Prototypical Contrastive Network for Imbalanced Aerial Image Segmentation

Keiller Nogueira[1], Mayara Maezano Faita-Pinheiro[2],
Ana Paula Marques Ramos[3], Wesley Nunes Gonçalves[4], José Marcato Junior[4], Jefersson A. dos Santos[5]

[1] University of Stirling, Stirling, FK9 4LA, Scotland, UK
[2] University of Western São Paulo (UNOESTE), Presidente Prudente, São Paulo, Brazil
[3] São Paulo State University (UNESP), Presidente Prudente, São Paulo, Brazil
[4] Federal University of Mato Grosso do Sul (UFMS), Campo Grande, Mato Grosso do Sul, Brazil
[5] University of Sheffield, Sheffield, S10 2TN, England, UK

keiller.nogueira@stir.ac.uk, mayarafaita@gmail.com, marques.ramos@unesp.br
jose.marcato@ufms.br, wesley.goncalves@ufms.br, j.santos@sheffield.ac.uk

Abstract

Binary segmentation is the main task underpinning several remote sensing applications, which are particularly interested in identifying and monitoring a specific category/object. Although extremely important, such a task has several challenges, including huge intra-class variance for the background and data imbalance. Furthermore, most works tackling this task partially or completely ignore one or both of these challenges and their developments. In this paper, we propose a novel method to perform imbalanced binary segmentation of remote sensing images based on deep networks, prototypes, and contrastive loss. The proposed approach allows the model to focus on learning the foreground class while alleviating the class imbalance problem by allowing it to concentrate on the most difficult background examples. The results demonstrate that the proposed method outperforms state-of-the-art techniques for imbalanced binary segmentation of remote sensing images while taking much less training time.

1. Introduction

In recent years, the rapid development of innovative sensor technologies has opened new opportunities to the remote sensing community, allowing a better understanding of the Earth's surface [6]. Towards this, various applications [8, 38, 52] are interested in identifying a specific category/object (such as cars, water surface, etc) in the images in order to better study and monitor particular events related to those. Such applications usually model this problem as a binary segmentation task in which the main goal is to classify each pixel of an image into one of two semantic categories, usually referenced as foreground/positive and background/negative [3, 12].

Despite being extremely important, such a task has several challenges. One is due to the fact that the background class is not clearly defined, being actually composed of samples of distinct semantic properties/categories, such as forests, cities, roads, etc, thus having a huge variance. Moreover, this also leads to another important challenge: class imbalance [23]. This is because the negative class usually has a lot (thousands, millions, or even billions) more samples than the foreground class, thus increasing the bias and, consequently, making learning difficult.

Most works proposed to tackle imbalanced remote sensing segmentation [17, 32, 35] are based on learning good representations for both classes, thus allowing the models to distinguish between them (using, for example, some learned decision boundary). The main problem is that this modeling makes learning extremely difficult, as techniques would need to capture most patterns of the negative class in order to distinguish it from the positive one, a complex process given the high diversity of this class. Furthermore, most approaches do not take into account that only the hard background samples are valuable for optimization and that, because of the visual dissimilarity, these are generally much less than the easy ones.

Motivated by this, in this paper, we propose a novel method to perform imbalanced binary segmentation of remote sensing images based on deep networks, prototypes [44], and contrastive loss [16, 46]. Specifically, the deep network extracts features that are used to learn prototypes [44] **only** for the positive class (instead of having prototypes for each label), allowing the model to focus on learning a good representation for this class. Such a model

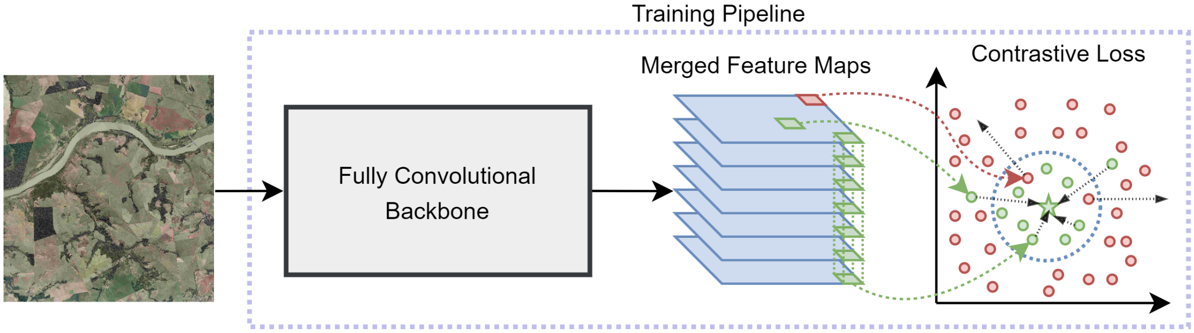


Figure 1. Training pipeline of the proposed approach. A fully convolutional backbone learns the patterns and extracts features for all pixels of the input image. Foreground samples (in green) are then used to learn its prototype (green star) while being pulled close together to it by a novel contrastive loss. Meanwhile, hard negative instances close to the prototype are pushed apart by this proposed cost function. Observe that, although only one positive prototype is represented in this image (for simplification purposes), the network can be trained using multiple positive prototypes depending on the problem.

is trained using a novel contrastive loss, which aims to pull positive samples closer to the prototypes whereas pushing background data apart. This new paradigm (Figure 1) allows, during training, the model to focus on the most difficult negative samples defined based on their distance to the prototypes, thus alleviating the impact of the class imbalance issue. In practice, we can summarize the main contributions of this paper as follows: (i) a novel approach that performs imbalanced binary semantic segmentation by focusing on the positive class, thus easing the network learning process, (ii) an intrinsic system that alleviates the imbalanced data issue, really common in those scenarios, during the learning process, and (iii) a new high-resolution remote sensing dataset for river segmentation that takes the unbalanced data issue to an extreme scenario.

