this might not be the most effective way to improve the segmentation results of an unsupervised model such as SmooSeg. However, the decision to maintain an unsupervised framework stems from academic interest in exploring the potential and recent advancements within this intriguing field. Unsupervised approaches are free from human bias, explore data-driven intrinsic patterns in the data, do not require any a-priori knowledge, and adapt easily to different domains.

2. Previous works

Semantic segmentation is a key area within deep learning and computer vision, where the goal is to classify each pixel in a given image. Numerous well-known supervised convolutional networks, like U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), and DeepLab (Chen et al., 2018) often utilize ResNet (He et al., 2016) backbones to effectively analyze remote sensing data. For instance, SegNet has been applied to the ISPRS Potsdam dataset in Song and Kim (2020). Remote sensing imagery frequently includes additional modalities beyond RGB, such as infrared and elevation data obtained from LiDAR or photogrammetry. The Normalized Digital Surface Model (NDSM), which represents height above ground, is particularly beneficial for distinguishing classes like buildings, ground, trees, and vehicles, as they usually share similar elevation characteristics. Although supervised convolutional networks are relatively straightforward to train and many studies explore the integration of additional channels (Qiu et al., 2022a; González-Santiago et al., 2023; Hong et al., 2020), their reliance on large labeled datasets poses a significant limitation.

In recent years, **self-supervised learning (SSL)** has gained traction in various fields, including remote sensing (Wang et al., 2022), following advancements in natural language processing. SSL allows for pre-training on extensive unlabeled datasets. A notable early example is SimCLR (Chen et al., 2020), which utilizes contrastive learning to maximize agreement between augmented versions of the same image while minimizing similarity between different images. Building on the principles of SSL, the work in González-Santiago et al. (2022) employs SSL for pixel-level classification of hyperspectral images, achieving strong performance with minimal labeled data.

Self-supervision techniques enable label free image segmentation networks, such as InfoSeg (Harb and Knobelreiter, 2021), which presents a method that maximizes mutual information between local and global high-level features through a two-step learning process. PiCIE (Cho et al., 2021) improves unsupervised image segmentation by employing pixel-level contrastive learning to enhance the understanding of pixel relationships. The previously mentioned methods are all CNN based. The introduction of vision transformers (Dosovitskiy et al., 2020) has further revolutionized computer vision methods by employing self-attention mechanisms. This has led to the development of DINO (Caron et al., 2021), which uses self-distillation and contrastive learning to generate high-quality visual representations from large datasets. Extending the capabilities of DINO, STEGO (Hamilton et al., 2022) introduces a segmentation head and CRF module for semantic segmentation. This approach distills unsupervised features from the frozen DINO backbone into semantic labels using a contrastive loss. Also based on DINO, HP (Seong et al., 2023) enhances unsupervised semantic segmentation by promoting task-specific training guidance and local semantic consistency. Similarly, EAGLE (Kim et al., 2024) focuses on object-centric representation learning, employing the spectral technique EiCue and object-centric contrastive loss using DINO to improve semantic accuracy. PriMaPs-EM (Hahn et al., 2024), also using DINO, enhances unsupervised semantic segmentation by using Principal Mask Proposals to decompose images into semantically meaningful masks, fitting class prototypes via a stochastic expectation-maximization algorithm. Lastly, SmooSeg (Lan et al., 2024) addresses the limitations of previous methods, like subpar global label coherence. It directly integrates a global smoothness prior into its framework, allowing for more coherent segmentation results.

Markov and **conditional random fields (CRFs)** have long been used for post-processing of semantic segmentation (Albert et al., 2017; Bulatov et al., 2019), especially for traditional methods like Random Forests or for specific landcover classes, like roads (Wegner et al., 2013). With the advancement of supervised, high-resolution deep networks, CRF have received less attention, since these networks already take neighborhood information into account, for example through their convolutional layers. However, if a network is purely image-based, the priors reflected in unary or pairwise potentials of the CRF can include additional channels or multi-modal information. For instance, in Qiu et al. (2022b), the incorporation of height information into a Markov Random Field (MRF) post-processing routine has proven beneficial for refining existing low-quality segmentations of cars.

3. Methods

Both SmooSeg and CRFs approach semantic segmentation as a dense prediction task, focusing on identifying a labeling function L that assigns a semantic category $l(\mathbf{f}) \subset \{1, 2, \dots K\}$ for each observation $\mathbf{f}$ (which may represent pixels, patches, or features) and where K is the number of categories or clusters. This can be framed within an energy minimization framework (Boykov et al., 2001) by:

$$L^* = \arg\min_L E(L), \text{ where } E_L = E_{\text{smooth}}(L) + E_{\text{data}}(L). \quad (1)$$

Here, E_{data} is a pointwise data term that quantifies the fit of $l(\mathbf{f})$ to the observation $\mathbf{f}$ while E_{smooth} serves as a pairwise smoothness term that encourages coherence among observations $\mathbf{f}$. We inject elevation information in the form of the NDSM into the smoothness terms of both SmooSeg and CRF. This smoothness term in SmooSeg enhances the consistency of label assignments across all $\mathbf{f}$ (global consistency) during training. In contrast, the smoothness term in CRFs focuses on smoothing labels in the spatial dimension by considering neighboring patches or pixels (local consistency). Figure 1 shows the overview of our proposed pipeline.

