Category-Wise Fusion and Enhancement Learning for Multimodal Remote Sensing Image Semantic Segmentation

Aihua Zheng, Jinbo He, Ming Wang, Chenglong Li, and Bin Luo, *Senior Member, IEEE*

***Abstract*— This article presents a simple yet effective method called Category-wise Fusion and Enhancement learning (CaFE), which leverages the category priors to achieve effective feature fusion and imbalance learning, for multimodal remote sensing image semantic segmentation. In particular, we disentangle the feature fusion process via the categories to achieve the category-wise fusion based on the fact that the feature fusion in the same category regions tends to have similar characteristics. The disentangled fusion would also increase the fusion capacity with a small number of parameters while reducing the dependence on large-scale training data. For the sample imbalance problem, we design a simple yet effective category-wise enhancement learning scheme. In particular, we assign the weight for each category region based on the proportion of samples in this region over the whole image. By this way, the learning algorithm would focus more on the regions with smaller proportion. Note that both category-wise feature fusion and imbalance learning are only performed in the training stage, and the segmentation efficiency is thus not affected. Experimental results on two benchmark datasets demonstrate the effectiveness of our CaFE against other state-of-the-art methods.**

***Index Terms*— Category-wise enhancement learning (CEL), category-wise fusion, imbalance learning, multimodal remote sensing, semantic segmentation.**

I. Introduction

HIGH spatial resolution (HSR) remote sensing images contain rich geographic objects, and the automatic identification (ID) of these objects has practical value for urban planning and monitoring. Remote sensing image semantic segmentation is a special semantic segmentation task, which aims to predict the category for each pixel in the image. It can provide semantic information and location information for the region of interest, thus has many applications in remote sensing, such as land use [1], [2], building or road extraction [3], [4], [5], [6], [7], [8], agriculture vision [9], and vehicle detection [10].

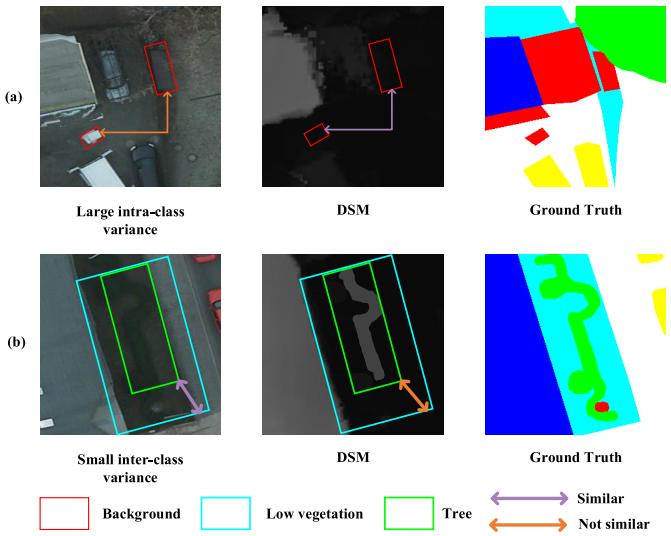

Fig. 1. Challenges in HSR imagery. (a) Large intraclass variance in the two red boxes. (b) Small interclass variance in the light blue and the red boxes.

Deep learning has made remarkable achievement on semantic segmentation of single-modal remote sensing images [11], [12], [13], [14]. However, due to the lack of rich and diverse information in HSR imagery, especially when the spectral characteristics of some foreground objects are similar in challenging scenarios, single-modal methods cannot achieve satisfactory results. Recent studies have shown that using elevation information in digital surface model (DSM) images can help the segmentation models overcome the problem of similar spectral features of foreground objects and larger intraclass variance of background. As shown in Fig. 1(a), the two red boxes are both "Background" categories but with significantly different appearance characteristics in HSR images. By contrast, the elevation information in the red boxes in the DSM image presents almost the same appearance. Meanwhile, as shown in Fig. 1(b), the light blue box and the red box with similar spectral characteristics correspond to two different categories, "Tree" and "Low veg.," respectively. By contrast, elevation information in the DSM images can better distinguish the difference.

Therefore, it is crucial to effectively integrate the information from the DSM images to improve the segmentation results. Sun and Wang [15] extract elevation information from

Manuscript received 1 August 2022; revised 9 October 2022; accepted 23 November 2022. Date of publication 1 December 2022; date of current version 14 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976002, Grant 61976003, Grant 61860206004, and Grant U20B2068. *(Corresponding author: Chenglong Li.)*

Aihua Zheng and Chenglong Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).