2. Related Work

We divided the related work into two parts: the first part presents the main works proposed for imbalanced segmentation of remote sensing images, while the second part presents the methods most similar to the proposed one, i.e., that combine prototypical and contrastive learning.

2.1. Imbalanced Segmentation

Some authors address the class imbalance issue by employing specific cost functions, such as weighted cross-entropy [20, 40], Focal loss [5, 10], Dice loss [54], or by combining distinct losses, including weighted cross-entropy and Lovász [17, 47], Focal and Tversky [14], cross-entropy and Tversky [32], dual cross entropy [29] and the Focal [18], Dice and cross-entropy [2, 51], and so on. In general, the main idea of these methods is to give more weight to hard examples.

Other authors handled this problem by proposing new and specific techniques. Ma *et al.* [35] combined features describing the whole image and the foreground objects using a dual-branch network. To deal with imbalanced data, they adapted the online hard example mining strategy (OHEM) [43] to dynamically select the relevant examples. In [12], the authors employed a novel strategy of combining tiles of multiple different images/classes (similar to existing image mixing data augmentation [36]) that deals with imbalanced data by replicating the minority class more often. Li *et al.* [28] proposed a point-wise propagation module that balances the learning by selecting the most salient background samples. Zheng *et al.* [53] developed a foreground-aware relation network that deals with imbalanced data via a relation-based and optimization-based modeling, that focuses on the hard examples of background during training. In general, such approaches seek to learn good representations for all classes, dealing with the class imbalance issue by proposing some technique to mine hard samples.

In contrast with aforementioned techniques, our approach focuses on learning features only for the foreground class, i.e., the approach does not focus on capturing patterns for the negative data, but on making them as different as possible from the positive ones. Furthermore, the proposed approach deals with the class imbalance issue by using an intrinsic system of selecting the most difficult background examples, a process that alleviates training time.

2.2. Prototypical Contrastive Learning

Prototypical learning [44, 48] seeks to associate each class to a representation, or prototype, and classify the observations according to the nearest prototype. On the other hand, contrastive learning tries to learn a space in which samples of the same class are close to each other, while instances of other classes are far apart.

Although such paradigms have been successfully employed separately for distinct tasks, such as few-shot [9, 39], domain adaptation [15], self-supervised learning [27], and segmentation [13, 22], relatively few works combine prototypical and contrastive learning. Precisely, only Yang and Ma [49] and Liu *et al.* [31] combined prototypical learning with contrastive learning to improve remote sensing image segmentation. Both works propose the use of prototypes for all classes and only use contrastive learning to optimize the embedding space, seeking to enhance intra-class compactness and inter-class separability, but without taking class imbalance into account. Unlike these methods, the proposed technique learns prototypes only for the foreground class while employing contrastive learning to optimize the representation space and deal with imbalanced data.

3. Methodology

In this section, we present the proposed method for imbalanced remote sensing image segmentation and provide all technical details.

3.1. Overview

The pipeline of the proposed method is presented in Figure 1. As introduced, the main idea of the approach is to have prototypes only for the foreground, pulling samples of this class closer to the prototypes while pushing background data apart through contrastive learning. Overall, this process alleviates the training, as the model does not need to focus on capturing patterns for the background class, which has a huge intra-class variance, but on making samples of this class different as possible from the positive one. Furthermore, in order to deal with imbalanced data, the proposed model is able to select the most difficult background samples just using their distance to the foreground prototypes, without having to resort to any other mining technique that would impact the training.

3.2. Training

During the training, an image $\mathcal{X}$ is used as input for a deep learning-based backbone, responsible for learning a function f_θ (with learnable parameters θ) that maps each input pixel i of $\mathcal{X}$ into a $\mathcal{N}$-dimensional representation $\mathcal{Z} \in \mathbb{R}^{\mathcal{N}}$. The ℓ_2-normalized feature representations $\|\mathcal{Z}\|_2$ of the foreground instances are then pulled closer to the nearest learnable prototype $\mathcal{P}^*$, whereas background data representations are pushed apart from all foreground prototypes at the same time, a process accomplished by optimizing the network with this novel contrastive loss function:

$$\mathcal{L} = \sum_i Y_i\, \mathcal{D}(\|f_\theta(\mathcal{X}_i)\|_2,\ \mathcal{P}^*)^2 + \sum_j^{\mathcal{C}} (1 - Y_i)\ \{max(0, m - \mathcal{D}(\|f_\theta(\mathcal{X}_i)\|_2,\ \mathcal{P}_j))\}^2 \tag{1}$$

where Y is the label (i.e., 0 for pixels of the negative class and 1 for those of the positive label), $\mathcal{C}$ is the number of positive prototypes, $\mathcal{P}_j \in \mathbb{R}^{\mathcal{N}}\ \forall\, j \in \mathcal{C}$ is a prototype, $m > 0$ is a margin, and $\mathcal{D}$ is a function that measures the distances between the normalized pixel representation $\|\mathcal{Z}\|_2$ and a prototype $\mathcal{P}_j$.

Observe that: (i) although standard ℓ_2 regularization is used with the proposed loss, it has been omitted from Equation 1 for simplification and readability purposes, and (ii) following the insights of [26, 50], the embedding function f_θ and the prototypes $\mathcal{P}$ are learned simultaneously. This differs from most other works [24, 44] in the literature that learns prototypes separately or defines them as centroids of the learned representations.