3.1 Integrating elevation into SmooSeg

SmooSeg's architecture consists of three main components: a feature extractor f_θ, a lightweight projector h_θ, and an asymmetric predictor g_θ, where θ are the learnable parameters. The frozen feature extractor (Caron et al., 2021) generates high-dimensional feature representations $X_i = f_\theta(I_i) \in \mathbb{R}^{C \times N}$, where C is the number of channels and I_i is the current (i-th, where $i \in \{1, \dots B\}$) image. For simplicity, we flatten the two spatial dimensions of an image into a single dimension N, the

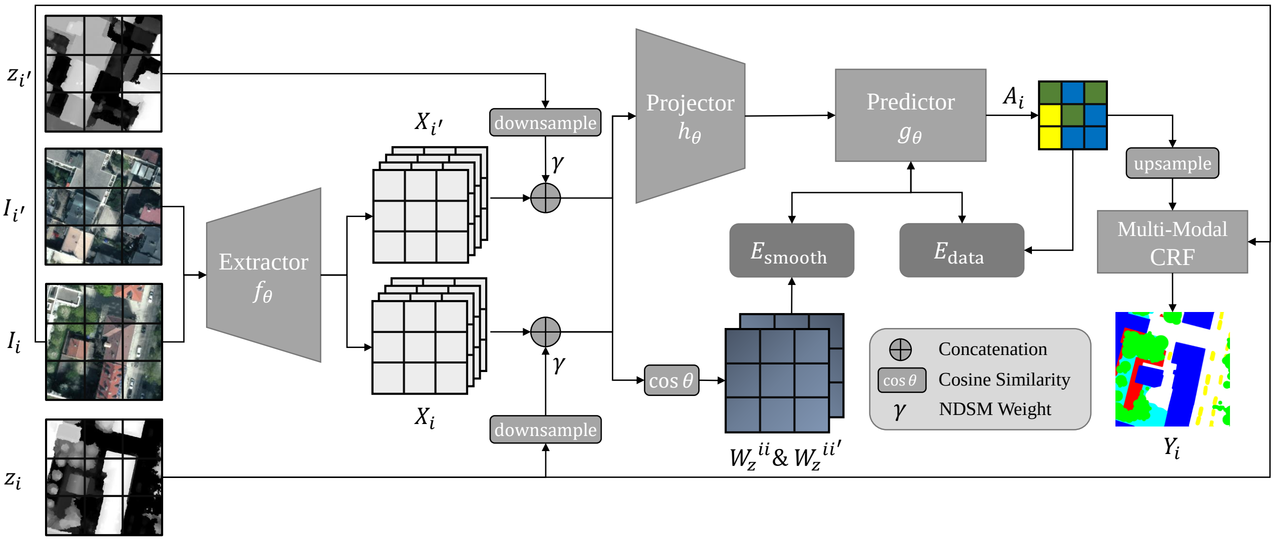


Figure 1. Overview of our elevation guided SmooSeg architecture for training on image I_i with elevation data z_i. Another image $I_{i'}$ is randomly chosen to act as a negative force for the smoothness prior. z is concatenated using a weight parameter γ with the respective output X from the frozen DINO extractor f_θ, fed to the projector and used to calculate the closeness matrices W. The latter are then used to calculate $E^{\text{SmooSeg}}_{\text{smooth}}$ for training the predictor g_θ and projector h_θ. The CRF is only applied during evaluation and also uses the elevation z_i in its smoothness term. The entire network is trained without any labels, including the frozen feature extractor f_θ.

number of patches of an image. The projector then maps X_i into a compact, low-dimensional embedding space, such that $Z_i = h_\theta(X_i) \in \mathbb{R}^{D\times N}$, where D is the reduced dimensionality. The output of DINO (we use the ViT-B/8 backbone) has a dimensionality of $C = 768$ and D in the projector is set to 64. The input image size during training is 224 × 224, yielding $N = 28 \cdot 28 = 784$ patches from f_θ. A smoothness term encourages the model to assign identical labels to patches with similar X_i, measured by closeness matrix W, effectively promoting global semantic continuity:

$$E^{\text{SmooSeg}}_{\text{smooth}} = \sum_{i=1}^{B} \sum_{p,q=1}^{N} W^{ii}_{pq} \cdot \delta(Y_{i,p}, Y_{i,q}), \tag{2}$$

where δ is a penalty function that takes the value of 1 if $Y_{i,p} \neq Y_{i,q}$, and 0 otherwise. $Y = \arg\max(A)$ are the predicted labels of the soft label A^t_i from the predictor. During training, A^t_i make δ differentiable, allowing δ to be defined in terms of the cosine similarity between the soft labels, where a larger δ indicates greater dissimilarity and thus a higher penalty. Furthermore, $W^{ii} \in \mathbb{R}^{N\times N}$ is the closeness matrix for the image I_i with itself to capture the relationships between its image patches based on their feature representations. Specifically, each element W^{ii}_{pq} of the closeness matrix is calculated using the cosine similarity of the feature vectors coming from f_θ corresponding to patches p and q of the image I_i, by:

$$W^{ii}_{pq} = \frac{X_{i,p} \cdot X_{i,q}}{\|X_{i,p}\| \, \|X_{i,q}\|}, \tag{3}$$

where $X_{i,p}$ and $X_{i,q}$ are the normalized high-level feature vectors for patches p and q, respectively. A higher value of W^{ii}_{pq} indicates a greater similarity between patch p and q, suggesting that these patches are likely to share one semantic label. While W^{ii} is used to calculate smoothness within images, $W^{ii'}$ calculates it across images i and a randomly chosen image i'. A second smoothness term with $W^{ii'}$ prevents the model from converging to a trivial solution.

