International Journal of Artificial Intelligence,
ISSN 0974-0635; Int. J. Artif. Intell.
2025, Volume 23, Issue Number 1

OTM-UNet: Optimized Semantic Segmentation of Remote Sensing Imagery with Learned Optimal Transport Maps

Abdelaadim Khriss[1] **and Aissa Kerkour Elmiad**[1] **and Mohammed Badaoui**[2]

[1]Lab. LARI, FSO, Mohammed Premier University, Oujda, MOROCCO
Email: abdel.abdelkrs@gmail.com
mid.kerkour@gmail.com

[2]Lab. LaMSD, ESTO, Mohammed Premier University, Oujda, MOROCCO
Email: med.badaoui@gmail.com

ABSTRACT

Deep learning techniques have recently shown remarkable effectiveness in the semantic segmentation of natural and remote sensing (RS) images. However, despite advances in conventional networks, accurate detection and segmentation of small objects in complex scenes remains a major challenge. The detection and segmentation of small features, such as vehicles and pedestrians, is complex due to the occlusion and density of contextual information. In this paper, we propose an enhanced UNet architecture, called Optimal Transport Maps (OTM-UNet), which uses optimal transport layers to compute learned transport maps that align feature maps from both the encoder and decoder. This alignment is critical for preserving spatial orientation and improving semantic consistency during segmentation. Optimized transport layers are strategically placed deep in the decoder and perform exhaustive transformations on feature maps sliced from encoders before concatenation. The resulting transport maps bridge the gap between encoder and decoder feature distributions, facilitating effective information transfer and preserving spatial detail throughout the architecture. The performance of OTM-UNet was evaluated using two publicly available Remote Sensing Imagery datasets, and a comprehensive quantitative and qualitative comparison was made with other models. Results from the evaluation of the Vaihingen dataset showed that the proposed model achieved an impressive average F1 score of 90.90% and an accuracy of 93.17%. In addition, the visual qualitative results showed a significant reduction in object class confusion, improved ability to segment different scales of object, and improved object integrity, highlighting the model's effectiveness in addressing small objects in remote sensing segmentation challenges.

Keywords: UNet, Unmanned aerial vehicle (UAV), Semantic segmentation, Small objects, Optimal transport maps.

Computing Classification System : Computing methodologies - Artificial intelligence - Computer vision - Computer vision problems - Image segmentation.

1 Introduction

Aerial imagery is important in various fields where complete images or landscapes are required, supporting the execution of key applications such as remote sensing data analysis

(Camps-Valls and Bruzzone, 2009). Remote sensing applications cover several domains, from urban planning, where having a global vision of the spaces is highly important for urban development (Navalgund, Jayaraman and Roy, 2007; Triharminto, Adji and Setiawan, 2013), to environmental monitoring (Khriss, Elmiad, Badaoui, Barkaoui and Zarhloule, 2024b; Long, Alexander and Huong, 2021), allowing an in-depth evaluation of ecological situations. It is significant because it provides complex information on a large scale and helps professionals arrive at knowledgeable perspectives. Aerial imagery is indispensable for critical tasks such as topographic mapping, assessing environmental changes, and efficiently and resourcefully managing urban populations, while panoramic tools are limited (Barbarella, Cuomo, Di Benedetto, Fiani and Guida, 2019). Currently, it not only influences the better visualization of people but is also involved in a strategic choice of environmental protection and intended urban development (Shao, Song, Mu, Tian, Chen, He and Kim, 2021). The recent emergence of deep convolutional neural networks has formed the industry standard for image classification, detection, and segmentation tasks significantly impacting a complicated domain like remote sensing (Hemanth and Estrela, 2017). The fact that these networks have been adopted is an interesting trend showing their great efficiency in several disciplines (Kim and Kim, 2017; Khriss, Elmiad, Badaoui, Barkaoui and Zarhloule, 2024a; Huong, Long, Kozlov, Tomin and Sidorov, 2021). This visibly reflects the influence of their wide-ranging ability to master sophisticated functions regarding aerial imaging (Li, Zhang, Xue, Jiang and Shen, 2018; Li, Wang, Zhang, Zhang, Zhao, Xu, Ben and Gao, 2022; Yuan, Shi and Gu, 2021). Despite their suitability for the task, the recognition and segmentation of small objects in large scenes have been a challenging problem (Liu, Sun, Wergeles and Shang, 2021). Small objects with limited spatial extent and complicated details have always been a problem for standard segmentation techniques (Tong, Wu and Zhou, 2020). Because these objects are fixed and loaded with dense representations, they have a limited spatial presence of their own, along with a complexity that requires certain special techniques to bring their importance to the fore.

In this context, our work has contributed to the understanding of the special features of small object segmentation in aerial imagery. Vehicles, pedestrians, and other small details are vital to complete the overall view, such as traffic monitoring disaster response or precision agriculture. Due to the size, occlusion and dense variety of contextual information so many objects are hard to segment accurately. We propose a U-Net architecture version with an enhanced feature alignment mechanism by implementing optimal transport layers. These layers compute learned optimal transport maps which allow aligning feature maps from the encoder as well as the decoder. This alignment is crucial to maintain the spatial orientation and enhance semantic consistency when upsampling. Deep into the decoder, after each transposed convolution operation optimized transport layers are carefully positioned and perform exhaustive transformation of feature maps cropped from encoders before concatenation. The gained optimum transport maps minimize the gap between encoder and decoder feature distributions, thus facilitating the appropriate fusion of combined features before further proceeding with more decoding layers. This approach of relating alignment strategies to features is identified as essential for enhancing a general performance in the network, transmitting effective information transfer, and maintaining appropriate preservation of spatial detail throughout the architecture.

2 Related Works

2.1 Fully Convolutional Networks

Fully Convolutional Networks have made a significant impact on semantic segmentation, which involves classifying each pixel in an image. Several architectures have been developed for FCNs, including those based on ResNet(He, Zhang, Ren and Sun, 2016) and DenseNet(Huang, Liu, Van Der Maaten and Weinberger, 2017). One notable architecture is FCN-8s(Long, Shelhamer and Darrell, 2015). It introduces skip connections from lower levels, combining high-level context with low-level details to improve segmentation quality. The result is a pixel-wise classification that provides a comprehensive segmentation of the image.