3.3. Testing

During inference, the input data is processed using the trained backbone and all sample representations $\|\mathcal{Z}_i\|_2$ are then projected onto the learned space (with the prototypes). Then, examples with a distance to any prototype smaller than the margin ($\mathcal{D}(\|\mathcal{Z}_i\|_2,\ \mathcal{P}_j) < m$) are classified as foreground whereas samples with a distance to all prototypes greater than the margin are considered as background.

3.4. Dealing with Imbalanced Data

As introduced, a common challenge related to binary segmentation is class imbalance. This is mainly due to the fact that the negative class aggregates samples from many different semantic categories, thus generally having far more samples than the positive class, which is composed of instances of just one (relevant) category. Such highly imbalanced scenarios make most machine learning models biased towards the majority class that, in severe cases, may completely ignore the minority category [23]. Moreover, highly imbalanced data is considered a particularly relevant issue for deep neural networks, because it directly impacts the gradients causing the model to get stuck in a slow convergence mode [4, 23].

However, although composed of a myriad of samples, only the hard part of the background examples is valuable for optimization. Moreover, given the clear difference in visual properties between most negative samples and the positive class, the hard background examples are usually much less than the easy ones. Seeking to take advantage of this, the proposed method intrinsically deals with the class imbalance issue by optimizing the model using only the most relevant (or hard/similar) negative samples. In practical terms, this is performed by training the model employing only the background instances with a distance to a prototype smaller than the margin ($\mathcal{D}(f_\theta(\mathcal{X}_i),\ \mathcal{P}_j) < m$, where $\mathcal{X}_i$ belongs to the negative category), as can be seen in Equation 1. This intrinsic system allows the proposed model to deal with imbalanced data without having to resort to any

Dataset	Sets	#Pixels		%	
		Foreground	Background	Foreground	Background
Vaihingen	Training	491,636	44,742,258	1.09	98.91
	Validation	174,982	9,475,136	1.81	98.19
	Testing	279,069	22,924,725	1.20	98.80
River	Training	358,203	190,787,097	0.19	99.81
	Validation	247,653	191,055,145	0.13	99.87
	Testing	175,552	191,288,546	0.10	99.90

Table 1. Number of pixels per class for the tested datasets.

other mining technique, further easing the training process.

Although this process may help alleviate this relevant problem, class imbalance issues can still persist depending on the dataset. Even though we did not observe that in the performed experiments (probably due to the aforementioned dissimilarity between most negative instances and the positive class), the proposed technique can be easily adapted to further tackle the class imbalance problem by selecting a specific amount of background samples based on their distance to the prototypes (or randomly), thus better balancing the classes.

4. Experiments

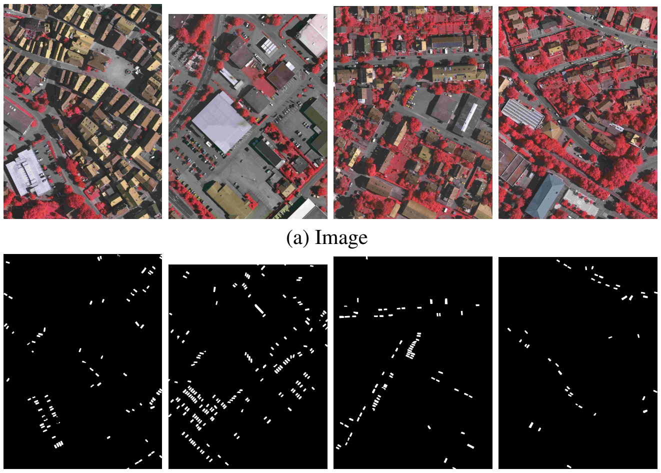
We evaluate the efficiency of our method in this section by carrying out a systematic evaluation using two datasets.

4.1. Datasets

To better evaluate the effectiveness of the proposed method, we carried out experiments using two high-resolution RGB remote sensing datasets with distinct properties: Vaihingen [1] and River. While the former dataset is considered one of the most important in the remote sensing domain, the latter is being proposed in this work and takes the imbalance issue to a new level. The pixel distribution of both datasets can be seen in Table 1.

Vaihingen. This publicly available dataset [1] is composed of a total of 16 image tiles (with an average size of 2494×2064 pixels), that are densely classified into six possible labels: impervious surfaces, buildings, low vegetation, tree, car, clutter/background. Each image is composed of near-infrared, red, and green channels (in this order) and has a spatial resolution of 0.9 meter. In order to simulate an imbalanced binary scenario with this dataset, we considered the rarest class, i.e., car, as the foreground and all remaining others as background, as presented in Figure 2.

For this dataset, we followed the protocol proposed by [37]. Precisely, 9 out of the 16 images were used to train the proposed model; 2 images (IDs 5 and 7) were used for validation; and the 5 remaining images (IDs 11, 15, 28, 30, 34) were employed for testing. In general, this protocol presents a scenario of considerable imbalance given that for each car sample there are 91 non-car pixels.


(a) Image

(b) Ground-Truth

Figure 2. Some images of the Vaihingen dataset [1] and their respective ground-truths for the car class.

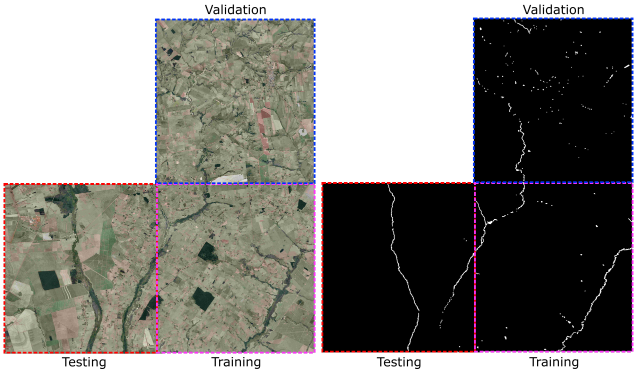


Figure 3. Images of the River dataset and their respective ground-truths. White areas represent water, while black regions are non-water. Note that the images of this dataset actually compose an orthomosaic that covers a large area in Brazil, including three rivers.