To introduce the elevation information z_i, we concatenate the extractor output X_i with z_i in the form of an NDSM. This NDSM is down-scaled bilinearly to match the spatial dimension of X_i. The NDSM is in its original units, where each value corresponds to real-world measurements (e.g., a value of one represents one meter elevation above ground). We use a parameter γ to weigh the NDSM channel.

$$X_{z,i} = \begin{bmatrix} X_i \\ \gamma z_i \end{bmatrix} \in \mathbb{R}^{(C+1)\times N}. \tag{4}$$

This vector $X_{z,i}$ is given to the projector h_θ and is used to calculate the new closeness matrices W^{ii}_z and $W^{ii'}_z$ according to Eq. (3). In computing cosine similarity of two patches, adding a zero entry in both vectors does not affect similarity. Non-zero similar values increase similarity due to better alignment, while differing values decrease it. This is beneficial for our approach, as we want to avoid misclassifying ground-level classes (e.g., grass, asphalt, car) based solely on similar elevation values. Instead, we aim for notable (dis-)similarity between patches with (non)similar high elevation values, as these primarily represent the building class. We therefore do not need an offset parameter. Figure 2 shows down-scaled NDSM patches before weighting and concatenation with X_i along with the RGB input and ground truth of the Potsdam-3 dataset (see Section 4.1).

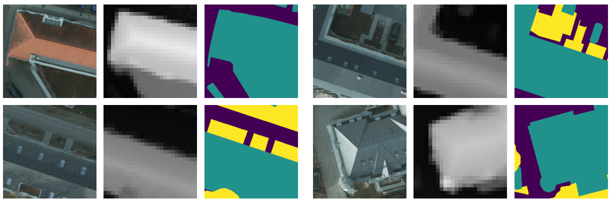


Figure 2. RGB (224 × 224), NDSM z (28 × 28) in Eq. (4) for training SmooSeg, and ground truth (only used for evaluation) of Potsdam-3. The usefulness of the NDSM in class delineation is immediately obvious, especially for buildings.

Like the original architecture, to ensure stability during optimization, zero-mean normalization is applied to the closeness matrices: $\bar{W}_{z,p} = W_{z,p} - \frac{1}{N}\sum_q W_{z,pq}$. This normalization balances the positive and negative influences in the optimization process, thereby enhancing the performance of the smoothness prior. Furthermore, dropout is applied to $X_{z,i}$ in the projector. The data term E_{data} remains unchanged wrt to Lan et al. (2024):

$$E_{\text{data}}^{\text{SmooSeg}} = -\sum_{i=1}^{B}\sum_{p=1}^{N}\sum_{k=1}^{K} I(Y_{t,i,p} = k)\log(A_{s,i,p,k}), \tag{5}$$

where summation takes place over cluster indices (K), patches (N) and images (B), and A_i^s represents the soft label assignments produced by the student branch in the predictor. There, SmooSeg employs an asymmetric teacher-student architecture, where the teacher predictor generates pseudo labels Y_t for the student predictor to learn from. Both teacher and student networks are updated using an exponential moving average to stabilize training. The smoothness within one image, across images, and the data term are summed and minimized according to Eq. (1). In evaluation and inference, only the teacher predictor that outputs $A_i^t \in \mathbb{R}^{K\times N}$ is used, where K is the number of clusters. For more details on the SmooSeg architecture, please refer to the original paper Lan et al. (2024).

We use the default Potsdam-3 settings and parameters as the authors of SmooSeg. As explained in more detail in Section 4.4, we set $\gamma = 2.5$. We train and evaluate on Windows and an Nvidia V100 GPU with 16 GB of memory.

3.2 Integrating elevation into CRF

The ViT-B/8 DINO extractor used in SmooSeg lowers the image dimension by a factor of 8 from 224×224 to 28×28, resulting in segmentations that are quite low resolution. For evaluation, the segmentations have to be upscaled, and local consistency has to be considered. This is necessary since SmooSeg operates on the patches individually without considering neighboring patches. Similarly to STEGO (Hamilton et al., 2022), the authors of SmooSeg add a CRF module for post-processing. CRFs usually consider the RGB values of neighboring pixels and their (pseudo) probabilities to ensure that close pixels with similar color values tend to have the same label. this refines a coarse segmentation map based on the higher resolution RGB image. Since neighboring pixels with similar elevation values also tend to have the same label, we also integrate the elevation z, here without downsampling, into the CRF. To clarify this, we provide a brief theoretical overview: The soft label assignment output of the teacher branch A_i^t is converted to unary potentials by applying the negative logarithm. These unary potentials, stored in $E_{\text{data}}^{\text{CRF}}$ (see Eq. (1)), must be balanced with piecewise-smoothness priors that promote label consistency among neighboring data points. These are stored in $E_{\text{smooth}}^{\text{CRF}}$:

$$\begin{aligned} E_{\text{smooth}}^{\text{CRF}} &= \sum_{\mathbf{x},\mathbf{y}\in\mathcal{N}} \omega_p\left(\mathbf{x},\mathbf{y},\mathbf{f}\right) d_p(l_{\mathbf{x}}, l_{\mathbf{y}}),\ \text{where} \\ \omega_p\left(\mathbf{x},\mathbf{y},\mathbf{f}\right) &= \lambda_{\mathbf{f}} \exp\left(-\frac{d_{\mathbf{xy}}^2}{2\sigma_{\mathbf{xyf}}^2} - \frac{\|\mathbf{f_x} - \mathbf{f_y}\|^2}{2\sigma_{\mathbf{f}}^2}\right) \\ &+\lambda_{\mathbf{xy}} \exp\left(-\frac{d_{\mathbf{xy}}^2}{2\sigma_{\mathbf{xy}}^2}\right), \text{and } d_p(l_{\mathbf{x}}, l_{\mathbf{x}}) = l_{\mathbf{x}} \neq l_{\mathbf{x}}, \end{aligned} \tag{6}$$

where $d_{\mathbf{xy}} = \|\mathbf{x} - \mathbf{y}\|$ (Euclidean distance), and the neighborhood $\mathcal{N}$ is theoretically fully connected; however, in practice, the decay in the negative exponent is rapid.

In brief, the equation reveals that the penalty for assigning different labels to adjacent pixels $\mathbf{x}$ and $\mathbf{y}$ is significantly stronger when these pixels are close together and/or share similar feature vectors. The nature of the relationship (whether it is an *and* or *or*) depends on the values of λ and σ; notably, $\exp(-\cdot/\sigma^2)$ approaches zero for small values and remains constant for larger values of σ. Consequently, the utilized implementation [1] of the CRF with this specific smoothness function offers substantial flexibility. Nonetheless, it is constrained by the inability to apply it to non-gridded graphs, use other than three-channel images, or employ location- and feature-dependent terms that deviate from Gaussian kernels. To minimize implementation complexity across the workflow, we utilized the CRF defined in the previous equations, with two terms representing the distinct data modalities:

$$\omega_p\left(\mathbf{x},\mathbf{y},\mathbf{f}\right) = \omega_p\left(\mathbf{x},\mathbf{y},\mathbf{j}\right) + \omega_p\left(\mathbf{x},\mathbf{y},\mathbf{z}\right), \tag{7}$$

where $\mathbf{j}$ and $\mathbf{z}$ represent the RGB image and the NDSM feature, respectively, converted into a three-channel format by replicating along the third dimension. The values for all eight parameters in the RGB+NDSM CRF configuration were determined through a parameter search method using validation data, as shown in Table 1. The parameters for the RGB CRF are taken from SmooSeg (Lan et al., 2024).

Table 1. The parameters for the default CRF with RGB and the multi-modal CRF with RGB + NDSM.

	Gaussian		**Pairwise RGB**			**Pairwise NDSM**		
CRF modality	$\sigma_{\mathbf{xy}}$	$\lambda_{\mathbf{xy}}$	$\sigma_{\mathbf{x}yj}$	$\sigma_{\mathbf{j}}$	$\lambda_{\mathbf{j}}$	$\sigma_{\mathbf{xyz}}$	$\sigma_{\mathbf{z}}$	$\lambda_{\mathbf{z}}$
RGB	1	3	67	3	4	-	-	-
RGB+NDSM	4	3.75	67	3	8	3.35	30	0.8

4. Results

4.1 Dataset

The dataset employed for this study is the photogrammetric ISPRS Potsdam dataset (Rottensteiner et al., 2014). This dataset features a ground sampling distance (GSD) of 5cm and comprises of 38 orthophotographic images, each sized at 6000 × 6000 pixels, covering approximately 2.5 square kilometers of the German city of Potsdam. In addition to RGB imagery, it also includes a near-infrared channel and a Digital Surface Model. The segmentation task involves six classes: buildings, cars, low vegetation, trees, impervious surfaces, and clutter.

Potsdam-3 is a version of this dataset for three category unsupervised image segmentation. The GSD is doubled to 10 cm by downsampling. This dataset consists of 8550 patches of size 200 × 200 and is also used by InfoSeg, STEGO and other unsupervised segmentation networks shown in Table 2. The patches are zero-padded to 224 × 224 for the DINO based networks since that is the resolution DINO expects. The dataset lacks elevation data. Therefore, we first derive the DTM (Digital Terrain Model) from the DSM by employing the methodologies outlined in Bulatov et al. (2014). Then, the NDSM is then calculated by subtracting the DTM from the DSM. We apply identical scaling and cropping operations as in Potsdam-3, to create the 8550 individual NDSM patches from the 38 aerial images. Subsequently, we matched and integrated the NDSM patches with the respective RGB patches, ensuring that the RGB images and the dataset split remain identical and therefore the models comparable. Hungarian matching (König, 1916)

[1] https://github.com/lucasb-eyer/pydensecrf

is used to align and fuse the six classes of Potsdam to the $K = 3$ predicted semantic groups of SmooSeg. Then, evaluation metrics such as mIoU (mean Intersection over Union) and Accuracy are computed using the three fused classes.