2.2 Encoder-decoder architecture

Segmentation tasks are one of the strengths associated with encoder-decoder models, which form a class of neural networks. Segmentation of medical images is usually performed using UNet(Ronneberger, Fischer and Brox, 2015), a well-known encoder-decoder model. This encoder-decoder model is also remarkable in image generation and object detection, so with the emergence of models that include MA-Unet(Cai and Wang, 2022) and SegNet(Badrinarayanan, Kendall and Cipolla, 2017). A variety of encoder-decoders, such as PSPNet(Zhao, Shi, Qi, Wang and Jia, 2017) and DeepLabV3(Chen, Zhu, Papandreou, Schroff and Adam, 2018), aim to improve accuracy by enlarging entire receptive fields. PSPNet uses a Pyramid Pooling Module (PPM) to collect additional global information. The Adaptive Feature Fusion UNet AFF-UNet(Wang, Hu, Shi, Hou, Xu and Zhang, 2023) improves semantic segmentation precision of remote sensing images by using dense skip connections, adaptive feature fusion, channel attention module, and spatial attention model. These operations enable the model to capture interdependence between representations on feature maps.

2.3 Attention mechanism

The addition of attention mechanisms has greatly improved the efficiency of neural networks, especially in image segmentation tasks. With these mechanisms, models are able to focus on specific regions in the input, improving their performance in tasks that require spatial hierarchies and understanding of relationships. SE-UNet(Hu, Shen and Sun, 2018) (Squeeze-and-Excitation) adds a squeeze and excitation block to the classic UNet. This block rescales the channel-wise responses of the functions in a way that takes into account the interdependencies between channels. This allows the network to pay more attention to informative features, leading to better segmentation results. DANet(Fu, Liu, Tian, Li, Bao, Fang and Lu, 2019) (Dual Attention Network) provides a dual attention mechanism that includes both position and channel association to capture global dependencies in both spatial and channel dimensions. This bidirectional attention enhances the model's ability to focus on important details, thus improving segmentation results. However, BAM(Park, Woo, Lee and Kweon, 2018) (Bottleneck Attention Module) is an attention module that can be integrated into network architectures. It generates

separate channel and spatial attention maps. After combining these attention maps, the feature map is rescaled using them as a mechanism to focus on the most important features for better overall accuracy.

2.4 Remote sensing image semantic segmentation

Semantic segmentation of remote sensing images has made tremendous progress with many different and innovative models. CAS-Net, as discussed in (Yang, Wu, Zhang, Zhang, Chen and Gao, 2023), is a significant development that performs coordinate attention (CA) and SPD convolution simultaneously. In terms of architecture, this approach marks a paradigm shift from traditional stepwise convolution with an added pooling layer to ensure that detailed information is preserved during feature extraction. The LPCUNet model (Liu, Wu, Bao and Zhong, 2023) adds to the landscape by focusing on a lightweight pure CNN UNet designed for urban scenario images. This model uses a large convolutional kernel to efficiently capture global context, and a simple fusion module to dynamically combine local and global features. Hi-ResNet, as proposed in (Chen, Fang, Yu, Zhong, Zhang and Li, 2023), is a high-resolution remote sensing network with a number of efficient design structures such as the funnel module, multi-branch module, and feature refinement module. On the other hand, a strategy to improve popular semantic segmentation network structures by merging ResNet-50 with a transformer hybrid model investigated in (Li, Du, Li et al., 2023) suggests a comprehensive approach to spatial distance correlation modeling and maintaining hierarchical nature. The proposed novel Transformer layered architecture, (Chen, Liu, Zhao, Huang and Yan, 2023), couples with CNN through the application of feature dimensionality reduction and a Transformer-style convolutional neural network module to achieve both shallow or deep features effectively. (Wang, Wang, Yang, Wang, Su and Chen, 2022) CFAMNet proposes a class feature attention mechanism integrated into an improved architecture of the Deeplabv3 network to solve typical challenges related to remote sensing images. In (Abdollahi, Pradhan, Shukla, Chakraborty and Alamri, 2021), MCG-UNet and BCL-UNet present new deep convolutional models for multi-object segmentation, where hewer presents the proposed feature of multi-level context gating and integrates bi-directional ConvLSTM at the same time. (Liu, Mi and Chen, 2020) Address VHR segmentation with a multi-level approach and scale-feature attention module. The SSAtNet (Zhao, Liu, Li and Zhang, 2021) proposes an end-to-end attention-based semantic segmentation network, which includes a pyramid attention pooling module for adaptive feature refinement, focusing on channel-wise and spatial path attentions. The multipath encoder structure in (Yang, Li, Chen, Chanussot, Jia, Zhang, Li and Chen, 2021) is designed to build features by providing a comprehensive feature fusion mechanism, the multipath attention-fused blocks. The integrated DCNN proposed in (Su, Li, Ma and Gao, 2022) includes DenseNet, U-Net, Diluted Convolution, and DeconvNet for semantic segmentation to capture subtle details as well as context. TdPFNet (Gu, Hao, Chen and Deng, 2021) is an up-down pyramid fusion network for high-resolution remote sensing semantic segmentation, which presents a multi-source feature extractor, a top-down pyramid fusion module, and also comes with its decoder. SCAttNet (Li, Qiu, Chen, Mei, Hong and Tao, 2020) presents an end-to-end semantic segmentation network that combines lightweight spatial and channel attention modules for adaptive feature refinement. Therefore,

the combination of PSPNet with DeepLabV3 and U-Net an adaptive feature selection module in (Xiang, Xie and Wang, 2021) presents a strong approach to semantic segmentation. MFFANet (He, Zhou, Zhao, Zhang, Yao, Liu and Li, 2021) is about shallow semantic segmentation that combines multiscale feature fusion and attention correction. R2SN (Wang, Wu, Nie and Huang, 2021) follows the classical encoding-decoding paradigm with convolutional layers present in both, which allows it to capture more local information. They include multiscale feature fusion and alignment in MFANet (Wang, Sun and Zhao, 2020), which uses a fully convolutional network, a multi-layer feature fusion block, and the result of building one on top of the other in an encoder. Scale-aware segmentation is addressed by SaNet (Wang, Zhang, Li, Duan, Meng and Atkinson, 2021) through its DCFPN module, which consists of a densely connected feature network. (Zheng, Huan, Xia and Gong, 2020) In EaNet, they developed a separate end-to-end edge-aware neural network for urban scene semantic segmentation and it also includes an LKPP module to acquire rich multi-scale context.(Liu, Zhu, Cao, Chen and Lu, 2021) propose an adaptive multi-scale module as well as the adaptive fuse module that includes channel attention and spatial attention to facilitate feature fusion. Finally, the efficient Hybrid Transformer EHT proposed in (Wang, Fang, Zhang, Li and Duan, 2021) implements a real-time urban scene segmentation by using a CNN-based encoder and a transformer-based decoder for learning globally localized tasks with less computation. Many of these models embody a variety of strategies and innovations that push the boundaries of what can be achieved in semantic segmentation with remotely sensed images.

3 Method

The workflow defined in this paper comprises three main phases, as shown in Figure 1. The first phase is data pre-processing, mainly focusing on image cropping and splitting of datasets. The second is the training, validation, and testing of models. The final section is related to the analysis and evaluation of the obtained results during experiments.

