APSIPA Transactions on Signal and Information Processing, 2024, 13, e202

Original Paper

AMBNet: Adaptive Multi-feature Balanced Network for Multimodal Remote Sensing Semantic Segmentation

Xiaochen Xiu[1], Xianping Ma[1], Man-On Pun[1*] and Ming Liu[2]

[1] *School of Science and Engineering, the Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong, Shenzhen, China*
[2] *MizarVision, Shanghai, China*

ABSTRACT

This work proposes an Adaptive Multi-feature Balanced network (AMBNet) for semantic segmentation in complex urban remote sensing scenarios. To fully exploit optical images and Digital Surface Models (DSM) data obtained from remote sensing sensors, a Depth Feature Extraction and Balancer (DFEB) module is devised to estimate and balance the depth information of all pixels by capturing detailed structural compositions of the ground surface. After that, a Parallel Multi-Stage Segmentator (PMSS) comprised of a dual-branch Encoder and Decoder with skip connections is constructed to perform effective segmentation by exploiting the balanced DSM (BDSM) and optical information. As a result, the proposed AMBNet can make effective use of optical images to complete depth information, so as to achieve multimodal information-assisted semantic segmentation for complex remote sensing scenes. Comprehensive experiments performed on the ISPRS Vaihingen and Potsdam remote sensing datasets confirm the segmentation performance of the proposed method.

*Corresponding author: Man-On Pun, SimonPun@cuhk.edu.cn.

Received 17 October 2023; Revised 20 December 2023
ISSN 2161-1823; DOI 10.1561/116.00000071

Keywords: Height estimation, Feature fusion, Multimodal semantic segmentation

1 Introduction

Semantic segmentation plays a pivotal role within the realm of remote-sensing image processing. Nevertheless, the semantic segmentation of remote sensing images is fraught with distinct technical intricacies when contrasted with its counterpart designed for natural images. For instance, extensive shadows cast by edifices can obscure ground areas whereas the dense canopy of vegetation impedes the identification of vehicles and structures. These intricacies necessitate the employment of more sophisticated techniques in the domain of remote sensing semantic segmentation. Driven by recent breakthroughs in the field of deep learning (DL), a spectrum of DL-based segmentation models have been devised for remote sensing applications [11, 5, 24, 18, 26]. These models can be divided into two main categories according to their network structures, namely the CNN-based networks [18, 17] and the Transformer-based networks [26, 4]. The former offers superior computational efficiency during training and test at the cost of a limited local receptive field, hindering their capability of effectively discerning high-level object semantics. In contrast, the latter enhances the segmentation performance through the utilization of Vit-based encoders [6], effectively capturing intricate long-range dependencies. Nevertheless, Transformer-based models come at a formidable computational cost.

In the meantime, recent researches have suggested that augmenting the input data with multimodal information, such as multispectral imagery (MSI) [14, 25], hyperspectral imagery (HSI) [20, 28], Digital Surface Models (DSM) [23, 30] and light detection and ranging (LiDAR) [21, 24], together with optical images, can significantly enhance the semantic segmentation accuracy, particularly for objects exhibiting similar chromatic attributes on the terrain [7, 29]. For instance, FuseNet [9] and vFuseNet [2] enhanced the encoding of ground elevation data by seamlessly integrating optical images with DSM. Furthermore, CMFNet [19] employed a crossmodal multi-scale Transformer to perform a comprehensive fusion of multimodal data. While the aforementioned methods demonstrate effectiveness, they are not without limitations. Firstly, data labeling poses a challenge as not all remote sensing images across different regions have accessible multi-source data for reference. Secondly, the fusion of multi-source hyperspectral data introduces complexities such as data registration, error correction, and consistency issues. The uncertainty and noise from diverse data sources may propagate into the fusion results, impacting the model's stability. Thirdly, handling shadows presents difficulties, as they often involve lighting changes, demanding high adaptability from the model. Accurately removing or correcting shadows can prove challenging. Furthermore, these methods assume the reliability of Digital Surface Model (DSM) data, which can be a practical concern. The precision of DSM data depends on various environmental factors, including the topography of the study area. To address this concern, this work proposes a depth estimation method that leverages detailed information from optical sources to generate more accurate and reliable DSM data. This refined approach, referred to as balanced DSM (BDSM), serves as the foundation for subsequent multimodal fusion processes.

Monocular depth prediction endeavors to extrapolate the spatial depth particulars of a scene from a solitary remote sensing image. This depth information can furnish the model with richer contextual insights, facilitating a more profound comprehension of the remote sensing imagery. Simultaneously, it serves to alleviate the impact of terrain irregularities and shadows, proving particularly advantageous in addressing challenges posed by uneven topography, mountainous terrains, and shaded regions. The derived depth information can be employed as an ancillary input for semantic segmentation tasks, enhancing the model's capacity to discriminate between diverse categories of ground objects. The outcomes of depth prediction may be harnessed to generate sophisticated semantic features, elevating the accuracy of ground object boundary delineation in segmentation tasks.

Multimodal fusion approaches have been developed for many years in the field of computer vision. For example, intensive investigations have been devoted to deriving multi-feature information from spatial geometry and focal length changes. Some pioneering works [3] have been established by utilizing multi-feature fusion to aid semantic segmentation. More specifically, different algorithms are employed in feature-based image data fusion to extract features from various data sources before fusing these features. For instance, color transformation [15] and spatial-to-frequency domain transformation [16] are utilized to extract geometric and spectral characteristics of the target, including range, shape, neighborhood, texture, relative positioning, and spectral information [8]. Compared with natural images, multimodal fusion for remote sensing images encounters more challenges due to the following facts: (1) Remote sensing images are mostly collected by unmanned aerial vehicles or satellites in orbit in the form of a top view of limited angles; (2) The objects in remote sensing images are more complex, including varying scales, irregular shapes or boundaries. These problems require more sophisticated network designs for remote sensing multi-feature fusion.

To address these challenges, an Adaptive Multi-Feature Balancing Network (AMBNet) is proposed for semantic segmentation of complex urban remote sensing scenes in this work. It utilizes a deep feature extraction and balancer (DFEB) for depth information extraction and balance before exploiting a parallel multi-stage segmentator (PMSS) for multimodal data fusion. Specifically, DFEB utilizes a height estimator and feature merge balancer to improve the quality of DSM by combining DSM and ground features for optical images, using Gaussian filtering [13] to smooth the DSM data. In particular, the proposed DFEB can effectively reduce noise while preserving image details. To assess the effectiveness of the feature fusion process, we propose to evaluate the similarity between remote sensing images and the corresponding enhanced BDSM features using performance indicators such as SSIM, perceptual hashing, LBP, and NMSE. After that, the generated BDSM and optical images are passed into the PMSS equipped with parallel ResNet Layers and decoder layers with skip connections for the final segmentation prediction. By leveraging depth estimation and multimodal data fusion, the proposed AMBNet opens up new prospects for semantic segmentation of remote sensing images. Extensive experiments are conducted on ISPRS Vaihingen and Potsdam remote sensing datasets to verify the performance of the proposed AMBNet in multimodal remote sensing segmentation tasks.