4.2 Quantitative findings

Table 2 compares the results on Potsdam-3 of various state-of-the-art unsupervised semantic segmentation models based on DINO with the elevation integrated SmooSeg. Reproducing the SmooSeg on our setup yields slightly worse (70.1% vs. 70.3% mIoU) results than reported in Lan et al. (2024). Integrating the NDSM into SmooSeg drastically improves the results by +4.0% to an mIoU of 74.1%. Further integrating the NDSM into the CRF improves the mIoU further by +0.4%, reaching an mIoU of 74.5%. The NDSM integrated SmooSeg surpasses all other RGB only SOTA methods by at least +3.4% based on the same backbone (DINO ViT-B/8), like EAGLE (Kim et al., 2024) or PriMaPs-EM (Hahn et al., 2024).

Table 2. SmooSeg performance with(out) NDSM and other unsupervised, RGB only models on the Potsdam-3 dataset. All models are based on the DINO ViT-B/8 backbone.

Method (DINO backbone)	**Acc.**	**mIoU**
DINO ViT-B/8 (Caron et al., 2021)	66.1	49.4
STEGO (Hamilton et al., 2022)	77.0	62.6
HP (Seong et al., 2023)	82.4	69.1
PriMaPs-EM+HP (Hahn et al., 2024)	83.3	71.0
EAGLE (Kim et al., 2024)	83.3	71.1
SmooSeg (Lan et al., 2024)	82.7	70.3
SmooSeg (reproduced)	82.6	70.1
+ NDSM in SmooSeg	85.1 (+2.5)	74.1 (+4.0)
+ NDSM in SmooSeg & CRF	85.3 (+2.7)	74.5 (+4.4)

Table 3 shows the mIoU scores broken down into the three semantic groups, approximately corresponding street, building and vegetation. The building class improves by +8.2% with NDSM in SmooSeg and further by +0.8% with the NDSM-integrated CRF. Similar improvements can be seen in the street class. Conversely, the integration of the NDSM leads to a slight decline of approximately -2% in the IoU scores for the vegetation class. This may be attributed to the heterogeneous nature of the vegetation class, which includes both low vegetation (grass), and high vegetation (trees). Consequently, the elevation characteristics of this class are more variable, whereas the building and street classes exhibit more consistent elevation values, resulting in more substantial improvements in their respective scores.

Table 3. Individual IoU scores for the three semantic groups, roughly corresponding to street, building and vegetation.

Method (DINO backbone)	**Street**	**Build.**	**Veg.**	**mIoU**
SmooSeg (reproduced)	64.1	68.1	**78.2**	70.1
+ NDSM in SmooSeg	70.0	76.3	76.0	74.1
+ NDSM in SmooSeg & CRF	**70.4**	**77.1**	76.0	**74.5**

4.3 Qualitative findings

Figure 3 shows predictions of some of the patches of the Potsdam-3 dataset using the baseline method SmooSeg with only RGB and our improved version and CRF with NDSM. The baseline SmooSeg classifies cars as part of the building class and struggles with erroneous detections of buildings, resulting in both false positives and false negatives, while also exhibiting unclear building boundaries. The NDSM-enhanced SmooSeg reclassifies cars into the street class, has more accurate building boundaries, and produces fewer erroneous detections. Vegetation is sometimes confused with street, reflecting the slightly worse scores for the vegetation class as mentioned in Section 4.2.