We propose a new semantic segmentation architecture called OTM-UNet 2 which is designed for small object segmentations in Remote Sensing images. As an innovative approach, we make use of learned optimal transport maps to address the problems related to aligning features posed by different scales within the network. The core structure of OTM-UNet is based on the UNet model that has encoder, bottleneck, and decoding with skip connections which makes a good base for effective feature extraction. This is a major change as the use of dynamically learned optimal transport maps for alignment features between encoder and decoder effectively reduces disparities in distributions.

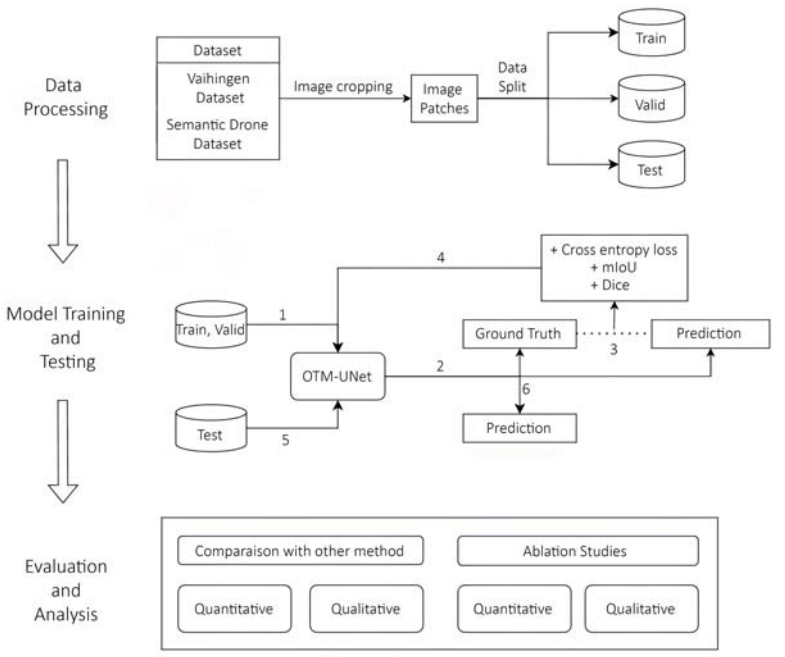

Figure 1: The paper workflow.

3.1 OTM-UNet Architecture

3.1.1 Encoder

The encoder in the OTM-UNet architecture 2 is essential for capturing the hierarchical features of the input image. The encoder consists of a chain of double convolutional blocks followed by max-pooling layers that systematically reduce the resolution to transform the input data into abstract, high-level representations. Each double convolution block includes convolution, batch normalization, and ReLU activations on the features to improve them level by level. Subsequent max-pooling layers effectively reduce spatial dimensions, allowing the model to focus on more salient features. The output of the encoder is a pyramidal arrangement of features, with each layer capturing information at different levels of abstraction. This hierarchical structure provides the basis for the following decoding steps in the OTM-UNet model.

3.1.2 Decoder

The decoder 2 is a set of up-transpose layers, double convolution blocks, and optimal transport layers in which the feature maps are systematically expanding while refining representations. The up-transpose layers carry out a per-pixel procedure to ensure that the model can recover spatial information that was lost during down-sampling. Each up-sampled feature map is then merged with the corresponding encoder features and optimized transport modification to enhance feature matching. Then, double convolution blocks work on the fused features and

enhance their ability to capture complicated details. During the completion of these series operations, a segmented output is presented that refines and sharpens upon giving an in-depth prediction.

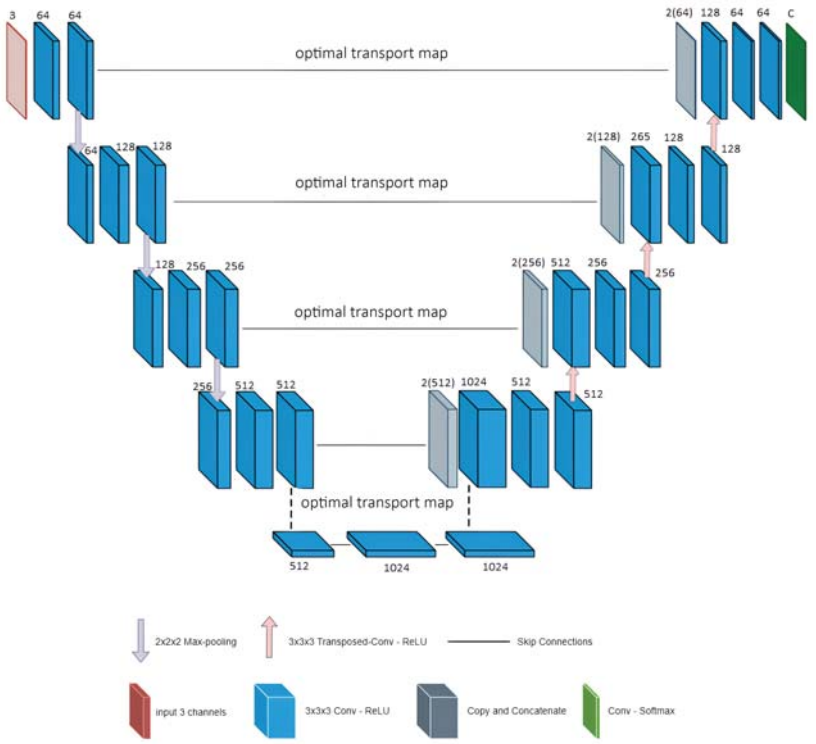

Figure 2: OTM-UNet Architecture.

3.1.3 Optimal transport map

The integration of optimal transport maps is a critical feature of the unique architecture in the OTM-UNet model for semantic segmentation. Optimal transport maps are perfectly squeezed into the UNet framework to solve scale misalignment issues and enhance feature alignment at different levels of abstraction. This integration takes place in the decoder, at various stages of upsampling features from previous layers to be combined with the encoder's feature. The choice to exploit optimal transport maps arises from the fact they can offer a proper scheme for feature alignment, enabling model information optimization transmission across scales. This decision is particularly applicable in cases of semantic segmentation tasks for remote sensing imagery where it is crucial to capture the context at different scales.

In semantic segmentation, the capability of a model to recognize and distinguish objects is connected with accurate feature alignment across scales and contexts. Traditional convolutional neural networks (CNNs) are good at local feature extracting but fail to align global contextual information (Liu, Sun, Wergeles and Shang, 2021). We get around this problem through the use of optimal transport maps that function as a transformational conduit enabling detailed

information to be effortlessly taken across channels.