2 Proposed AMBNet

2.1 Adaptive Multi-Feature Balancing Network (AMBNet)

Figure 1 illustrates the schematic of the proposed AMBNet consisting of two novel modules, namely DFEB and PMSS. To explore multimodal information more efficiently, the proposed AMBNet is optimized in an end-to-end manner. As shown in Figure 1, both optical and DSM are input into AMBNet in which the DFEB module generates the BDSM by exploiting auxiliary crossmodal information. More specifically, the proposed DFEB integrates optical and DSM information using a variety of techniques, including Laplacian pyramid decomposition [27], Gaussian blur, and weighted fusion. Multiple fusion evaluation metrics can be employed, including SSIM, Perceptual Hashing, LBP analysis, and NMSE, to facilitate the feature matching between the fused feature map and the pre-processed optical image, thereby assessing the quality of the fusion process. Given the inherent imprecision and lack of fine-grained detail in much of the DSM data, especially with regard to objects such as vehicles, the richness of detailed feature information is primarily embedded in the optical image. However, the original optical image is burdened with excessive high-frequency information that does not always align with the fusion outcome. To address this challenge, comparison is drawn between the Laplacian pyramid-pre-processed results and the fused DSM data. Finally, the resulting BDSM is fed into PMSS along with the raw optical images for final semantic segmentation. The proposed PMSS undertakes multi-dimensional feature extraction from both the enhanced DSM information and the original optical images, culminating in the production of the final semantic segmentation output after upsampling.

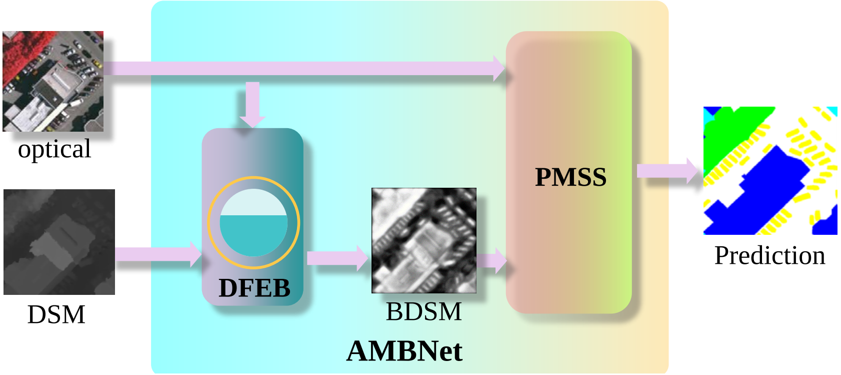

Figure 1: Structure of the proposed Adaptive Multi-feature Balanced Network (AMBNet).

2.2 Deep Feature Extraction and Balancer (DFEB)

The proposed DFEB module, as depicted in Figure 2, embodies two distinct components, namely the height estimator and the feature merge balancer. The former is designed for estimating the height of ground objects, consisting of a classical encoder-decoder structure that can output highly recognizable estimated DSM (EDSM) based on optical images. Upon receiving EDSM and DSM, the latter derives reliable BDSM by weighing EDSM and DSM with the Laplacian pyramid.

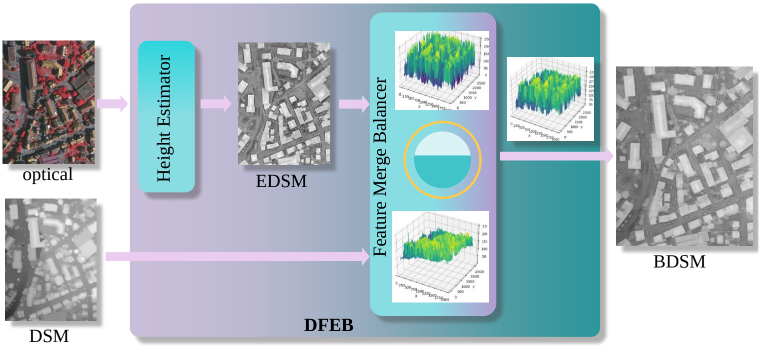

Figure 2: The structure of Deep Feature Extraction and Balancer (DFEB).

Height Estimator: Stemmed from its foundational ResNet [10], the height estimator capitalizes on the pivotal MidasNet [22] while introducing feature screening together with channel attention mechanisms [12]. Furthermore, the amalgamation of feature channels is optimized through the back-propagation process. The structure of the height estimator is shown in Figure 3. It is

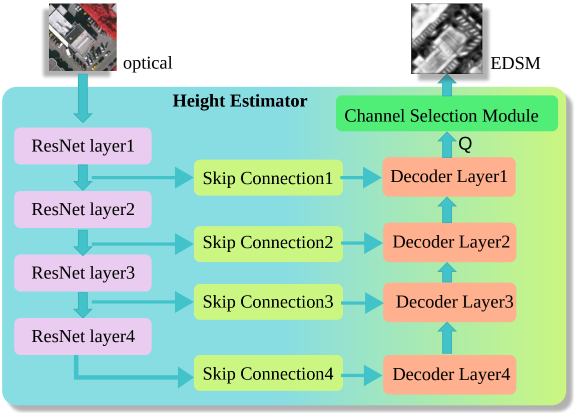

Figure 3: The structure of the proposed height estimator module.

comprised of four ResNet layers, four skip connections, four corresponding Decoder layers and a channel selection module. More specifically, ResNet layers processes the optical image input denoted as $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$. The feature maps of four different scales are obtained by four ResNet layers that also serve as skip connections. Four decoder layers recover the spatial and contextual information to generate the corresponding depth feature maps denoted as $\mathbf{Q} \in \mathbb{R}^{H \times W \times 64}$. Finally, a channel attention selection module is implemented to assign weighting coefficients to useful information. We denote by $\mathbf{F}_{se}(\cdot)$ the squeeze and excitation function, and the final EDSM can be generated as follows:

$$\mathrm{EDSM} = \mathbf{F}_{se}\left(\mathbf{Q}\right). \tag{1}$$

More specifically, the channel attention mechanism incorporated in the framework of monocular depth estimation draws its inspiration from SENet (Squeeze and Excitation Network), wherein the conventional Rectified Linear Unit (ReLU) activation function is substituted with the *Tanh* function. This strategic replacement allows for the comprehensive utilization of both positive and negative tensor information present in the depth feature map. It is imperative to note that this is a channel-centric attention model specifically designed to capture and quantify the significance of individual feature channels. Consequently, it facilitates the selective amplification or attenuation of specific channels tailored to different task requirements. Following the convolutional operation, a branch for bypass is introduced, commencing with the Squeeze operation. This operation condenses the spatial dimensions into feature values, effectively collapsing each two-dimensional feature map into a scalar value. Essentially, it approximates a pooling operation with a global receptive field. Remarkably, the total count of feature channels remains unaltered throughout this process.

2.3 Feature Merge Balancer (FMB)

As shown in Figure 4, the proposed Feature Merge Balancer employs a multi-faceted approach to fuse the EDSM and DSM data by exploiting the Laplacian pyramid method and the Gradient domain fusion (GDF). More specifically, three fusion images denoted as IMG1, IMG2 and IMG3 are generated and compared against the RGB image separately in terms of four metrics, namely Structural Similarity Index (SSIM), Perceptual Hashing, LBP, and NMSE. After that, the model derives the Overall Similarity (OS) from the four resulting indicators before the optimal fusion solution is selected. In particular, since SSIM is specifically designed to quantify the similarity between two images, it assesses the perceived quality of an image by considering various aspects of the structural information and luminance conditions that are important factors for human visual perception. Finally, adaptive weights are assigned to the four indicators to provide comprehensive evaluation on the Overall Similarity.