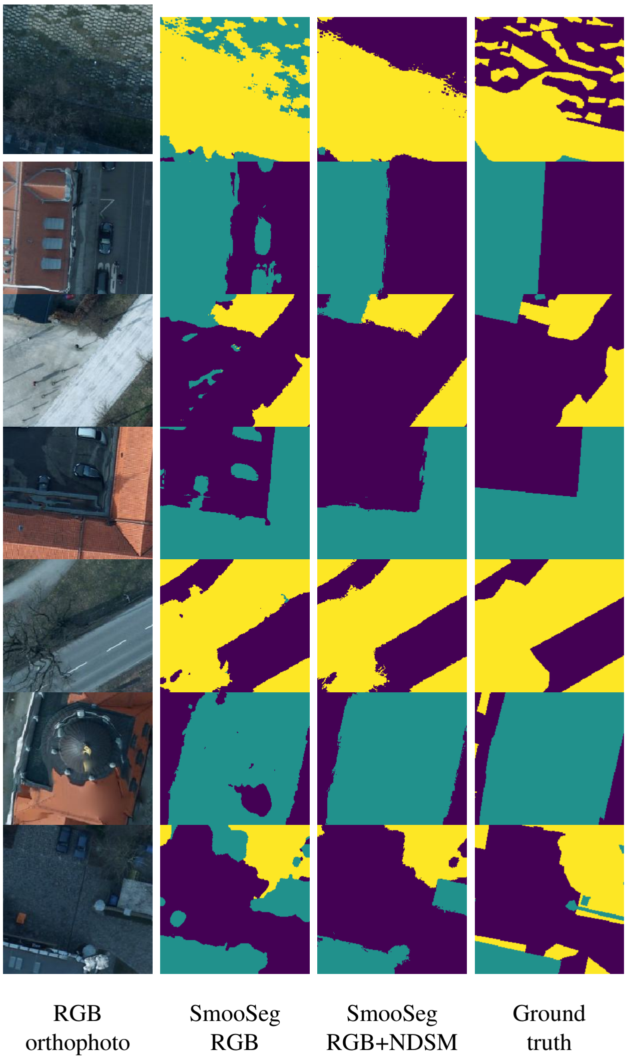
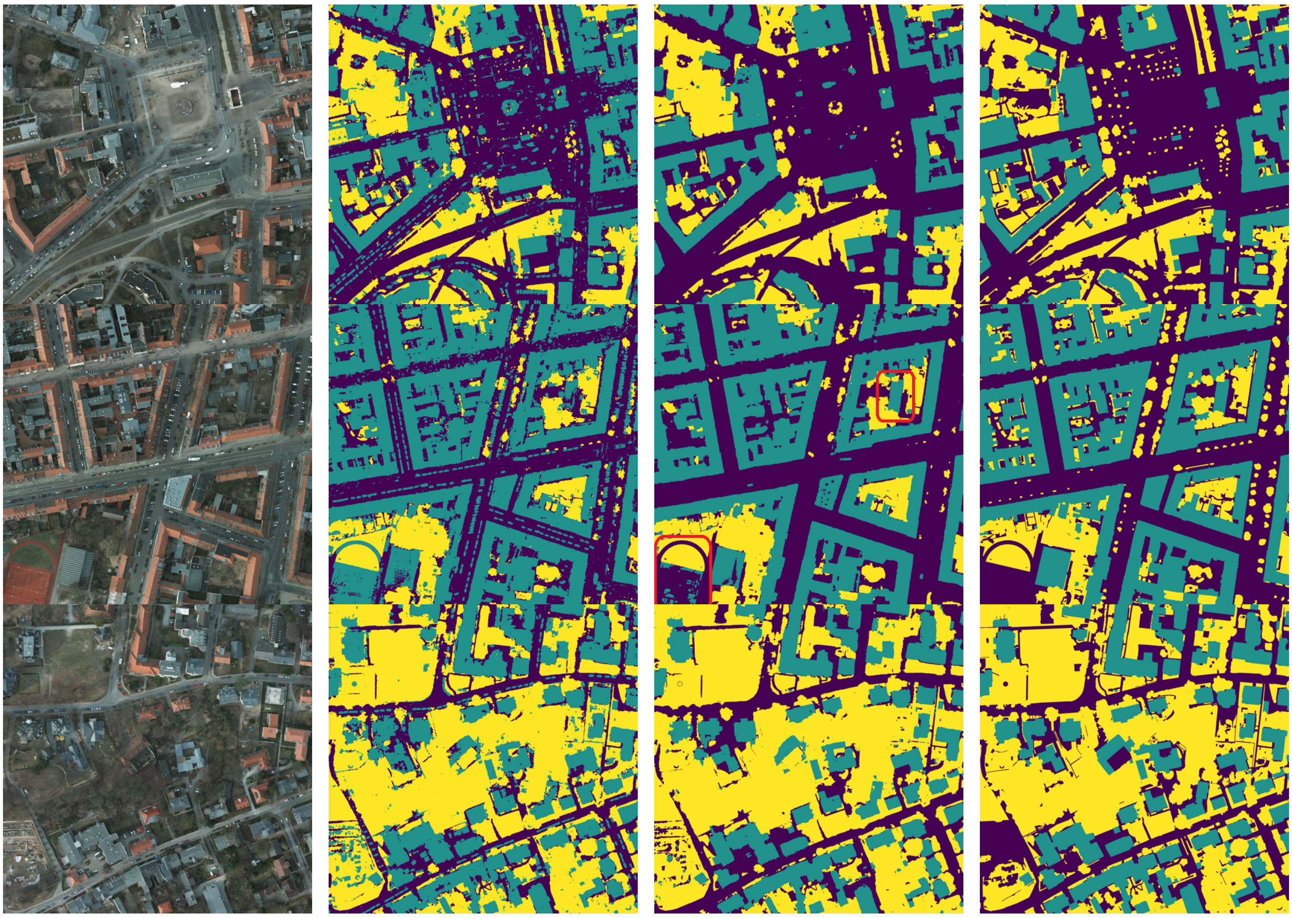


Figure 3. Qualitative results of baseline SmooSeg with CRF and our NDSM enhanced SmooSeg and multi-modal CRF compared to the ground truth on patches of the Potsdam-3 dataset. Color legend: **Street**, **Building**, **Vegetation**.

We also run inference on the full Potsdam images (scaled to 3000 × 3000). We employ an overlap of 40 pixels (equivalent to 4m GSD) when dividing the full images into patches of 200 × 200 for inference, ensuring a seamless transition in the stitched prediction image. CRF is then applied across the entire image. Figure 4 shows the resulting predictions of the default RGB SmooSeg and our NDSM-enhanced SmooSeg and CRF. Similar to the individual patches, the results on the full images demonstrate a significant improvement with our method compared to the baseline, with better building outlines and fewer artifacts. For instance, on the lower left of the second image (marked in red), a soccer field and running track are incorrectly classified as a building but are accurately identified as a street using our method. Additionally, a building located in the interior block of the upper right quadrant (marked in red) is only fully detected with our approach. SmooSeg, in general, struggles with trees, as the Potsdam dataset was captured in winter. Furthermore, street and low vegetation are occasion-



RGB orthophoto | SmooSeg (RGB) | SmooSeg (RGB+NDSM) | Ground truth

Figure 4. RGB orthophotos and qualitative results of baseline SmooSeg and our NDSM enhanced SmooSeg and multi-modal CRF compared to the ground truth on full images of the ISPRS Potsdam dataset. Color legend: **Street**, **Building**, **Vegetation**.

ally confused due to their ambiguous characteristics, like on unpaved roads.