Figure 3 shows the process of the optimal transport map. It starts with an input feature map, denoted as x, with dimensions $d \times h \times w$. This input feature map is transformed by a learned matrix, denoted as $T(x)$, which is initially an identity matrix but evolves through mapping - the use of an identity matrix is primarily attributed to its simplicity and adaptability. The identity matrix, characterized by ones on the main diagonal and zeros elsewhere, presents a straightforward and uncomplicated structure that serves as an excellent starting point for the learning process. It imposes no initial transformation on the input features, allowing the model to commence from a neutral point without any bias introduced by the initial transformation. Furthermore, the identity matrix's adaptability is a significant advantage. It enables the model to progressively adapt the optimal transport map parameters during training, thereby capturing the intricate relationships between channels for improved performance -. The transformation results in an output feature map, denoted as u. This process rearranges the dimensions of the matched features to align them for concatenation properly. The down-sampling process is connected to the up-sampling process by a skip connection, which indicates that certain features or data will be bypassed during processing to be used later. After the up-sampling process, the adjusted features are concatenated with the up-sampled features from the corresponding layer in the encoder. This entire process enhances the ability of the model to capture context and detail from different scales in the decoding process.

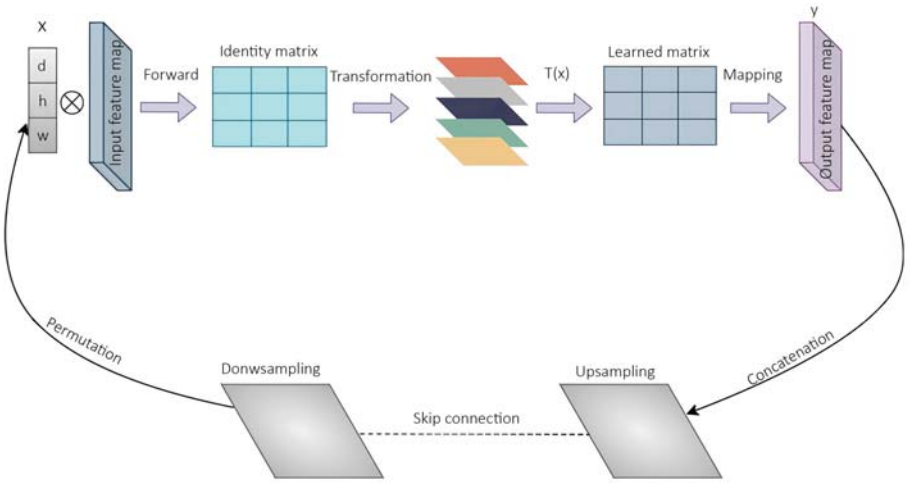

Figure 3: Detailed composition of the OTM.

- **Optimal transport layer forward pass :** Within the optimal transport layer forward pass, the feature map adjustment process y is essential to rectify the scale misalignment problems inherent in semantic segmentation tasks. We consider an input tensor x characterized by dimensions $h \times w \times d$ and introduce a learnable optimal transport matrix T with dimensions $d \times d$. The fitting operation is formulated as follows:

$$y_{ijc} = \sum_{k=1}^{d} x_{ijk} \cdot T_{ck} \tag{3.1}$$

This formula defines the accurate fitting of each element y_{ijc} into the resulting tensor y. The adjustment is performed by a summation process through the channels (k) of the input tensor x. The individual element x_{ijk} is multiplied by the corresponding element T_{ck} of the optimal transport matrix. This per-element adjustment mechanism ensures optimal transport of information across various channels, promoting better feature alignment and ameliorating the challenges associated with scale mismatches in the semantic segmentation process.

- **Shared optimal transport map initialization :** A key step in implementing the learning process within the OTM-UNet architecture is the initialization of the shared optimal transport map. This initialization consists of defining the optimal transport map T as an identity matrix.

$$T_{ij} = \delta_{ij} \tag{3.2}$$

In this equation, δ_{ij} represents the Kronecker delta function, ensuring that each element T_{ij} in the matrix T is 1 when $i = j$ (on the main diagonal) and 0 otherwise. This choice of initialization designates T as an identity matrix, indicating that, initially, no transformation is imposed on the input features. The shared attribute of this optimal transport map underscores its consistent application across different decoding stages, fostering uniformity in the learning process of feature alignments. This deliberate initialization strategy establishes a common starting point, allowing the model to progressively adapt the optimal transport map parameters through training, capturing the intricate relationships between channels for enhanced semantic segmentation performance.

- **Permutation and concatenation :** After obtaining the matched feature maps y characterized by dimensions $h \times w \times d$, a meticulous permutation and concatenation procedure is implemented to optimize the features for eventual concatenation with upsampled counterparts. This multi-step process begins with a permutation operation ($y_{cij} = y_{ijc}$) that strategically rearranges the dimensions of the matched features. The permutation aligns the channel dimension (c) with the second spatial dimension (j), ensuring proper synchronization for subsequent operations. Next, the adjusted features are concatenated with the upsampled features derived from the corresponding layer in the encoder. This concatenation facilitates the fusion of information across different scales, enhancing the model's ability to capture contextual details during the decoding phase. By integrating matched and upsampled features, the model can effectively refine its understanding of the input, fostering a comprehensive representation that spans multiple scales for improved semantic segmentation performance.

- **Parameter update :** The parameter update process, an integral part of the training dynamics, involves the adaptive refinement of the elements within the optimal transport

map T through the mechanism of backpropagation. This complex adaptation is fundamental for the continuous learning of the model during training iterations. The specifics of this update depend on the chosen optimization algorithm, and the crux is to iteratively modify the elements of T to minimize a given loss function. The essence of this dynamic adaptation is to steer the optimal transport map toward configurations that facilitate the most effective feature alignments for the semantic segmentation task. During backpropagation, gradients are computed to the loss function, and the elements of T are updated accordingly to minimize the discrepancy between predicted and ground truth segmentation results. This dynamic adaptation mechanism allows the model to iteratively fine-tune its feature adaptation strategy by learning optimal channel interactions. During training, the optimal transport map evolves to capture intricate relationships between channels, resulting in an increasingly refined and effective representation of semantic features. This continuous refinement contributes to the model's ability to optimize its performance over time, ultimately improving its ability to perform semantic segmentation tasks.

3.2 Datasets

3.2.1 Vaihingen dataset

The Vahingen dataset, referenced as (International Society for Photogrammetry and Remote Sensing, n.d.), serves as a testing ground for digital aerial camera evaluations conducted by the German Society for Photogrammetry and Remote Sensing (DGPF). The dataset also includes airborne laser scanner (ALS) data consisting of 10 ALS strips collected by Leica Geosystems using a Leica ALS50 system with a 45° field of view positioned 500 m above ground. The ALS data captures multiple echoes and intensities, with strip adjustment applied in the background to correct for systematic errors within the ground reference. Designed primarily for analytical purposes, the Vaihingen dataset serves as a benchmark for evaluating urban object extraction techniques.