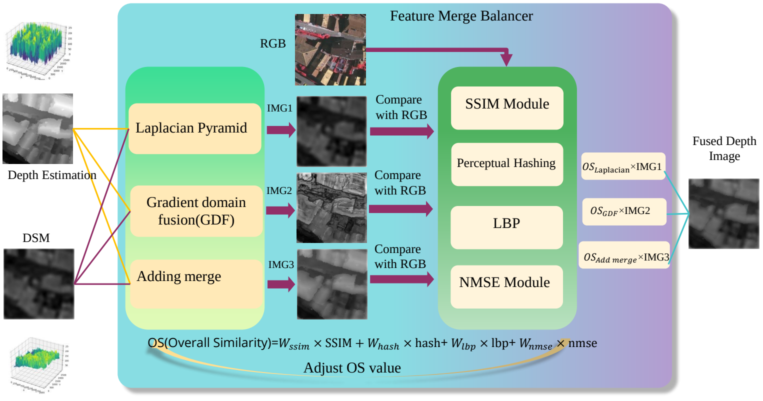

Figure 4: The structure of Feature Merge Balancer.

The Laplacian pyramid shown in Figure 5 is a multi-scale image representation. Firstly, the original image is downsampled for multiple times to generate a series of images of different resolutions, each of which is a blurred version of the original image, which develops a Gaussian pyramid [1]. We denote by $\mathbf{G}_0$ and $\mathbf{G}_i$ the input image and the i-th level image of the Gaussian pyramid, respectively. It is worth noting that the higher-level images have lower resolution.

Next, we create a Laplacian pyramid by subtracting the Gaussian image of an adjacent level from one level in the Gaussian pyramid. This step generates a series of images, each of which contains details of the original image at different spatial frequencies. This process can be mathematically modeled as:

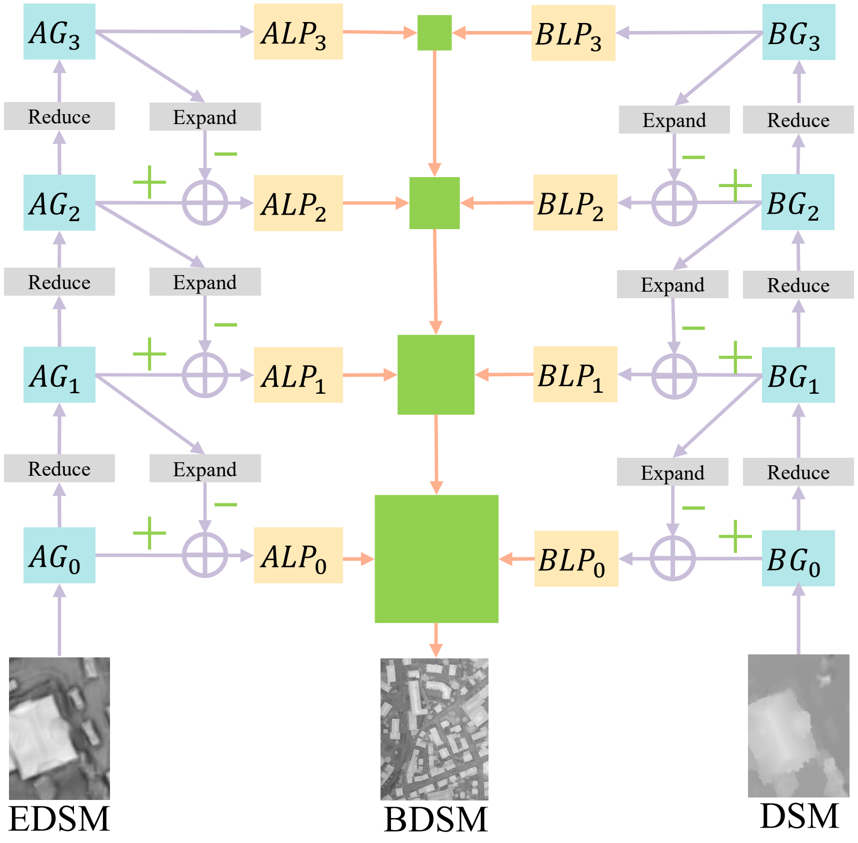

Figure 5: The structure of Laplace fusion pyramid.

$$L_i = \mathbf{G}_i - Expand\left(\mathbf{G}_{i+1}\right) \otimes \kappa_{5\times 5}, \tag{2}$$

where $Expand\left(\cdot\right)$ models the process of upsampling, mapping the pixel at position (x, y) in the source image to the position $(2x+1, 2y+1)$ of the target image. Furthermore, $\otimes$ stands for the convolution operator.

To fuse two images, the specific structure described above performs the same operation on the Laplacian pyramids of two inputs before the Laplacian images at the corresponding levels are added together with weights. The weights are usually determined by a fusion mask that defines the image's information that should dominate at each level. Note that the variance is large for the current local block in the detail areas of the Laplacian image. In contrast, the variance is small for the smooth areas.

2.4 Parallel Multi-Stage Segmentator (PMSS)

The proposed PMSS is designed to perform the multimodal feature fusion and subsequently, generate the final segmentation map. As shown in Figure 6, two parallel encoders comprised of multiple ResNet Layers are utilized to encode BDSM and optical images. More specifically, the encoders are divided into four blocks with the output feature maps being denoted as $\mathbf{M}_{optical} \in \mathbb{R}^{H \times W \times C_i}$ and $\mathbf{M}_{Depth} \in \mathbb{R}^{H \times W \times C_i}$ where $C_i \in \{64, 128, 256, 512\}$ for $i = 1, 2, 3, 4$. After each downsampling operation, the BDSM is fused into the optical encoder as the auxiliary information to the next downsampling operation. Mathematically, the resulting post-fusion feature map $\mathbf{M}_{Fusion}$ can be modeled as follows:

$$\begin{aligned} \mathbf{M}_{Fusion} \quad &= \quad \mathbf{M}_{optical} \times \mathbf{F}_{se}\left(\mathbf{M}_{optical}\right) + \\ & \quad\quad \mathbf{M}_{Depth} \times \mathbf{F}_{se}\left(\mathbf{M}_{Depth}\right). \end{aligned} \tag{3}$$

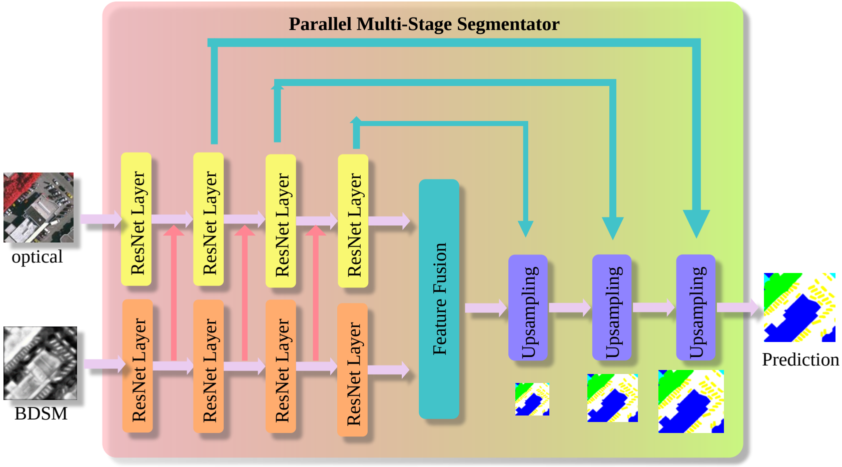

Figure 6: Structure of the proposed PMSS.