River. This dataset, proposed in this work, is composed of three high-resolution RGB orthoimages between the states of São Paulo and Paraná, Brazil. The images, manually annotated by experts into water and non-water classes, have an average size of $13,358 \times 14,322$ pixels and 1 meter of spatial resolution, thus totalling more than 500 millions pixels and more than 500 square kilometers.

For this dataset, each of the images is used separately for training, validation, and testing, as can be seen in Figure 3. This protocol takes the imbalance issue to an extreme scenario wherein for each water pixel there are approximately 631 pixels of non-water.

4.2. Implementation Details

The proposed technique employed a fully convolutional [34] DenseNet-121 [19] as backbone, which has been pre-trained on the ImageNet dataset. It is important to highlight that, in order to aggregate more contextual information and enhance discriminative power, feature representations from all dense blocks are extracted, upsampled (if necessary), and merged, thus fusing low, mid, and high semantic-level data. Finally, we employed the squared Euclidean distance as $\mathcal{D}$ for the proposed loss function (Equation 1) based on an analysis performed by Snell *et al.* [44], who concluded that such a distance works better for prototype learning because it is a regular Bregman divergence.

Additionally, the proposed method[1] was implemented using PyTorch. During training, the proposed approach used the following hyper-parameters: patch size of 128 × 128 pixels, 100 epochs, weight decay of 0.005, batch size equal to 32, Adam [25] as optimizer, learning rate of 0.01, and step decay of 0.1 every 25 epochs. Finally, all experiments were performed on a machine with an Intel i7 4960X with 3.6GHz of clock, 64GB of RAM memory, and Ubuntu operating system version 18.04.3 LTS. Four GeForce GTX Titan X with 12GB of memory, under an 11.4 CUDA version, were employed in this work. Note, however, that each GPU was used independently and that all models employed in this work can be trained using only one GPU.

4.3. Baseline Methods

For both dataset, we compare our method with various baseline models proposed for imbalanced binary segmentation of remote sensing images, including: (i) DeepWaterMap [20], which addresses class imbalance by employing a weighted cross entropy loss function, (ii) BASNet [5], that handles unbalanced scenarios by using the focal loss [30], (iii) U-Net++ [17, 47], that tries to handle unbalanced data by combining the weighted cross-entropy and Lovász cost functions [21], (iv) DUPnet [32], which tackles the class imbalance issue by using a combination of the cross-entropy and Tversky [41] losses, (v) DFL [18, 55], that uses a combination of dual cross entropy [29] and the Focal loss, (vi) UFL [42, 51], which combines the Dice and cross-entropy losses to address class imbalance. It is important to highlight that all baselines used the same backbone employed in the proposed approach, i.e., a fully convolutional [34] DenseNet-121 [19].

4.4. Evaluation Metrics

Three different metrics, Precision, Recall, and F1 score (defined in Equation 2), have been selected based on other related works [32, 47] and used to assess the performance of the proposed algorithm and baselines. Specifically, Precision measures the fraction of correct positive outcomes (i.e., True Positive – TP) out of the total positive predictions, being a metric more related to the False Positives (FP). On the other hand, Recall measures the proportion of correct positive predictions (TP) out of the total actual positive instances in the dataset, a measure more related to the False Negatives (FN). Finally, F1 score, is defined as the harmonic mean of precision and recall.

$$
\begin{gathered}
Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \\
F1 - Score = \frac{2 \times Precision \cdot Recall}{Precision + Recall}
\end{gathered} \tag{2}
$$