3.2.2 Semantic drone dataset

This dataset (of Computer Graphics and Vision, n.d.) is an extensive database of images captured from the air intended for increasing semantic understanding regarding urban settings. The dataset provides a distinctive view that contains over 20 house collectors from different altitudes between and above meters. The imagery is created with high resolution, using a 6000×4 pixels 24 megapixels.

Table 1: Details regarding datasets.

Dataset	Platform	Type of View	Flight Altitude
Vaihingen dataset	Airborne Laserscanner	aerial-view	500 m
Semantic drone dataset	UAV	bird-view	5–30 m

3.3 Experimental setting details

3.3.1 Training Settings

The study used a robust computing configuration to train convolutional neural networks (CNNs), using two NVIDIA Tesla T4 GPUs, each equipped with 16 GB of memory. The loss function chosen was the cross-entropy loss, and the optimization used the Adam optimizer, renowned for its efficiency in optimizing neural networks, with a learning rate fixed at 1×10^{-4} (0.0001). The experiments were carried out in a POSIX-compatible environment with 28 GB of RAM. The TensorFlow 2.2 software was chosen for modeling in the experiments. This combination of hardware and software provides a well-organized and efficient infrastructure for research and experimentation.

3.3.2 Evaluation criteria

Evaluating the performance of the different approaches includes five key measures: overall accuracy (OA), F1 score per class, average F1 score, mean intersection over union (mIoU), and Dice coefficient. The OA quantifies the accuracy of pixel identification over the entire dataset. The F1 score per class evaluates the harmonic mean of precision and recall for individual classes, providing a detailed measure of performance. The average F1 score provides an overview by taking the average of the F1 scores across all classes. mIoU measures the average correlation between actual and predicted results at the class level. The introduction of the Dice coefficient further enhances the evaluation by capturing the agreement between predicted and actual segmentations, providing a comprehensive assessment of the overall performance of the model.

$$Overall\ Accuracy = \frac{Number\ of\ Correctly\ Classified\ Pixels}{Total\ Number\ of\ Pixels} \tag{3.3}$$

$$Precision_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \tag{3.4}$$

$$Recall_{class} = \frac{TP_{class}}{TP_{class} + FN_{class}} \tag{3.5}$$

$$F1_{class} = \frac{2 \times Precision_{class} \times Recall_{class}}{Precision_{class} + Recall_{class}} \tag{3.6}$$

$$AvgF1 = \frac{1}{N} \sum_{i=1}^{N} F1_{class_i} \tag{3.7}$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_{class_i}}{TP_{class_i} + FP_{class_i} + FN_{class_i}} \tag{3.8}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.9}$$

4 Experimental result and analysis

4.1 Compare with state-of-the-art methods

we provide a comprehensive comparison with other semantic segmentation methods. This includes our proposed OTM-UNet, which we benchmark with well-known architectures: UNET, DeepLabv3, DANet, PSPNet, and AFF-UNET. Our evaluation uses the Vaihingen dataset for evaluation purposes. Table 2 serves as a comprehensive overview, revealing that the proposed method is the leading performer. It achieves the highest scores across all key metrics: Avg. F1 90.90%, OA 93.17%, mIoU 77.69%, and Dice coefficient 89.5%. These results underscore the proposed method's robust capability in accurately delineating objects within the complex Vaihingen dataset.

A comparative analysis with baseline models, particularly noting the lower performance of UNet, emphasizes the pivotal role of advanced architectures in semantic segmentation tasks. Noteworthy is the proposed method's consistent outperformance DeepLabv3, DANet, and PSPNet. This underscores its effectiveness in capturing intricate spatial relationships and semantic details inherent in the dataset.

Delving deeper into the per-class F1 scores presented in Table 3 provides insights into the proposed method's performance across specific object categories. The method consistently excels in segmenting Buildings, Roads, and Cars, showcasing its versatility in handling diverse structures. While AFF-UNet demonstrates proficiency in Trees, the proposed method remains competitive and, notably, surpasses AFF-UNet in other classes. This balanced performance across various object categories highlights the adaptability of the proposed method. The observed improvements in segmenting Trees and Cars with the proposed method indicate its capacity to handle a spectrum of complex object classes.

Table 2: Performance comparison between our proposed method and different approaches on the Vaihingen dataset.

Method	Avg. F1 (%)	OA (%)	mIoU (%)	Dice (%)
UNet	71.47	74.05	71.81	81.1
DeepLabv3	79.77	76.09	75.41	83.58
DANet	84.12	90.5	76.33	85.33
PSPNet	85.25	90.7	75.2	85.73
AFF-UNet	89.48	92.23	76.78	87.65
Proposed	**90.90**	**93.17**	**77.69**	**89.5**

Table 3: Quantitative comparisons.

Method	Per-class F1 score (%)			
	Buildings	Roads	Trees	Cars
UNet	91.68	85.32	67.81	31.1
DeepLabv3	93.24	86.54	74.73	63.58
DANet	93.39	89.77	76.02	79.33
PSPNet	94.03	90.01	75.25	83.73
AFF-UNet	95.86	92.43	**79.78**	88.85
Proposed	**96.22**	**93.2**	79.69	**91.5**

In addition to the quantitative evaluation 2,3, a more in-depth qualitative comparison was also conducted to assess the visual performance of our proposed method in comparison to other state-of-the-art approaches on the Vaihingen dataset. This analysis 4 was meticulously designed to focus on three key aspects: segmenting confused object classes, segmenting different sizes of object classes, and maintaining the overall integrity of objects.

The first aspect, segmenting confused object classes, is a challenging task due to the similarities between certain classes. In this regard, UNet was observed to have obvious false segmentations, misidentifying impervious surfaces and vehicles as the background class. This indicates a lack of discriminative power in the model to differentiate between similar classes. DeepLabv3, DANet, and PSPNet also exhibited similar issues, suggesting a common challenge across these models. However, AFF-UNet demonstrated good performance in avoiding these misidentifications, indicating its superior discriminative ability. Our proposed method consistently showed accurate segmentations with minimal confusion between classes, demonstrating its robustness in handling class confusion.

The second aspect, segmenting different sizes of object classes, is crucial for maintaining the granularity of the segmentation. UNet, DeepLabv3, DANet, and PSPNet were observed to be unable to effectively consider the segmentation accuracy of both large and small-size objects, leading to suboptimal results. This suggests a limitation in these models in handling objects of varying sizes. In contrast, AFF-UNet showed improved segmentation accuracy for different sizes of objects compared to the previous models, indicating its ability to handle size variations. Our proposed method outperformed all models, effectively considering segmentation accuracy for both large-size buildings and small-size vehicles, showcasing its versatility.