To augment the preservation of intricate long-term details, skip connections have been strategically incorporated, facilitating seamless information flow from the encoder to the decoder components. Ultimately, the process culminates with the fusion of depth information and the upsampled feature maps, which is shown to be a critical precursor for improving the segmentation performance in the following operation.

2.5 Loss and Activation functions

The following loss function L_{FMB} is employed in this work:

$$\begin{aligned} L_{FMB} = w_1 * SSIM + w_2 * PerceptualHashing \\ +w_3 * LBP + w_4 * NMSE, \end{aligned} \tag{4}$$

where $\{w_i\}$ with $i = 1, 2, 3, 4$ stands for the weighting coefficients designed to reflect the relative importance of the corresponding performance metric during the training process.

In addition, the softmax function is used as the activation function in the classification process, which maps the output of multiple neurons to the probability within the interval of $(0, 1)$ for multi-class classification. Furthermore, the softmax function is used in conjunction with the cross entropy loss function to avoid the problem of numerical overflow.

The following softmax activation function S_i and softmax loss function L_{sft} are employed in this work:

$$S_i = \frac{e^{p_i}}{\sum\limits_{c=1}^{C} e^{p_c}}, \tag{5}$$

$$L_{sft} = -\sum_{i=1}^{C} q_i \log{[P]_i}, \tag{6}$$

where p_i is the output value at the i-th node while C is the total number of output nodes or classes. Furthermore, $[P]_i$ is the i-th value of the output P, which represents the probability that this sample is classified into the i-th category. Finally, $q_i = 1$ if the sample actually belongs to the i-th category. Otherwise, we set $q_i = 0$.

Using the softmax function S_i as the probability distribution of p_i, we can obtain the Cross Entropy (CE) loss function as follows:

$$L_{CE} = -\sum_{i=1}^{C} q_i S_i. \tag{7}$$

If there are multiple samples, the CE loss function can be computed as the mean value over all samples.

Finally, the following overall loss function combines L_{FMB} and L_{CE} to enable diverse optimization objectives.

$$L_{Total} = \gamma L_{FMB} + (1 - \gamma) L_{CE}, \tag{8}$$

where $\gamma \in [0, 1]$ is a weighting coefficient. Adjustment of γ allows the network to learn various aspects of knowledge. As a result, the network can enhance image fusion while simultaneously improving pixel-level classification accuracy.

3 Fusion Quality Metrics

3.1 SSIM (Structural Similarity)

SSIM is a measure designed to detect the similarity of two given images of the same size by comparing the brightness, contrast, and structure of the two images.

$$S(x,y) = f\left(I(x,y), c(x,y), s(x,y)\right), \tag{9}$$

where $I(x,y)$, $c(x,y)$ and $s(x,y)$ are the brightness contrast function, the contrast function and the structural contrast function, respectively. These functions are defined as follows:

$$\begin{aligned} I(x,y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, & (10) \\ c(x,y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, & (11) \\ s(x,y) &= \frac{\mu_{xy} + C_3}{\mu_x\mu_y + C_3}, & (12) \end{aligned}$$

where μ_x, σ_x and μ_{xy} are given below:

$$\begin{aligned} \mu_x &= \frac{1}{N}\sum_{i=1}^{N} x_i, & (13) \\ \sigma_x &= \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}}, & (14) \\ \mu_{xy} &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(-\mu_y). & (15) \end{aligned}$$

Finally, constants C_1, C_2 and C_3 are introduced for mathematical stability by preventing the division-by-zero problem in the denominator.

It is worth noting that these formulae possess inherent clarity, necessitating the computation of mean and variance for each of the two images independently, followed by the derivation of their covariance, which is subsequently incorporated into the SSIM formula for evaluation. The resultant computation output represents an image, effectively illustrating the aliasing artifacts present in both images.

3.2 Local Binary Pattern (LBP)

LBP is an image-processing metric for texture analysis. The calculation of the LBP features is based on the neighborhood of pixels, usually using 3 × 3 or 5 × 5 neighborhoods. We consider a central pixel with gray value of I_c surrounded by K pixels whose gray values are denoted as $I_{p_1}, I_{p_2}, \cdots, I_{p_K}$. Then, the LBP value of the central pixel is computed by comparing its gray value against those of its neighboring pixels. If the gray value of the neighboring pixel is greater than or equal to the central pixel, then we count the pixel in the LBP calculation with pre-defined weighting. Otherwise, the contribution of the neighboring pixel to the LBP value of the central pixel is ignored. Mathematically, the computation of the LBP value can be expressed as follows:

$$LBP(I_c) = \sum_{k=1}^{K} s(I_k - I_c) * 2^k, \tag{16}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Finally, a K-bit binary LBP code is formed by connecting these binary codes together.

After obtaining the LBP code, we can utilize the histogram intersection kernel for similarity measurement. More specifically, we denote by H_1 and H_2 two histograms of LBP features with each comprising ℓ bins, e.g. $\ell = 256$ for 8-bit LBP representation. The calculation of the histogram intersection kernel is delineated as follows:

$$Similarity = \sum_{i=1}^{\ell} \min\left(H_1[i], H_2[i]\right). \tag{18}$$

Note that Eq. (18) stands for the similarity between two histograms by measuring their overlaps. A larger similarity value indicates that the texture features of the two images are more similar.

In summary, SSIM and LBP serve as two image quality evaluation metrics grounded in human perception. They are primarily employed for assessing the structural similarity and perceptual quality of images, showcasing relevance to the accuracy of semantic segmentation in deep learning, particularly in the context of multi-feature fusion for remote sensing scenarios.

4 Experiments Configuration

4.1 Datasets

The experiments involves the utilization of the well-known Potsdam and Vaihingen datasets provided by ISPRS for training and validation purposes. These datasets are commonly considered for urban classification and 3D building reconstruction projects. They incorporate a digital terrain model (DSM) derived from high-resolution orthophotos and dense image-matching technology. Both dataset areas encompass urban environments, with Vaihingen representing a compact village featuring numerous standalone structures and small multi-story buildings, while Potsdam epitomizes a historical city characterized by grand edifices, narrow thoroughfares, and a densely populated layout. Each dataset has undergone manual categorization into the five most prevalent land cover classes. To evaluate the efficacy of our AMBNet, the DSM data from both the Potsdam and Vaihingen datasets is chosen as comparative experiments. More specifically, we divided the Vaihingen dataset into the training and test sets as follows:

- Training set: 1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, 37;
- test set : 5, 21, 15, 30.

Furthermore, the Potsdam dataset was divided into the training and test sets as follows:

- Training set: 6_10, 7_10, 2_12, 3_11, 2_10, 7_8, 5_10, 3_12, 5_12, 7_11, 7_9, 6_9, 7_7, 4_12, 6_8, 6_12, 6_7, 4_11;
- Test set: 2_11, 3_10, 4_10, 5_11, 6_11, 7_12.

4.2 Segmentation Evaluation Method

Overall accuracy (OA), mean MIoU, and F1 Score are used for semantic segmentation accuracy analysis. We first define “Precision” as the proportion of all samples with positive predictive values and positive real values whereas “Recall” the proportion of samples with positive predicted values among all samples with positive real values:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{FP}_c + \text{TP}_c}, \tag{19}$$

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{FN}_c + \text{TP}_c}, \tag{20}$$

where TP_c, FP_c and FN_c denote true positives, false positives and false negatives for the c-th class, respectively.