4.5. Results and Discussion

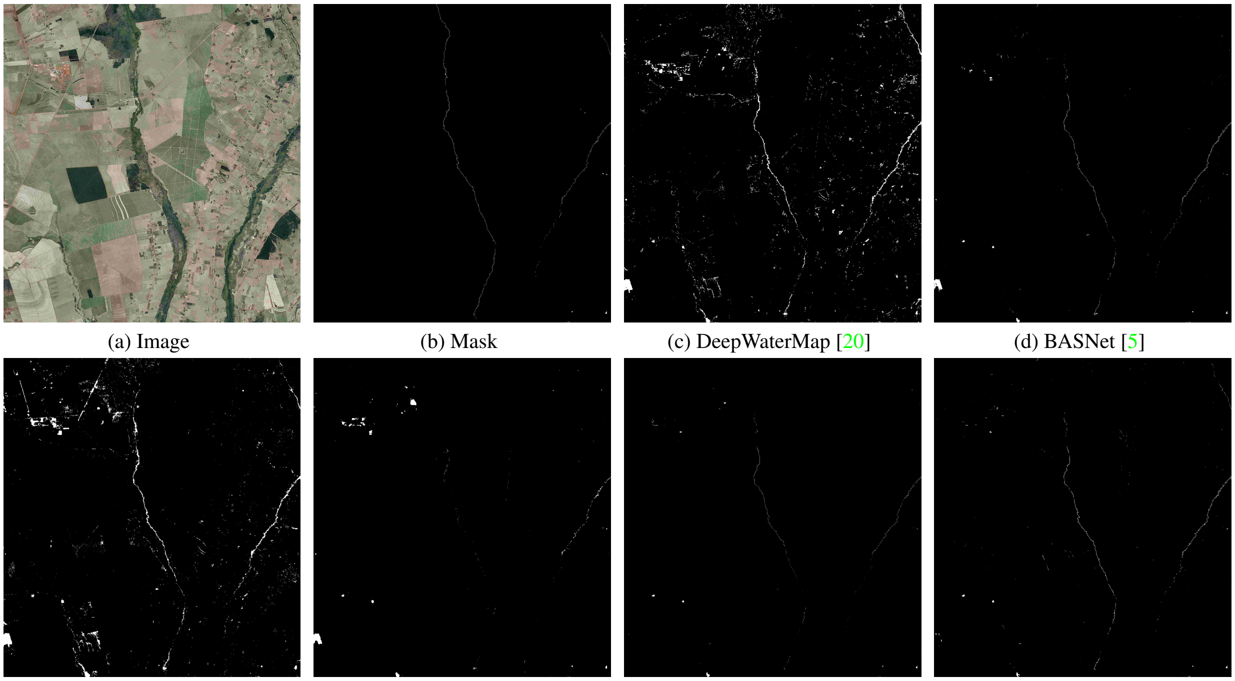
4.5.1 State-of-the-Art Comparison

In Table 2, we report the overall results (on the test set) and training time of the proposed approach and all employed baselines. Compared to the state-of-the-art methods, the proposed approach achieves improvements, in terms of F1-score, on all experimented datasets. Precisely, for the Vaihingen dataset, the proposed method obtained 79.07% of F1-score, with a 4.44% relative gain. As for the River dataset, our approach yielded 78.78% of F1-score, a relative gain of 5.43%.

Despite achieving the best results in terms of F1-score, the proposed approach yielded the second-best result in terms of both Precision and Recall for the two tested datasets, being outperformed by other techniques. This is due to the fact that such methods focus explicitly on reducing false positives and/or negatives by giving more weight to such related samples (thus generating more impact on the aforementioned metrics), while the proposed technique does not optimize the model based on specific sample weights, dealing with false positives and negatives equally, thus better balancing the learning (and such metrics). As a result, the proposed technique is capable of producing outcomes with fewer false positives and false negatives, being more consistent and generating better results in terms of F1-score. Such an analysis can be better visualized with the qualitative results presented in Figures 4 and 5.

In addition to the results, it is important to highlight that our method is very computationally efficient as it takes less than half the training time when compared to all other competing approaches (potentially, because of the mining process that reduces the number of training samples). Overall, the obtained results show that the proposed technique can effectively focus on learning the positive class, showing better generalization in representation learning and semantic understanding, while efficiently dealing with unbalanced datasets, taking much less training time.

[1] The code and the proposed River dataset have been made publicly available at https://github.com/keillernogueira/proto_contrastive_net_imbalanced.



(a) Image (b) Mask (c) DeepWaterMap [20] (d) BASNet [5]

(e) U-Net++ [17,47] (f) DUPnet [32] (g) DFL [18,55] (h) Ours

Figure 4. The River test image, its respective ground-truth, and the prediction maps generated by the proposed algorithm, as well as the best baselines. The white areas represent water, while the black regions are non-water.

Dataset	Method	Training Time (h)	Precision	Recall	F1-Score
Vaihingen	**DeepWaterMap** [20]	20.6	77.95	67.44	70.38
	BASNet [5]	26.6	68.44	79.96	71.52
	U-Net++ [17,47]	20.8	76.00	68.40	69.98
	DUPnet [32]	35.0	68.16	60.34	60.35
	DFL [18,55]	22.3	**93.30**	68.23	74.63
	UFL [42,51]	33.3	69.36	**82.70**	73.04
	Ours	8.3	78.78	81.04	**79.07**
River	**DeepWaterMap** [20]	125.2	52.51	90.05	54.39
	BASNet [5]	139.1	63.76	79.66	68.78
	U-Net++ [17,47]	127.0	52.89	**92.12**	55.10
	DUPnet [32]	145.4	54.60	62.01	56.64
	DFL [18,55]	124.0	**76.81**	70.69	73.35
	UFL [42,51]	143.1	52.41	55.19	54.28
	Ours	66.6	69.61	90.74	**78.78**

Table 2. Obtained results achieved on the test set by the proposed method and baselines for both unbalanced datasets.

4.5.2 Number of Prototypes

The number of prototypes for the foreground class is a critical parameter of the proposed algorithm. Motivated by this, in this Section, we assess the proposed model varying the number of prototypes in order to define the most suitable value for each dataset.

Precisely, we vary the number of foreground prototypes from 1 to 3, evaluating all possibilities using both datasets. For a fair comparison, we preserve all other parameters as described in Section 4.2 and use margin m equal to 3 for all experiments. The experimental results (on the validation set) are shown in Table 3. From this table, we can observe that the number of prototypes has an influential effect on the final outcome for the two tested datasets. Overall, we can conclude that using only one positive prototype produces the best results for both datasets. Such an outcome corroborates with other works in the literature which concluded that using more than one prototype does not bring effective gains, just increasing the processing time [7,26,44]. Finally, it is important to highlight that the outcomes of this analysis have been used for all other experiments in this work.

4.5.3 Margin Analysis

The margin hyperparameter m, introduced in Equation 1, plays a crucial role in the optimization of the proposed method, given that if it is set too low, the model might not be able to learn a meaningful space as classes might have similar representations, whereas if it is set too high, the model might become too conservative, and might not be able to generalize to unseen data. Due to this, in this Section, we investigate the effects of different margin values for the cost function to define the most appropriate for each dataset.