4.4 Ablation study

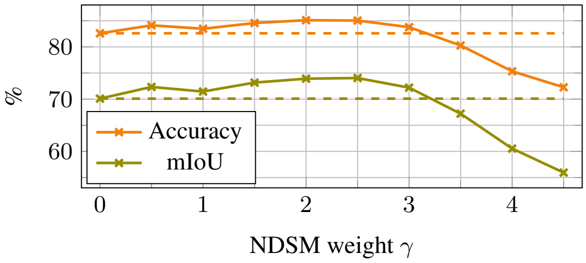


Figure 5. Sensitivity study of the NDSM weight γ from Eq. (4) and the Accuracy and mIoU scores with DINO in the NDSM enhanced SmooSeg with the default CRF. The dotted lines represent the baseline (Lan et al., 2024) without NDSM.

Figure 5 shows our sensitivity study of the NDSM parameter γ on Accuracy and mIoU. $\gamma = 0$ is equivalent to the RGB-only original SmooSeg. Since the semantic groups are fairly balanced, both the Accuracy and mIoU follow a similar pattern. With increasing γ, the scores increase at the beginning, remain quite stable in a large interval between 1.5 and 3, and then drop off. The best result is achieved with $\gamma = 2.5$.

Since the elevation z is injected separately into the projector h_θ and smoothness matrices, we evaluated a configuration with two separate weights. However, after optimizing these weights, the mIoU and accuracy scores remain similar. The parameter space is fairly broad as well, as a similarly wide range of values produces good performance. Due to this stability and to avoid an extra parameter, we use a single value $\gamma = 2.5$ to obtain the results in Sections 4.3 and 4.2 of this paper.

The input $X_{z,i}$ into the projector is subject to random dropout during training. While that makes sense for the DINO features, it might not for a measurable, physical feature like the NDSM. Still, we found no discernable difference to the final performance of applying dropout to the NDSM or not and decided to keep it for (theoretically) improved adaptability and reliability of the model towards the NDSM.

5. Discussion and Conclusion

A clear and strong improvement of SmooSeg is achieved by our method of incorporating the NDSM into the unsupervised learning mechanism. The NDSM is injected into the smoothness term and the projector with a weight γ, and also in the CRF post-processing step. We achieve a combined improvement of +4.4% in mIoU on the Potsdam-3 dataset over the baseline (Lan et al., 2024). Interestingly, the addition of NDSM pushes the car class from the building into the street class, which happens to be a welcome change from an applications perspective. This is not an issue for evaluation because of Hungarian matching. EAGLE (Kim et al., 2024), for example, also assigns cars to the street class. The building class shows clear improvements, featuring sharper edges, reduced false detections, fewer missing parts, and less confusion with the street class, which in turn enhances the performance of the latter as well. The vegetation class, unfortunately, performs slightly worse. With minimal added complexity, training (20 min on Potsdam-3) and inference – 4 min per 3000×3000 image with overlap, due to CRF – remain virtually unchanged. The only additional time-consuming step is computing the DTM once if it is not already available.

Improvements through the multi-modal CRF, however, are modest. The DSM of the Potsdam dataset is somewhat blurry and contains artifacts and outliers, which hinders a high-resolution refinement through a CRF. Using the multi-modal CRF on a dataset with a higher quality DSM is expected to result in a more significant improvement. In future work, CRFs with advanced, possibly learnable terms may represent an interesting research direction as well.

In Qiu et al. (2022a), the enhancement through the addition of NDSM with RGB in two supervised deep learning models on the ISPRS Potsdam dataset was relatively modest at around 1%. Similarly, Koppanyi et al. (2019) report only minor improvement by adding NDSM information using supervised learning. Improvements by only +0.5% were observed in the middle fusion experiment of Audebert et al. (2018) on a different dataset. Meanwhile, in this study, including NDSM alongside unsupervised segmentation significantly improved the mIoU by $+4.4\%$. While the numbers are not directly comparable, this seems to suggest that NDSM data plays a more important role in unsupervised learning scenarios, where the model has to rely on the input data itself for effective class separation rather than on reference data, indicating that the NDSM is a valuable modality for this purpose. However, as observed in the vegetation class, the NDSM may also negatively affect performance. This could be mitigated with methods such as adaptive weighting or hierarchical clustering. Further research is also necessary to investigate the influence of NDSM when segmenting into more than three semantic groups, especially when some of these groups may exhibit less distinct and less characteristic elevations.

One consideration of our approach is the necessity to determine the weight γ for the NDSM, as well as the parameters for the multi-modal CRF. This challenge is common in unsupervised learning, as the original SmooSeg also requires tuning multiple parameters, among them two in the smoothness term that we adopted. Luckily, γ seems to exhibit a convex search space and performs well across a broad range of values. Another general consideration in state-of-the-art unsupervised learning approaches, particularly those based on DINO, is the low-resolution input. This results in a limited receptive field, posing challenges for detecting large objects like buildings that have to be divided across multiple, separate inputs. We partially address this issue by downscaling the images, employing overlap for qualitative analysis, and running CRF across the stitched prediction images. Moreover, DINO is designed for image-level classification, leading to even smaller output feature maps due to the absence of a decoder, which necessitates the use of CRFs or similar methods for refinement. Additionally, different datasets, number of clusters K, and different feature extractors like DINOv2 (Oquab et al., 2024) tend to yield varying semantic clusters, which is inherent to unsupervised methods. This inconsistency highlights the need for further research into retaining semantic groups across datasets, transfer learning, domain adaptation techniques, practical (real-time) applicability, and scalability of these models.