Subsequently, OA, IoU and F1 Score can be defined as follows:

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{FN} + \text{FP} + \text{TP} + \text{TN}}, \tag{21}$$

$$\text{IoU}_c = \frac{1}{C}\sum_{c=1}^{C} \frac{\text{TP}_c}{\text{FN}_c + \text{FP}_c + \text{TP}_c}, \tag{22}$$

$$\text{F1}_c = \frac{2 * \text{Precision}_c * \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{23}$$

where TN_c is the true negatives for the c-th class. mIoU and mF1 Score are the mean of IoU and F1 Score of the five main categories, respectively.

5 Results and Discussions

5.1 Fusion Example Derived from Laplacian Pyramid

Figure 7 illustrates an example derived from the Laplacian pyramid fusion in the proposed FMB. As shown, the DSM and EDSM (depth estimation) information is input into the Laplacian pyramid. The input is then represented in multi-scale space before image fusion is performed between adjacent layers. Finally, the fused image is generated by summing the output from each level of the Laplacian pyramid.

5.2 Depth Estimation Example

The proposed AMBNet consists of two parts, DFEM and PMSS. The former estimates the height and produces the BDSM while the latter performs the semantic segmentation by exploiting optical and BDSM. In this section, we will demonstrate and analyze the performance of depth estimation and semantic segmentation.

In DFEB, the height estimator can effectively explore the detailed information from the optical images before the feature merge balancer improves the quality of DSM. Figure 8 shows the results of DFEB where two red boxes are added to highlight the differences. It is observed from the right box that the car elevation information is missing in the original DSM information. This makes it difficult to recognize the small vehicles on the road, and subsequently challenging to semantically segment the vehicles. In sharp contrast, the vehicle information is correctly captured by the height estimator as shown in Figure 8(c). Furthermore, the left box represents the height information of the building. It is observed that the DSM information is very irregular at the boundary of the building. After exploiting the optical information, the EDSM at the same location became more regular. In conclusion, the height information is substantially enhanced after the Laplacian pyramid and weighted fusion being applied, which is evidenced from the fact that the BDSM in Figure 8(d) has more details than DSM and is smoother than EDSM.

Figure 9 presents the three-dimensional visualization, showcasing the negative influence that terrain has made on DSM information as well as the improvement on the DSM quality provided by the proposed AMBNet. Note that the DSM values of ground objects on the earth surface are heavily affected by the terrain. As a result, if a single image for semantic segmentation contains a large ground area range, it is possible for ground objects of lower

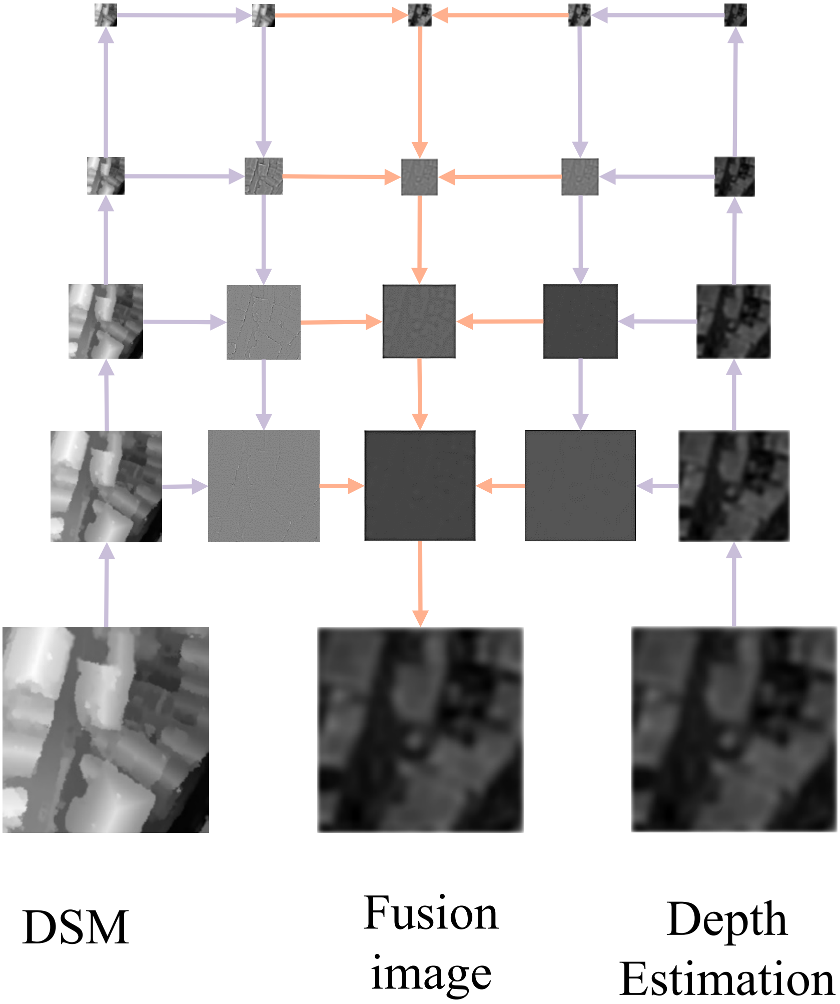

Figure 7: An output example of the Laplace pyramid fusion.

relative heights to possess higher DSM values. For instance, buildings and roads may have similar DSM values if the road is on a hillside, which can pose a major challenge to multimodal tasks especially when the ground objects exhibit similar colors. Furthermore, some trees may have low height information, which may result in mis-classification as low vegetation. In summary, inaccurate DSM information may incur poor segmentation performance. In this work, the proposed DFEB improves the DSM accuracy by exploiting the

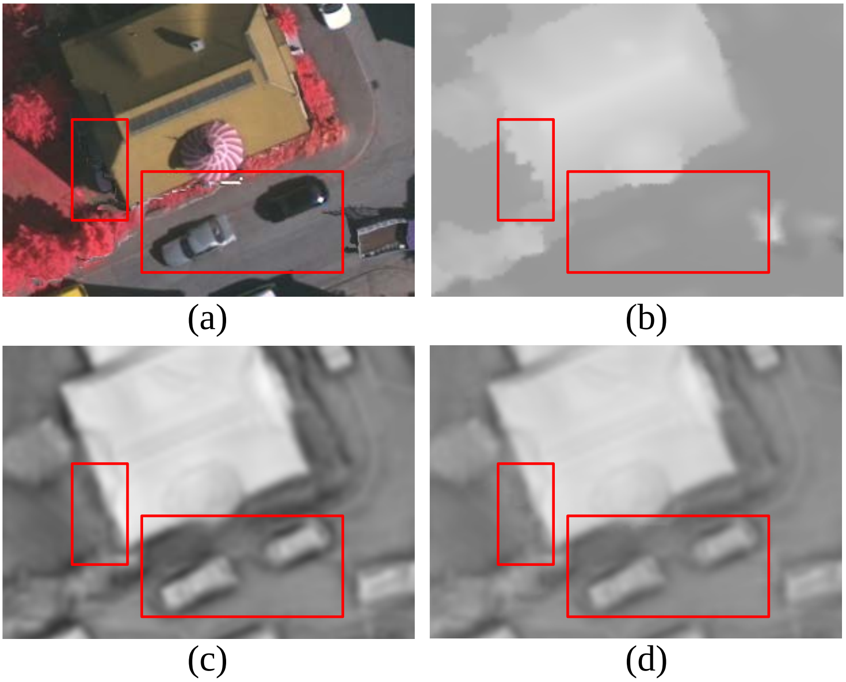

Figure 8: The EDSM and BDSM generated by DFEB. (a) optical images; (b) DSM; (c) EDSM; (d) BDSM.

ground optical information as compensation. This is evidenced by comparing Figure 9(b) and Figure 9(d) in which the optimization effect of DFEB can be observed. More specifically, in the red box of Figure 9(b), a higher DSM value can be observed, which may be due to the terrain or one of the taller trees. However, in the ground truth, we can observe that most of the ground objects in the red box area are trees. As a result, inaccurate DSM information may mislead segmentation models. To cope with this problem, the proposed DFEB provides improved BDSM as indicated in the areas highlighted by the red box in Figure 9(d), which helps improve the semantic segmentation performance.