Specifically, we vary the value of the margin from 1 to

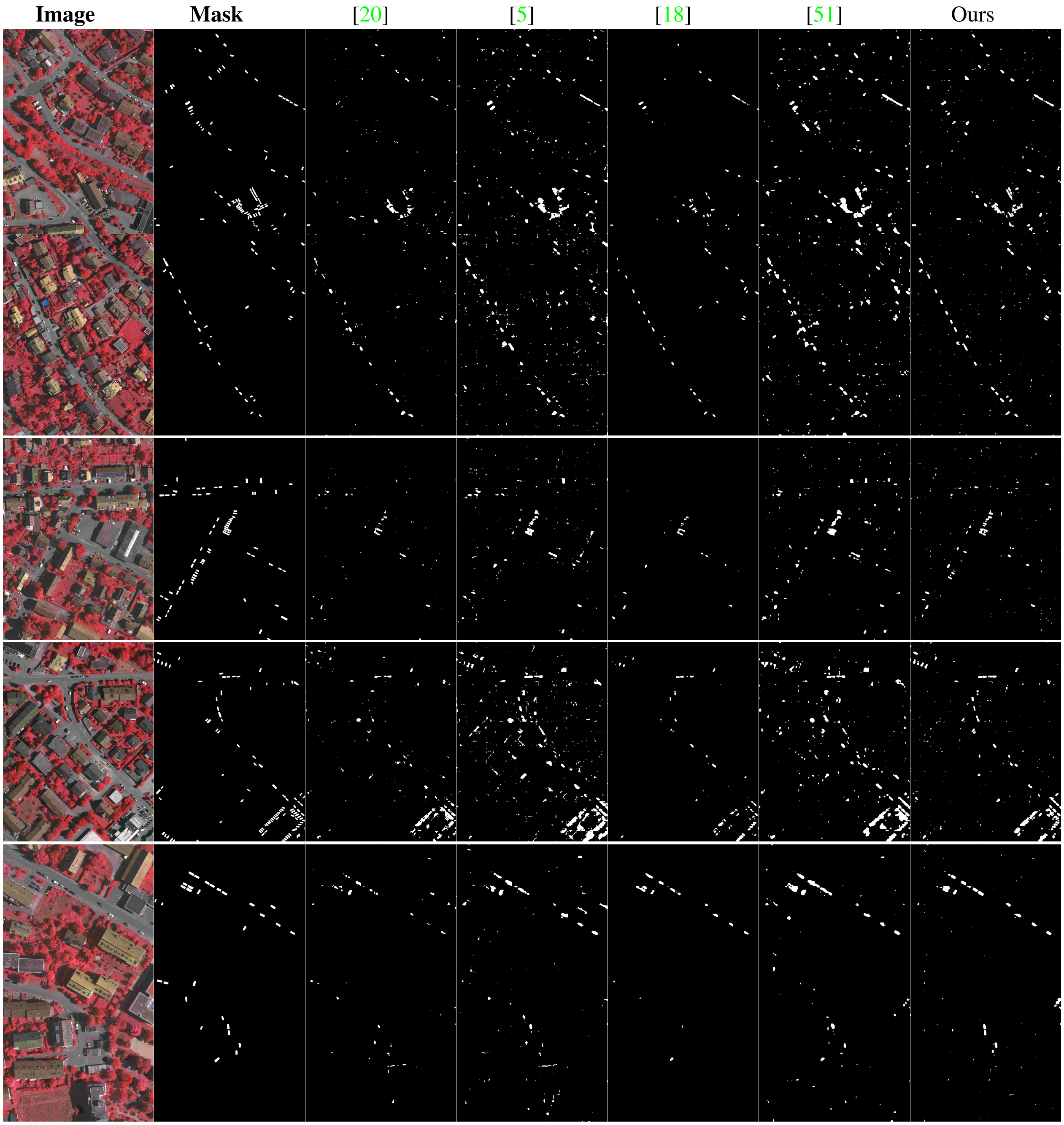


Figure 5. Predictions for the test set of the Vaihingen dataset. White areas represent the car class, while black regions are the background.

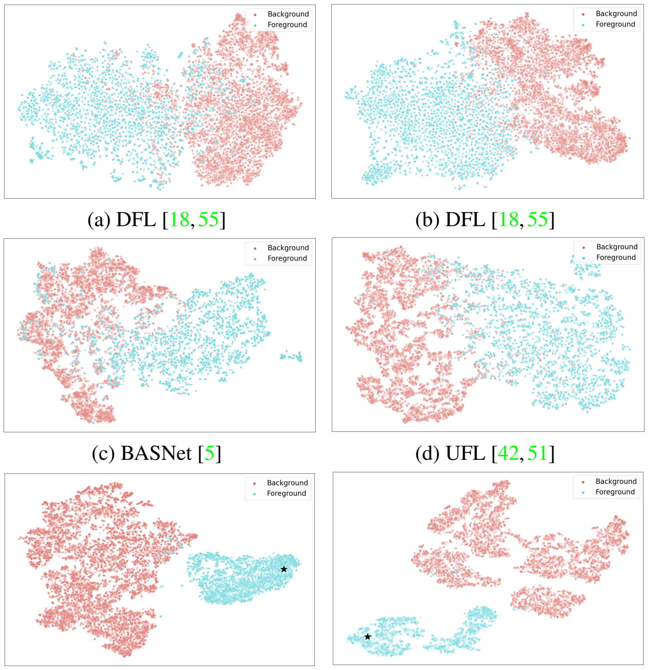
5, evaluating all possibilities using both datasets. Again, for a fair comparison, we preserve all other parameters as described in Section 4.2 and use only one prototype for the foreground class (based on the outcomes of the previous Section). The experimental results are shown in Table 4. From this table, we can observe that the margin value drastically affects the final outcome for both datasets. In any case, we can conclude that, for the Vaihingen dataset, the best value for the margin parameter is 4, whereas for the River dataset, the optimal margin is 3. Again, it is important to emphasize that the outcomes of this analysis have been used for all other experiments in this work.