Ultimately, we present a novel approach for enriching networks built on frozen foundation models, which are typically limited to RGB inputs, with additional modalities. Our fully unsupervised method requires neither labeled data nor handcrafted rules, yet delivers notable performance improvements. As we could see on the current dataset, self-supervised approaches, and our method in particular, readily support applications such as land-sealing analysis and building-density estimation.

References

Albert, L., Rottensteiner, F., Heipke, C., 2017. A higher order conditional random field model for simultaneous classification of land cover and land use. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130, 63–80.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 20-32. Geospatial Computer Vision.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11), 1222-1239.

Bulatov, D., Burkard, E., Ilehag, R., Kottler, B., Helmholz, P., 2020. From multi-sensor aerial data to thermal and infrared simulation of semantic 3D models: Towards identification of urban heat islands. *Infrared Physics & Technology*, 105, 103233.

Bulatov, D., Häufel, G., Lucks, L., Pohl, M., 2019. Land cover classification in combined elevation and optical images supported by OSM data, mixed-level features, and non-local optimization algorithms. *Photogrammetric Engineering & Remote Sensing*, 85(3), 179–195.

Bulatov, D., Häufel, G., Meidow, J., Pohl, M., Solbrig, P., Wernerus, P., 2014. Context-based automatic reconstruction and texturing of 3D urban terrain for quick-response tasks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 157–170.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*, PMLR, 1597–1607.

Cho, J. H., Mall, U., Bala, K., Hariharan, B., 2021. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16794–16804.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

González-Santiago, J., Schenkel, F., Gross, W., Middelmann, W., 2022. Deep self-supervised pixel-level learning for hyperspectral classification. *2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–5.

González-Santiago, J., Schenkel, F., Gross, W., Middelmann, W., 2023. Deep self-supervised hyperspectral-lidar fusion for land cover classification. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 5910–5913.

Hahn, O., Araslanov, N., Schaub-Meyer, S., Roth, S., 2024. Boosting Unsupervised Semantic Segmentation with Principal Mask Proposals. *Transactions on Machine Learning Research (TMLR)*.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W. T., 2022. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*.

Harb, R., Knobelreiter, P., 2021. Infoseg: Unsupervised semantic image segmentation with mutual information maximization. *German Conference on Pattern Recognition*.

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

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2020. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5), 4340–4354.

Kim, C., Han, W., Ju, D., Hwang, S. J., 2024. Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Koppanyi, Z., Iwaszczuk, D., Zha, B., Saul, C. J., Toth, C. K., Yilmaz, A., 2019. Chapter 3 - multimodal semantic segmentation: Fusion of rgb and depth data in convolutional neural networks. M. Y. Yang, B. Rosenhahn, V. Murino (eds), *Multimodal Scene Understanding*, Academic Press, 41–64.

König, D., 1916. Über Graphen und ihre Anwendung auf Determinantentheorie und Mengenlehre. *Mathematische Annalen*, 77, 453-465. http://eudml.org/doc/158740.

Lan, M., Wang, X., Ke, Y., Xu, J., Feng, L., Zhang, W., 2024. SmooSeg: smoothness prior for unsupervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic Segmentation of Aerial Images with an Ensemble of CNSs. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-3, 473–480.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, 1–31.

Qiu, K., Budde, L. E., Bulatov, D., Iwaszczuk, D., 2022a. Exploring fusion techniques in u-net and deeplab v3 architectures for multi-modal land cover classification. *Earth Resources and Environmental Remote Sensing/GIS Applications XIII*, 12268, SPIE, 190–200.

Qiu, K., Bulatov, D., Lucks, L., 2022b. Improving car detection from aerial footage with elevation information and markov random fields. *International Conference on Signal Processing and Multimedia Applications, SIGMAP*, 112–119.

Qiu, K., Wagenbach, L., Bulatov, D., Iwaszczuk, D., 2025. Improving self-supervised segmentation of urban scenes using ndsm-enhanced crf. *Proc.SPIE*, 13263, 132630L.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, Springer.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 256–271.

Seong, H. S., Moon, W., Lee, S., Heo, J.-P., 2023. Leveraging Hidden Positives for Unsupervised Semantic Segmentation.

Song, A., Kim, Y., 2020. Semantic segmentation of remote-sensing imagery using heterogeneous big data: International society for photogrammetry and remote sensing Potsdam and Cityscape datasets. *ISPRS International Journal of Geo-Information*, 9(10), 601.

Tuia, D., Camps-Valls, G., 2011. Urban Image Classification With Semisupervised Multiscale Cluster Kernels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(1), 65-74.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., Zhu, X. X., 2022. Self-Supervised Learning in Remote Sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4), 213-247.

Wegner, J. D., Montoya-Zegarra, J. A., Schindler, K., 2013. A higher-order crf model for road network extraction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1698–1705.

Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169, 114417.