Finally, Figure 10 compares the results before and after the LBP processing to exemplify the differences between DSM and BDSM. Inspection of Figure 10 indicates that BDSM derived by the proposed DFEB more closely resembled the raw optical image as compared to the post-LBP DSM.

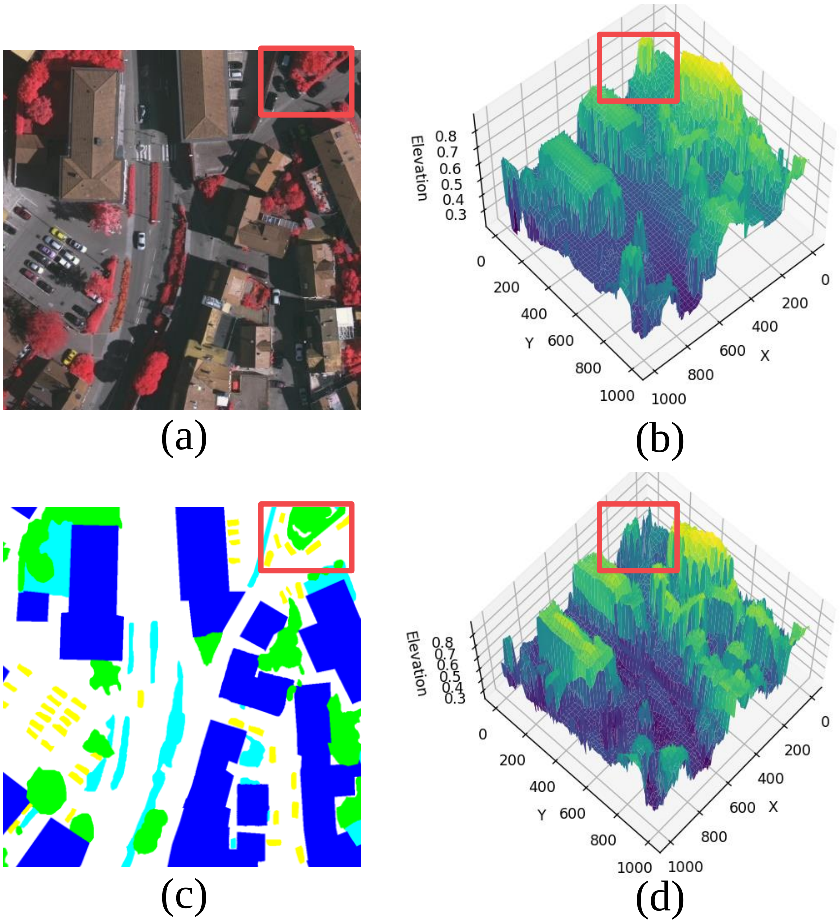

Figure 9: The effect of altitude on DSM. (a) optical images; (b) DSM; (c) Ground Truth; (d) BDSM.

5.3 Semantic Segmentation Result Analysis

5.3.1 Single-modal models versus AMBNet

We benchmark the proposed AMBNet against four state-of-the-art models, namely DABNet, ABCNet, MAResUnet and UNetformer on Potsdam and Vaihingen datasets. The experimental results on the Vaihingen dataset are summarized in Table 1 and visualized in Figure 11. It is evidenced from Table 1 and Figure 11 that the proposed AMBNet achieved the best segmentation performance for most categories, which confirmed the effectiveness of the proposed height estimation and multimodal fusion method. In particular,

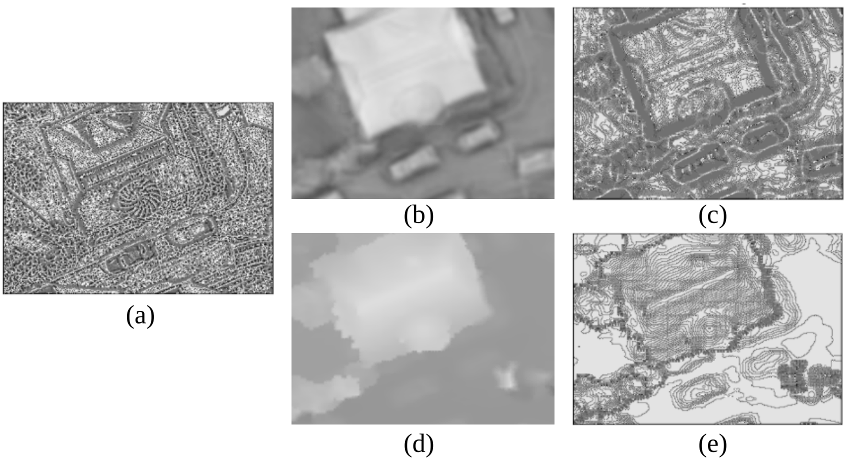

Figure 10: LBP similarity comparison. (a) IRRG images after LBP; (b) BDSM; (c) BDSM after LBP; (d) DSM; (e) DSM after LBP.

Table 1: Segmentation results on the Vaihingen dataset (%).

Method	road	building	low veg.	tree	car	OA	MIoU	mF1
DABNet	92.4	94.97	80.7	90.99	85.66	90.83	78.56	87.61
ABCNet	92.38	97.17	**85.88**	84.57	88.62	89.26	82.04	89.94
MAResUnet	90.84	95.79	78.1	91.0	81.81	90.2	79.61	88.36
UNetformer	91.53	95.93	77.64	90.86	79.71	90.3	79.4	88.23
AMBNet	**92.93**	**97.76**	82.61	**91.18**	**90.1**	**92.22**	**83.41**	**90.72**

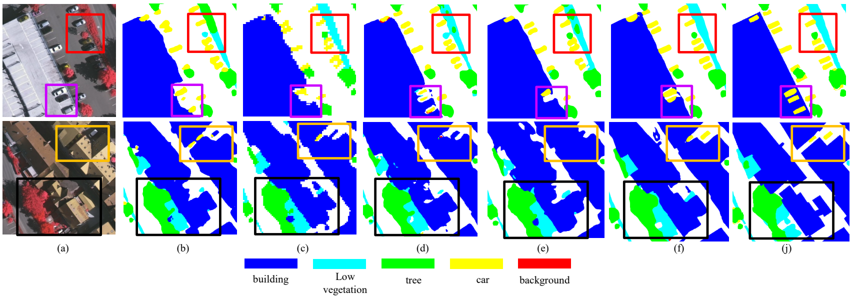

Figure 11: Visual Segmentation Results on the Vaihingen dataset. (a) optical images; (b) DABNet; (c) ABCNet; (d) MAResUnet; (e) UNetformer; (f) AMBNet; (g) Ground Truth.