Dataset	Number of Prototypes	Precision	Recall	F1-Score
Vaihingen	**1**	**75.95**	**80.92**	**77.41**
	2	70.43	79.21	73.35
	3	64.76	83.17	69.70
River	**1**	**68.19**	**88.08**	**76.56**
	2	55.71	74.05	63.61
	3	55.96	60.17	57.29

Table 3. Results (on the validation set) of the proposed method trained using different numbers of prototypes for the foreground class.

Dataset	Margin	Precision	Recall	F1-Score
Vaihingen	**1**	73.04	79.30	76.06
	2	74.57	**82.26**	77.07
	3	75.95	80.92	77.41
	4	**77.21**	80.59	**78.22**
	5	75.78	75.76	74.93
River	**1**	50.27	85.56	46.75
	2	54.47	85.90	65.67
	3	**68.19**	**88.08**	**76.56**
	4	66.38	80.83	71.38
	5	63.80	81.15	71.44

Table 4. Results (on the validation set) of the proposed method trained using distinct values for the margin hyperparameter m of the proposed loss (Equation 1).



(a) DFL [18, 55]

(b) DFL [18, 55]

(c) BASNet [5]

(d) UFL [42, 51]

(e) Ours

(f) Ours

Figure 6. Visualizations of the embedding spaces learned by the proposed method and some baselines for both datasets. First column is for the Vaihingen dataset whereas the second one is for the proposed River dataset. The black star represents the learned prototype for the foreground class.

4.5.4 Embedding Space

To better analyze the optimization process of the proposed method, we include visualizations of the embedding space generated by the best models for both datasets. Furthermore, to allow for a comparison, we also include visualizations of the spaces generated by some of the best-performing baselines. To create such visualizations, we randomly selected approximately 6,000 samples of each class from the validation set of each tested dataset. Then we projected the related representations into a 2-dimensional space using t-SNE [45], as shown in Figure 6.

Overall, we can observe that the proposed technique is able to efficiently group samples of the foreground class around the prototype, presenting better compactness for such a class when compared to the baselines, while being able to effectively push away instances of the negative class.

5. Conclusions

In this paper, we propose a novel approach for extremely imbalanced binary segmentation of remote sensing images. The proposed method, trained using a new contrastive loss, is able to effectively learn a good representation for the foreground class while efficiently dealing with imbalanced data, showing better generalization in representation learning and semantic understanding. Experiments were conducted using two high-resolution remote sensing datasets with very distinct properties. Results demonstrate the effectiveness and computational efficiency of the proposed method which outperforms several state-of-the-art techniques while taking much less training time. In the future, we plan to better analyze more up-to-date backbones, such as ViT [11] or Swin Transformers [33], and different reduction strategies for the loss (for example, mean instead of the sum). We also plan to investigate the effectiveness of the proposed approach for different datasets and applications.

Acknowledgments

This research was partially funded by CAPES (88887.901123/2023-00, 88881.311850/2018-01), FAPEMIG (APQ-00449-17), CNPq (310517/2020-6, 306955/2021-0, 405997/2021-3, 308481/2022-4, 305296/2022-1), Fundect (p. 71/009.436/2022), and the Serrapilheira Institute (grant R-2011-37776).

References

[1] International society for photogrammetry and remote sensing (isprs). https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx. Accessed: 2023-05-09. 4

[2] Abolfazl Abdollahi, Biswajeet Pradhan, and Abdullah Alamri. Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access*, 8:179424–179436, 2020. 2

[3] Shubhra Aich, William van der Kamp, and Ian Stavness. Semantic binary segmentation using convolutional networks without decoders. In *IEEE/CVF Computer Vision and Pattern Recognition Workshop*, pages 197–201, 2018. 1

[4] Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4(6):962–969, 1993. 3

[5] Yanbing Bai, Wenqi Wu, Zhengxin Yang, Jinze Yu, Bo Zhao, Xing Liu, Hanfang Yang, Erick Mas, and Shunichi Koshimura. Enhancement of detecting permanent water and temporary water in flood disasters by fusing sentinel-1 and sentinel-2 imagery using deep learning algorithms: Demonstration of sen1floods11 benchmark datasets. *Remote Sensing*, 13(11):2220, 2021. 2, 5, 6, 7, 8

[6] Yifang Ban, Peng Gong, and Chandra Giri. Global land cover mapping using earth observation satellite data: Recent progresses and challenges, 2015. 1

[7] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 15333–15342, 2021. 6

[8] Ila Chawla, L Karthikeyan, and Ashok K Mishra. A review of remote sensing applications for water security: Quantity, quality, and extremes. *Journal of Hydrology*, 585:124826, 2020. 1

[9] Gong Cheng, Liming Cai, Chunbo Lang, Xiwen Yao, Jinyong Chen, Lei Guo, and Junwei Han. Spnet: Siamese-prototype network for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 2

[10] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *IEEE/CVF Computer Vision and Pattern Recognition Workshop*, 2018. 2

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020. 8

[12] Kun Fu, Wanxuan Lu, Wenhui Diao, Menglong Yan, Hao Sun, Yi Zhang, and Xian Sun. Wsf-net: Weakly supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sensing*, 10(12):1970, 2018. 1, 2