The third aspect, maintaining the overall integrity of objects, is essential for preserving the semantic coherence of the segmentation. In this regard, AFF-UNet demonstrated better maintenance of the integrity of object classes, especially for buildings, indicating its ability to preserve object boundaries. UNet, DeepLabv3, DANet, and PSPNet, however, produced obvious errors in the segmentation of buildings, resulting in fragmented results, suggesting a lack of spatial coherence in these models. Our proposed method maintained object integrity well, even in challenging scenarios, with results closely aligned with the ground truth, demonstrating its robustness in preserving object integrity.

Overall, our proposed method consistently outperformed other comparison models in qualitative aspects, showcasing improved performance in avoiding confusion, considering different

object sizes, and maintaining object integrity. The success of our proposed model is attributed to several innovative features such as dense skip connections, which allow for better information flow; channel attention convolutional blocks, which enable the model to focus on relevant features; adaptive fusion attention modules, which allow for better fusion of multi-scale features; and spatial attention modules, which enable the model to focus on relevant spatial locations.

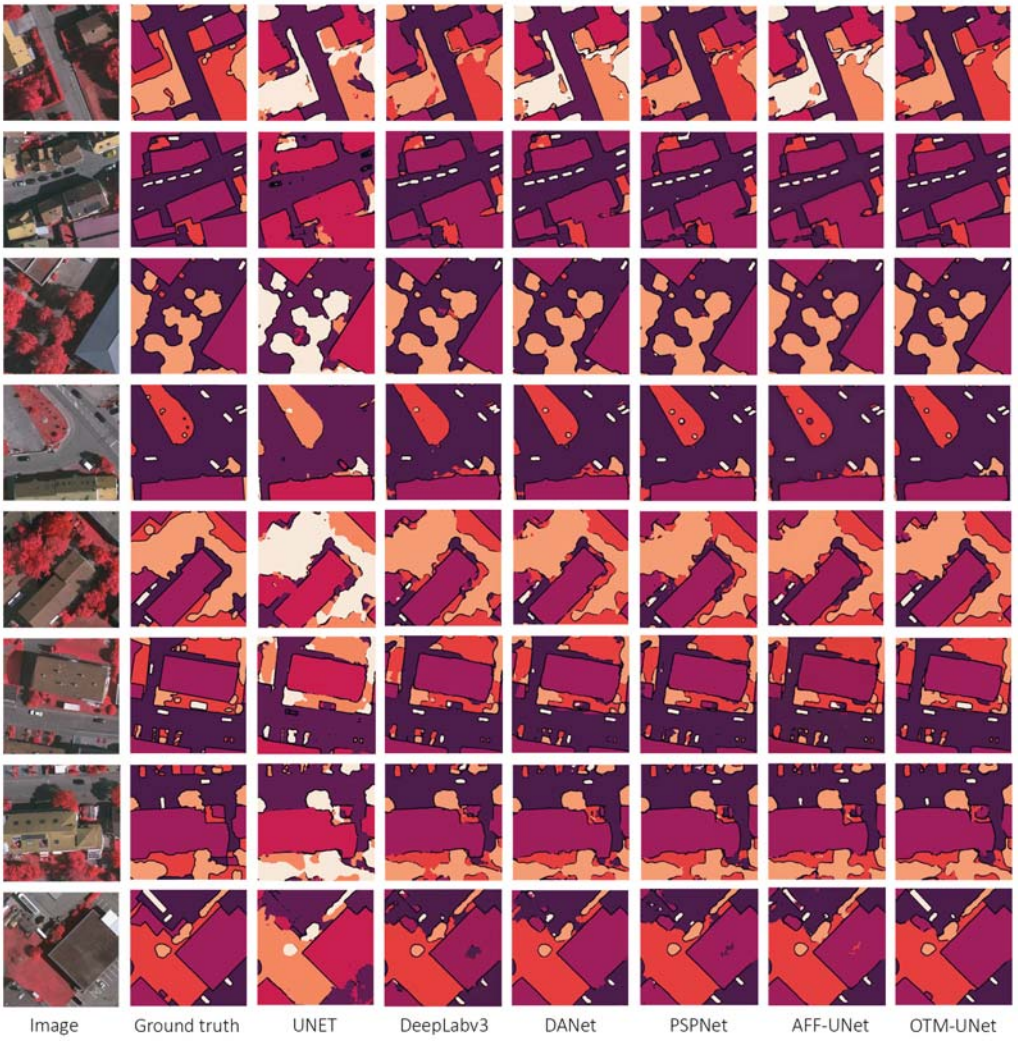

Figure 4: Visualization of the comparison experiments on the Vaihingen dataset.

4.2 Ablation studies

4.2.1 Effect of the number of skip connections

In the first set of experiments 4, we varied the number of skip connections in the OTM-UNet architecture. We evaluated three configurations: 1-SKIP, 2-SKIP, and 3-SKIP. The results indicated that introducing an additional skip connection slightly improved both Average Overall Accuracy (AO) and mean Intersection over Union (mIoU) in the 2-SKIP configuration. However, further increasing the number of skip connections to 3-SKIP resulted in a minor decrease in performance. This suggests that while skip connections enhance feature propagation and information flow, their impact saturates beyond a certain point. Additionally, we observed a linear increase in training time with the number of skip connections, emphasizing the trade-off between improved accuracy and computational efficiency.

4.2.2 Effect of the number of Optimal Transport Maps (OTM)

In the second set of experiments 4, we examined the influence of optimal transport maps on the OTM-UNet architecture. We compared configurations with 1-OTM and 2-OTM. The results demonstrated that adding an extra optimal transport map 2-OTM led to notable improvements in both AO and mIoU. This highlights the effectiveness of OTM in enhancing feature alignment across different scales. However, introducing an additional OTM significantly increased training time, suggesting a trade-off between improved segmentation performance and computational efficiency.

Table 4: Evaluation metrics results of ablation experiments on the Vaihingen dataset with different configurations of the network.

Model	AO (%)	mIoU (%)	Training Time
1-SKIP	93.17	77.69	-
2-SKIP	93.26	77.72	$\approx \times 1.3$
3-SKIP	92.05	76.54	$\approx \times 1.7$
1-OTM	93.17	77.69	-
2-OTM	93.82	78.23	$\approx \times 1.8$

4.2.3 Model complexity and the stability

In our experiments, with the same hardware environment and the same amount of training data, Our model has demonstrated significant improvements in several key areas:

- Reduced Category Confusion: The model has successfully avoided confusion in defining categories. This improvement has allowed for more accurate classifications and predictions, thus raising the efficiency of model functioning.
- Improved Segmentation Results for Different Sizes of Targets: The model can segment the targets with different sizes. This optimization ensures that the model can detect and

delineate targets regardless of their scale facilitating its suitability in a broad range of datasets.