AMBNet provided the most significant improvement on the *car* class with an increase of 4.4% as compared to the existing method DABNet. This can be explained by the fact that the increased details of BDSM can more effectively distinguish *car* from *road*. Furthermore, the classification accuracy for *road* and *building* has been improved by 2.88% and 0.13%, respectively as these two categories have the most uniform and significant elevation information. For instance, the DSM values of buildings are generally higher while those of roads are generally lower. As a result, the enhanced DSM, namely EDSM can improve these two categories. In terms of the overall performance, AMBNet excelled in terms of three key overall performance metrics, namely OA, MIoU, and mF1. Remarkably, the AMBNet achieved OA of 92.06%, MIoU of 83.33%, and mF1 of 90.67%, which stands for an increase of 1.02%, 2.46% and 1.55% respectively, as compared to the corresponding performance of other comparative methods. These results confirmed that the proposed AMBNet achieved better generalization performance by enhancing the DSM with more details.

Furthermore, the experimental results on the Potsdam dataset are summarized in Table 2 and visualized in Figure 12. In particular, Figure 12 presents two visualization examples of the segmentation results generated by all five methods with prediction differences being highlighted with rectangle boxes. The first example shown in the first row indicate is very challenging as the road and the building have similar optical characteristics. Since DABNet and UNetformer do not effectively use the DSM information, the road was misclassified as a building. Furthermore, the segmentation performance on *road* by ABCNet and MAResUnt was not satisfactory. In contrast, the proposed AMBNet achieved much better segmentation results with smoother borders and fewer impurities compared with other methods.

Table 2: Segmentation results on the Potsdam dataset (%).

Method	road	building	low veg.	tree	car	OA	MIoU	mF1
DABNet	90.22	97.17	81.84	87.18	86.60	91.04	80.87	89.12
ABCNet	68.33	91.20	67.57	94.37	60.50	83.80	68.30	80.81
MAResUnet	87.94	90.30	77.53	91.06	83.49	90.12	79.84	88.52
UNetformer	87.07	90.98	**82.85**	86.71	82.58	87.42	76.44	86.39
Proposed method	**93.10**	**97.30**	81.86	**91.25**	**91.00**	**92.06**	**83.33**	**90.67**

The second example shown in the second row in Figure 11 is another challenging semantic segmentation case. Some of the leaves of the trees in this area have fallen, which caused severe interference. Meanwhile, the buildings exhibit almost identical colors as the road while being partially covered by trees. It can be seen from Figure 11 that only the proposed AMBNet produced good semantic segmentation of the buildings under trees. In contrast, other methods could not accurately distinguish *tree* and *building*. Furthermore,

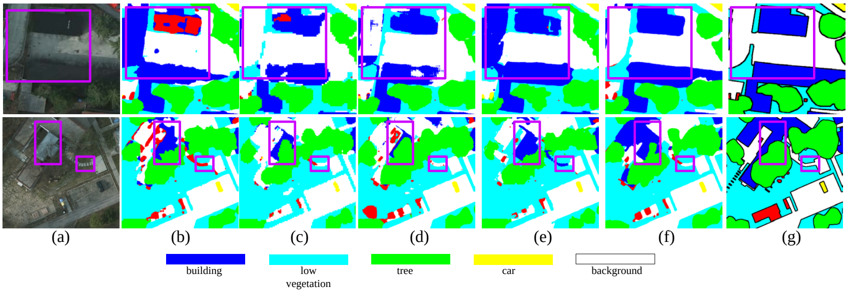

Figure 12: Visual Segmentation Results on the Potsdam dataset. (a) optical images; (b) DABNet; (c) ABCNet; (d) MAResUnet; (e) UNetformer; (f) AMBNet; (g) Ground Truth.

another box shows a path through the grass. In this area, despite that the colors of *road* or *low vegetation* are clearly distinguishable, their DSM values are similar. It is observed that with the help of BDSM, AMBNet achieved the best segmentation of the narrow path in the middle of grass. Table 2 also confirms AMBNet's excellent semantic segmentation performance in categories such as car, low veg, tree, building and road. In summary, the proposed AMBNet demonstrated improved performance by exploiting enhanced DSM as compared to other existing semantic segmentation methods.

5.3.2 Existing multi-modal models versus AMBNet

To demonstrate the efficacy of our multi-modal AMBNet, we conducted a comparative analysis with other multi-modal semantic segmentation models, namely CMGFNet, ESANet, FUSENet, and CMFNet. Figure 13 presents the test results on the Vaihingen dataset. Notably, distinguishing between rooftop parking lots and roads poses a significant challenge due to their similar colors and structural features, as illustrated in the purple box where cars were parked on the roof. This segmentation and recognition difficulty is particularly pronounced in most models, with only AMBNet exhibiting commendable results, outperforming others that struggle with low recognition rates. Further examination of the results reveals that AMBNet excelled in recognizing vehicles, trees, and shrubs. As detailed in Table 3, AMBNet attained exceptional accuracy across all categories, boasting an Overall Accuracy (OA) of 92.06% and a mean Intersection over Union (MIoU) of 83.33%. A higher MIoU indicates clearer object outline edges.

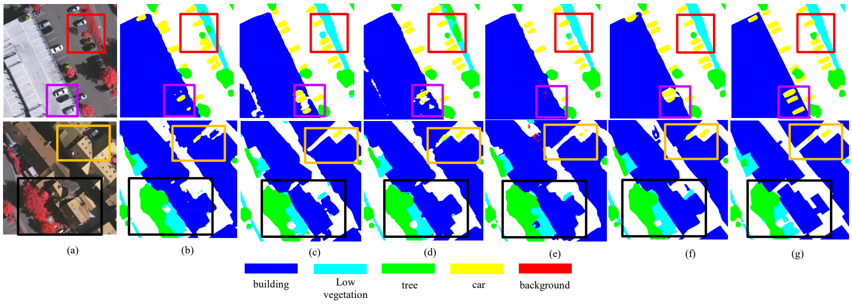

Figure 13: Visual Segmentation Results on the Vaihingen dataset. (a) optical images; (b) CMGFNet; (c) ESANet; (d) CMFNet; (e) FUSENet; (f) AMBNet; (g) Ground Truth.

Table 3: Segmentation results on the Vaihingen dataset (cross modals vs AMBNet) (%).

Method	road	building	low veg.	tree	car	OA	MIoU	mF1
CMFNet	92.98	95.28	79.6	91.26	85.95	90.97	81.54	89.6
ESANet	92.83	94.23	78.7	91.39	84.98	90.48	78.69	87.75
CMFNet	92.98	95.28	79.6	91.26	85.95	90.97	81.54	89.6
Proposed method	**93.10**	**97.30**	81.86	**91.25**	**91.00**	**92.06**	**83.33**	**90.67**

5.3.3 Abalation experiments of AMBNet

This section reports the ablation experiments conducted on the AMBNet model, featuring three essential components, namely DFEB, FEB, and PMSS. The objective of these experiments is to scrutinize the individual contributions of each structure to the model's performance, providing insights into the functioning of the model.

Figure 14 depicts the results of the ablation experiment, revealing the distinct significance of each module within AMBNet. In Figure 14 (b), the impact of removing DFEB is evident, leading to confusion in the recognition of vehicles within the purple box and irregular outlines for buildings in the black box. In Figure 14 (c), the consequences of removing FMB are highlighted. Without FMB's feature selection for fusion, the absence of the channel attention mechanism may result in the model confusing roofs with roads, manifesting as large areas of white space (representing roads) in the purple box. Table 4 presents the data from the ablation experiment. It is notable that upon deleting DFEB, FMB, and PMSS, the accuracy was reduced to 91.45%, 90.64%, and 90.84%, respectively. While deleting a specific module may enhance accuracy in a particular category, this improvement comes at the cost of reduced accuracy in other categories, indicating an imbalance in the model's recognition capabilities.