[13] Pedro Henrique Targino Gama, Hugo Neves Oliveira, Jose Marcato, and Jefersson Dos Santos. Weakly supervised few-shot segmentation via meta-learning. *IEEE Transactions on Multimedia*, 2022. 3

[14] Chengling Gao, Hailiang Ye, Feilong Cao, Chenglin Wen, Qinghua Zhang, and Feng Zhang. Multiscale fused network with additive channel–spatial attention for image segmentation. *Knowledge-Based Systems*, 214:106754, 2021. 2

[15] Kuiliang Gao, Anzhu Yu, Xiong You, Chunping Qiu, and Bing Liu. Prototype and context enhanced learning for unsupervised domain adaptation semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE/CVF Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006. 1

[17] Max Helleis, Marc Wieland, Christian Krullikowski, Sandro Martinis, and Simon Plank. Sentinel-1-based water and flood mapping: benchmarking convolutional neural networks against an operational rule-based processing chain. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2023–2036, 2022. 1, 2, 5, 6

[18] Md Sazzad Hossain, John M Betts, and Andrew P Paplinski. Dual focal loss to address class imbalance in semantic segmentation. *Neurocomputing*, 462:69–87, 2021. 2, 5, 6, 7, 8

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 4, 5

[20] Furkan Isikdogan, Alan C Bovik, and Paola Passalacqua. Surface water mapping by deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(11):4909–4918, 2017. 2, 5, 6, 7

[21] Shruti Jadon. A survey of loss functions for semantic segmentation. In *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2020. 5

[22] Xufeng Jiang, Nan Zhou, and Xiang Li. Few-shot segmentation of remote sensing images using deep metric learning. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 3

[23] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019. 1, 3

[24] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *Advances in Neural Information Processing Systems*, 34:1192–1203, 2021. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Loic Landrieu and Vivien Sainte Fare Garnot. Leveraging class hierarchies with metric-guided prototype learning. In *British Machine Vision Conference*, 2021. 3, 6

[27] Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, and Chao Tao. Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 3

[28] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 4217–4226, 2021. 2

[29] Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao. Dual cross-entropy loss for small-sample fine-grained vehicle classification. *IEEE Transactions on Vehicular Technology*, 68(5):4204–4212, 2019. 2, 5

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 5

[31] Quanyong Liu, Jiangtao Peng, Yujie Ning, Na Chen, Weiwei Sun, Qian Du, and Yicong Zhou. Refined prototypical contrastive learning for few-shot hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 3

[32] Zhiheng Liu, Xuemei Chen, Suiping Zhou, Hang Yu, Jianhua Guo, and Yanming Liu. Dupnet: Water body segmentation with dense block and multi-scale spatial pyramid pooling for remote sensing images. *Remote Sensing*, 14(21):5567, 2022. 1, 2, 5, 6

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 8

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4, 5

[35] Ailong Ma, Junjue Wang, Yanfei Zhong, and Zhuo Zheng. Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 1, 2

[36] Humza Naveed. Survey: Image mixing and deleting for data augmentation. *arXiv preprint arXiv:2106.07085*, 2021. 2

[37] Keiller Nogueira, Mauro Dalla Mura, Jocelyn Chanussot, William Robson Schwartz, and Jefersson Alex Dos Santos. Dynamic multicontext segmentation of remote sensing images based on convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):7503–7520, 2019. 4

[38] Keiller Nogueira, Samuel G Fadel, Ícaro C Dourado, Rafael de O Werneck, Javier AV Muñoz, Otávio AB Penatti, Rodrigo T Calumby, Lin Tzy Li, Jefersson A dos Santos, and Ricardo da S Torres. Exploiting convnet diversity for flooding identification. *IEEE Geoscience and Remote Sensing Letters*, 15(9):1446–1450, 2018. 1

[39] Gokul Puthumanaillam and Ujjwal Verma. Texture based prototypical network for few-shot semantic segmentation of forest cover: Generalizing for different geographical regions. *Neurocomputing*, 538:126201, 2023. 2

[40] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 2

[41] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017. 5

[42] Weipeng Shi, Wenhu Qin, Zhonghua Yun, Allshine Chen, Kai Huang, and Tao Zhao. Land cover classification in foggy conditions: Toward robust models. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 5, 6, 8

[43] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 761–769, 2016. 2

[44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2, 3, 5, 6

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[46] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 2495–2504, 2021. 1

[47] Marc Wieland, Sandro Martinis, Ralph Kiefl, and Veronika Gstaiger. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sensing of Environment*, 287:113452, 2023. 2, 5, 6

[48] Ziheng Xia, Penghui Wang, Ganggang Dong, and Hongwei Liu. Spatial location constraint prototype loss for open set recognition. *Computer Vision and Image Understanding*, 229:103651, 2023. 2

[49] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 1809–1818, 2022. 3

[50] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 3474–3482, 2018. 3

[51] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022. 2, 5, 6, 7, 8

[52] Wang Zhang, Chunsheng Liu, Faliang Chang, and Ye Song. Multi-scale and occlusion aware network for vehicle detection and segmentation on uav aerial images. *Remote Sensing*, 12(11):1760, 2020. 1

[53] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery.

In *IEEE/CVF Computer Vision and Pattern Recognition*, pages 4096–4105, 2020. 2

[54] Lichen Zhou, Chuang Zhang, and Ming Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *IEEE/CVF Computer Vision and Pattern Recognition Workshop*, pages 182–186, 2018. 2

[55] Zheng Zhou, Change Zheng, Xiaodong Liu, Ye Tian, Xiaoyi Chen, Xuexue Chen, and Zixun Dong. A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data. *Remote Sensing*, 15(7):1768, 2023. 5, 6, 8