- 3. Improved Target Integrity: The model has also enhanced the integrity of the targets. This entails that the model is superior in terms of sustaining completeness and unity, which are important for preserving meaningfulness or relevance among segmented targets.

These improvements reflect the advantages of our model in dealing with challenging aerial image semantic segmentation tasks and ensuring high-accuracy results. 5,6 illustrates the trend of accuracy and loss validations on the Vaihingen dataset.

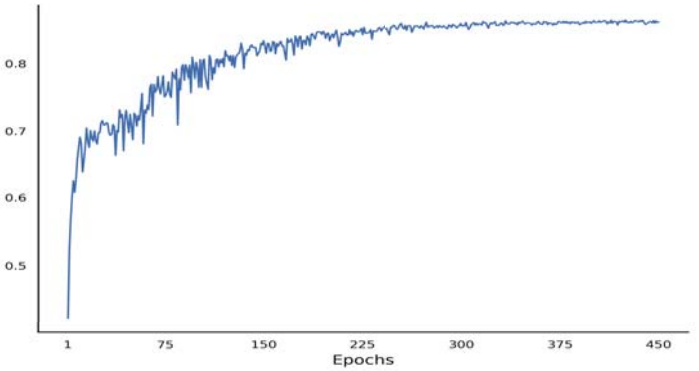

Figure 5: Validation Accuracy.

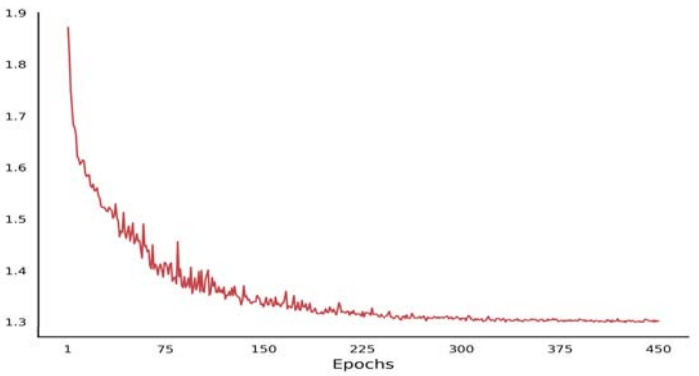

Figure 6: Loss Validation.

4.2.4 Experiment results on SDD dataset

To demonstrate the generalization ability of OTM-UNet, we train our model on the Semantic Drone Dataset (SDD), a task that involves semantic segmentation for aerial imagery. The results obtained underscore the effectiveness of OTM-UNet in this context, with notable performance metrics. The achieved Average F1 score of 91.85% reflects the model's precision and

recall balance, while the Overall Accuracy (OA) reaches an impressive 94.03%, highlighting the model's proficiency in correctly classifying pixels across diverse drone images. Moreover, the mean Intersection over Union (mIoU) score of 81.62% indicates the model's accuracy in delineating object boundaries. Notably, the Dice coefficient attains a high value of 91.15%, further emphasizing the precision of segmentation outcomes. These results collectively demonstrate the OTM-UNet robustness and efficacy in achieving superior semantic segmentation performance on aerial imagery from the Semantic Drone Dataset.

5 Conclusion

This study presents the OTM-UNet, a novel semantic segmentation architecture designed to address small object detection in aerial imagery. Leveraging learned optimal transport maps, our model demonstrated its ability to align both encoder and decoder feature maps to achieve better spatial alignment and semantic consistency during sampling. The improved U-net architecture with optimal transport layers addressed a persistent problem in the identification and segmentation of small objects in the scenes. The precise feature alignment mechanism proved critical to the successful transfer of information, resulting in a better appreciation of spatial detail across the architecture. The ability of OTM-UNet to overcome these limitations makes it a reliable solution for applications where accurate and well-defined segmentation in an aerial image is highly desired. What these contributions have achieved in this paper goes beyond simply advancing the field of semantic segmentation, but also provides significant insight into addressing specific challenges in detecting and segmenting small objects in remote sensing. However, with additional efforts in the research process, the innovative features and approaches incorporated in OTM-UNet will significantly benefit the existing deep learning methods for remote sensing applications.

Data Availability Statement

We utilize open data from the following datasets:
Vaihingen dataset: `https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx`. Semantic segmentation drone dataset: `https://www.tugraz.at/index.php?id=22387`

References

Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S. and Alamri, A. 2021. Multi-object segmentation in complex urban scenes from high-resolution remote sensing data, *Remote Sensing* **13**(18): 3710.

Badrinarayanan, V., Kendall, A. and 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.

Barbarella, M., Cuomo, A., Di Benedetto, A., Fiani, M. and Guida, D. 2019. Topographic base maps from remote sensing data for engineering geomorphological modelling: An application on coastal mediterranean landscape, *Geosciences* **9**(12): 500.

Cai, Y. and Wang, Y. 2022. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation, *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, Vol. 12167, SPIE, pp. 205–211.

Camps-Valls, G. and Bruzzone, L. 2009. *Kernel methods for remote sensing data analysis*, John Wiley & Sons.

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

Chen, Y., Fang, P., Yu, J., Zhong, X., Zhang, X. and Li, T. 2023. Hi-resnet: A high-resolution remote sensing network for semantic segmentation, *arXiv preprint arXiv:2305.12691* .

Chen, Y., Liu, P., Zhao, J., Huang, K. and Yan, Q. 2023. Shallow-guided transformer for semantic segmentation of hyperspectral remote sensing imagery, *Remote Sensing* **15**(13): 3366.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H. 2019. Dual attention network for scene segmentation, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3146–3154.

Gu, Y., Hao, J., Chen, B. and Deng, H. 2021. Top-down pyramid fusion network for high-resolution remote sensing semantic segmentation, *Remote Sensing* **13**(20): 4159.

He, K., Zhang, X., Ren, S. and Sun, J. 2016. Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, X., Zhou, Y., Zhao, J., Zhang, M., Yao, R., Liu, B. and Li, H. 2021. Semantic segmentation of remote-sensing images based on multiscale feature fusion and attention refinement, *IEEE Geoscience and Remote Sensing Letters* **19**: 1–5.

Hemanth, D. J. and Estrela, V. V. 2017. *Deep learning for image processing applications*, Vol. 31, IOS Press.