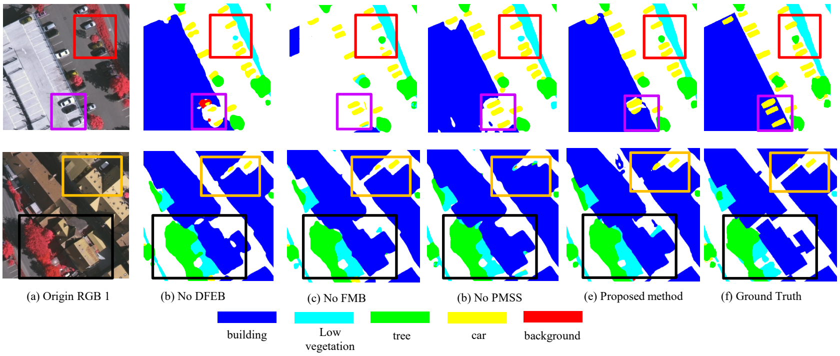
(a) Origin RGB 1
(b) No DFEB
(c) No FMB
(b) No PMSS
(e) Proposed method
(f) Ground Truth

Figure 14: Visual Segmentation Results on the Vaihingen dataset. (a) optical images; (b) No DFEB; (c) No FMB; (d) No PMSS; (e) AMBNet; (f) Ground Truth.

Table 4: Segmentation results on the Vaihingen dataset (ablation experiments) (%).

Method	road	building	low veg.	tree	car	OA	MIoU	mF1
NO DFEB	91.88	97.02	80.08	91.04	83.19	91.45	81.99	89.81
NO FMB	94.65	93.55	**83.4**	87.35	87.86	90.64	81.41	89.56
NO PMSS	**95.24**	94.99	76.39	90.48	89.88	90.84	81.26	87.65
Proposed method	93.10	**97.30**	81.86	**91.25**	**91.00**	**92.06**	**83.33**	**90.67**

The preceding ablation experiments comprehensively illustrated the significance and individual contributions of each module. Notably, FMB and DFEB emerged as particularly crucial components, with FMB playing a pivotal role in the effective fusion of channel multi-dimensional features, and DFEB contributing significantly to the optimization of DSM information.

6 Conclusion

In this work, an Adaptive Multi-feature Balanced Network named AMBNet has been proposed to perform depth estimation and multi-feature fusion for remote sensing images. Specifically, the proposed AMBNet utilizes a Depth Feature Extraction Module (DFEB) to accurately estimate ground object heights, leading to the creation of a more precise Digital Surface Model (DSM) referred to as BDSM. This enhanced DSM mitigates the adverse effects of terrains on raw DSM data, providing improved and reliable multimodal auxiliary information. Additionally, AMBNet incorporates a Parallel Multi-Stage Network (PMSS) to harness the combined power of BDSM and optical images for semantic segmentation. The results demonstrate that AMBNet excels in handling building shadows and detecting smaller ground targets hidden by canopies, showcasing commendable performance in semantic segmentation. These findings are substantiated through comprehensive experiments conducted on the Vaihingen and Potsdam datasets.

There are several extensions of this study that can be further explored. First, it is of great practical interest to further improve the height estimation performance to generate more accurate EDSM. Furthermore, it is interesting to consider how to efficiently utilize the EDSM information as the monocular depth estimation module produces high-dimensional output of multiple channels. Finally, end-to-end designs of semantic segmentation and other downstream tasks will be explored in future research.

References

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing", *RCA engineer*, 29(6), 1984, 33–41.

[2] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks", *ISPRS journal of photogrammetry and remote sensing*, 140, 2018, 20–32.

[3] J. Chen, Y. Zhao, C. Meng, and Y. Liu, "Multi-feature aggregation for semantic segmentation of an urban scene point cloud", *Remote Sensing*, 14(20), 2022, 5134.

[4] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation", *arXiv preprint arXiv:2102.04306*, 2021.

[5] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data", *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 2020, 94–114.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*, 2020.

[7] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art", *IEEE Geoscience and Remote Sensing Magazine*, 7(1), 2019, 6–39.

[8] F. G. Hall, Y. E. Shimabukuro, and K. F. Huemmrich, "Remote sensing of forest biophysical structure using mixture decomposition and geometric reflectance models", *Ecological applications*, 5(4), 1995, 993–1013.

[9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based cnn architecture", in *Asian conference on computer vision*, Springer, 2016, 213–28.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–8.

[11] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder-decoder networks for classification of hyperspectral and LiDAR data", *IEEE Geoscience and Remote Sensing Letters*, 2020.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 7132–41.

[13] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems", *IEEE transactions on automatic control*, 45(5), 2000, 910–27.

[14] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning", *ISPRS journal of photogrammetry and remote sensing*, 145, 2018, 60–77.

[15] H. Kim, S.-M. Choi, C.-S. Kim, and Y. J. Koh, "Representative color transform for image enhancement", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, 4459–68.

[16] T. Kim, H. Lee, H. Son, and S. Lee, "SF-CNN: a fast compression artifacts removal via spatial-to-frequency convolutional neural networks", in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, 3606–10.

[17] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images", *IEEE Geoscience and Remote Sensing Letters*, 19, 2021, 1–5.

[18] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery", *ISPRS journal of photogrammetry and remote sensing*, 181, 2021, 84–98.

[19] X. Ma, X. Zhang, and M.-O. Pun, "A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 2022, 3463–74.

[20] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "DSSNet: A simple dilated semantic segmentation network for hyperspectral imagery classification", *IEEE Geoscience and Remote Sensing Letters*, 17(11), 2020, 1968–72.

[21] X. Pan, L. Gao, A. Marinoni, B. Zhang, F. Yang, and P. Gamba, "Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network", *Remote Sensing*, 10(5), 2018, 743.

[22] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer”, *IEEE transactions on pattern analysis and machine intelligence*, 44(3), 2020, 1623–37.

[23] W. Sun and R. Wang, “Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM”, *IEEE Geoscience and Remote Sensing Letters*, 15(3), 2018, 474–8.

[24] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, “Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data”, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2021, 1–18.

[25] C. Tao, Y. Meng, J. Li, B. Yang, F. Hu, Y. Li, C. Cui, and W. Zhang, “MSNet: multispectral semantic segmentation network for remote sensing images”, *GIScience & Remote Sensing*, 59(1), 2022, 1177–98.

[26] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, “UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery”, *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 2022, 196–214.

[27] W. Wang and F. Chang, “A Multi-focus Image Fusion Method Based on Laplacian Pyramid.”, *J. Comput.*, 6(12), 2011, 2559–66.

[28] H. Xu, W. He, L. Zhang, and H. Zhang, “Unsupervised spectral–spatial semantic feature learning for hyperspectral image classification”, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022, 1–14.

[29] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, “Multisource remote sensing data classification based on convolutional neural network”, *IEEE Transactions on Geoscience and Remote Sensing*, 56(2), 2017, 937–49.

[30] S. Zhou, Y. Feng, S. Li, D. Zheng, F. Fang, Y. Liu, and B. Wan, “DSM-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery”, *IEEE Transactions on Geoscience and Remote Sensing*, 2023.