Hu, J., Shen, L. and Sun, G. 2018. Squeeze-and-excitation networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q. 2017. Densely connected convolutional networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huong, N. T., Long, N. T., Kozlov, , Tomin, N. and Sidorov, D. 2021. Deep learning methods for classification of road defects, *International Journal of Artificial Intelligence* **19**(1): 178–192.

International Society for Photogrammetry and Remote Sensing n.d.. Vaihingen dataset. Accessed: June 5, 2024.
URL: *https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx*

Khriss, A., Elmiad, A. K., Badaoui, M., Barkaoui, A.-E. and Zarhloule, Y. 2024a. Exploring deep learning for underwater plastic debris detection and monitoring, *Journal of Ecological Engineering* **25**(7): 58–69.

Khriss, A., Elmiad, A. K., Badaoui, M., Barkaoui, A. and Zarhloule, Y. 2024b. Advances in machine learning and deep learning approaches for plastic litter detection in marine environments, *Journal of Theoretical and Applied Information Technology* **102**(5).

Kim, Y. and Kim, Y. 2017. Optimizing neural network to develop loitering detection scheme for intelligent video surveillance systems, *International Journal of Artificial Intelligence* **15**(2): 30–39.

Li, H., Qiu, K., Chen, L., Mei, X., Hong, L. and Tao, C. 2020. Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images, *IEEE Geoscience and Remote Sensing Letters* **18**(5): 905–909.

Li, J., Du, Q., Li, W. et al. 2023. Mcafnet: a multiscale channel attention fusion network for semantic segmentation of remote sensing images. remote sens 15: 361.

Li, Y., Zhang, H., Xue, X., Jiang, Y. and Shen, Q. 2018. Deep learning for remote sensing image classification: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(6): e1264.

Li, Z., Wang, Y., Zhang, N., Zhang, Y., Zhao, Z., Xu, D., Ben, G. and Gao, Y. 2022. Deep learning-based object detection techniques for remote sensing images: A survey, *Remote Sensing* **14**(10): 2385.

Liu, B., Wu, H., Bao, X. and Zhong, Z. 2023. Lpcunet: A lightweight pure cnn unet for efficient urban scene remote sensing semantic segmentation, *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, IEEE, pp. 57–61.

Liu, R., Mi, L. and Chen, Z. 2020. Afnet: Adaptive fusion network for remote sensing image semantic segmentation, *IEEE Transactions on Geoscience and Remote Sensing* **59**(9): 7871–7886.

Liu, Y., Sun, P., Wergeles, N. and Shang, Y. 2021. A survey and performance evaluation of deep learning methods for small object detection, *Expert Systems with Applications* **172**: 114602.

Liu, Y., Zhu, Q., Cao, F., Chen, J. and Lu, G. 2021. High-resolution remote sensing image segmentation framework based on attention mechanism and adaptive weighting, *ISPRS International Journal of Geo-Information* **10**(4): 241.

Long, J., Shelhamer, E. and Darrell, T. 2015. Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Long, N., Alexander, A. and Huong, N. 2021. Segmentation of forest fire images based on convolutional neural networks, *International Journal of Artificial Intelligence* **19**(1): 21–35.

Navalgund, R. R., Jayaraman, V. and Roy, P. 2007. Remote sensing applications: An overview, *current science* pp. 1747–1766.

of Computer Graphics, T. I. and Vision, G. U. o. T. n.d.. Semantic drone dataset. Accessed: June 5, 2024.
URL: *https://www.tugraz.at/index.php?id=22387*

Park, J., Woo, S., Lee, J.-Y. and Kweon, I. S. 2018. Bam: Bottleneck attention module, *arXiv preprint arXiv:1807.06514* .

Ronneberger, O., Fischer, P. and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, pp. 234–241.

Shao, H., Song, P., Mu, B., Tian, G., Chen, Q., He, R. and Kim, G. 2021. Assessing city-scale green roof development potential using unmanned aerial vehicle (uav) imagery, *Urban Forestry & Urban Greening* **57**: 126954.

Su, Z., Li, W., Ma, Z. and Gao, R. 2022. An improved u-net method for the semantic segmentation of remote sensing images, *Applied Intelligence* **52**(3): 3276–3288.

Tong, K., Wu, Y. and Zhou, F. 2020. Recent advances in small object detection based on deep learning: A review, *Image and Vision Computing* **97**: 103910.

Triharminto, H., Adji, T. and Setiawan, N. 2013. 3d dynamic uav path planning for interception of moving target in cluttered environment, *International Journal of Artificial Intelligence* **10**(S13): 154–163.

Wang, C., Wu, D., Nie, J. and Huang, L. 2021. R2sn: Refined semantic segmentation network of city remote sensing image, *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*, Springer, pp. 380–396.

Wang, L., Fang, S., Zhang, C., Li, R. and Duan, C. 2021. Efficient hybrid transformer: Learning global-local context for urban sence segmentation, *arXiv e-prints* pp. arXiv–2109.

Wang, L., Zhang, C., Li, R., Duan, C., Meng, X. and Atkinson, P. M. 2021. Scale-aware neural network for semantic segmentation of multi-resolution remote sensing images, *Remote sensing* **13**(24): 5015.

Wang, X., Hu, Z., Shi, S., Hou, M., Xu, L. and Zhang, X. 2023. A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved unet, *Scientific Reports* **13**(1): 7600.

Wang, Y., Sun, Z. and Zhao, W. 2020. Encoder-and decoder-based networks using multiscale feature fusion and nonlocal block for remote sensing image semantic segmentation, *IEEE Geoscience and Remote Sensing Letters* **18**(7): 1159–1163.

Wang, Z., Wang, J., Yang, K., Wang, L., Su, F. and Chen, X. 2022. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with deeplabv3+, *Computers & Geosciences* **158**: 104969.

Xiang, S., Xie, Q. and Wang, M. 2021. Semantic segmentation for remote sensing images based on adaptive feature selection network, *IEEE Geoscience and Remote Sensing Letters* **19**: 1–5.

Yang, X., Li, S., Chen, Z., Chanussot, J., Jia, X., Zhang, B., Li, B. and Chen, P. 2021. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery, *ISPRS Journal of Photogrammetry and Remote Sensing* **177**: 238–262.

Yang, Z., Wu, Q., Zhang, F., Zhang, X., Chen, X. and Gao, Y. 2023. A new semantic segmentation method for remote sensing images integrating coordinate attention and spd-conv, *Symmetry* **15**(5): 1037.

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

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. 2017. Pyramid scene parsing network, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890.

Zhao, Q., Liu, J., Li, Y. and Zhang, H. 2021. Semantic segmentation with attention mechanism for remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing* **60**: 1–13.

Zheng, X., Huan, L., Xia, G.-S. and Gong, J. 2020. Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss, *ISPRS Journal of Photogrammetry and Remote Sensing* **170**: 15–28.