Jinbo He, Ming Wang, and Bin Luo are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: hjinb977@163.com; wangming@stu.ahu.edu.cn; luobin@ahu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3225843

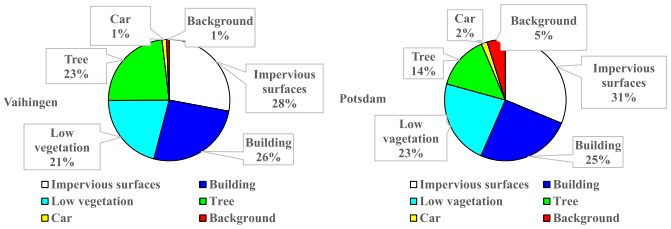

Fig. 2. Sample imbalance in HSR imagery. Compared with other categories, "Car" and "Background" account for a relatively small proportion.

DSM images to modify segmentation results of red, green, blue (RGB) optical images. However, the backend-based processing cannot be trained end-to-end, and thus difficult to converge. Cao et al. [16] propose to use a lightweight feature extractor [DSM fusion (DSMF)] to extract features from DSM images and investigate four fusion strategies. DSMF contains only a small number of parameters and can be flexibly applied to other semantic segmentation networks. However, the DSMF network structure is simple, the features in the DSM images cannot be fully extracted. Audebert et al. [17] verify that integrating multimodal information based on residual correction strategy improves semantic segmentation by allowing the network to learn robust multimodal features. However, the methods mentioned above simply sum the feature maps of two modalities and concatenate them in the channel direction while ignoring the fact that the feature fusion based on the same category region often has similar features.

To effectively integrate the DSM information to better solve the problem of similar spectral features of foreground objects and larger intraclass variance of background, we propose a category-wise feature fusion (CFF) module to enhance the feature fusion of the same category and achieve category-level feature fusion in this article. Based on the fact that the same category regions tend to have similar characteristics, we propose to use category mask to enhance feature fusion of the same category in both modalities. We first derive the binary mask of the different categories from the label map. The value "1" on the corresponding category mask selectively aggregates the feature of the same category in different channels from the two modalities. Meanwhile, the value "0" suppresses features on different channels that do not belong to the category mask. Then, the feature maps aggregated by each category mask on the two modalities perform element-wise summation after concatenation on the channel. Finally, the channel attention (CA) fuses the features on different channels to obtain a more salient feature representation.

In addition, the sample imbalance problem hinders the performance of segmentation in remote sensing images. As shown in Fig. 2, the "Car" and "Background" categories account for a small proportion compared with other categories. Faced with the challenge of small foreground objects in remote sensing datasets and sample imbalance problem, Zheng et al. [11] propose a foreground perception network using the scene vector to combine the object context with improving the characteristics of the foreground object in the instance segmentation in aerial images dataset (isaid) [18] dataset. This method proposes the foreground perception network which consists of a 1-D scene vector. However, it ignores the semantic gap when modeling the context of the feature maps in different layers of the encoder. Ma et al. [12] propose a foreground activation network branch that improves the characteristics of foreground objects. However, the structure of the dual decoder under collaborative probability loss cannot converge well during optimization.

To extract more discriminative features for the samples with a small proportion to overcome the sample imbalance problem in remote sensing images, we propose a category-wise enhancement learning (CEL) module to enhance the network learning ability, which enforces the network to focus on the categories features with a small proportion. We use prior knowledge of the category distribution to enhance network training, enforcing the network to focus on categories with a small proportion. We first derive the weight factor of the category from the label map and embed this prior knowledge to help the network learn the distribution information of the image. The categories with a small proportion are assigned with small weights, by this way, the network pays more attention to learning this region under the optimization of the focal loss [19].

Pyramid structure has been noted as a practical scheme to alleviate the gradient vanishing problem and improve feature extraction capability, which has been widely used in related computer vision and machine learning tasks, including Rednet [20], Deeplabv3 [21], and pyramid scene parsing network (PSPNet) [22]. To preserve the multiscale feature of objects in the multilayer decoder, we propose to employ the pyramid loss supervision (PLS) to optimize the output of each layer of the decoder.

As summary, we propose a novel network framework called category-wise fusion and enhancement learning (CaFE) based on the pyramid structure for multimodal remote sensing image semantic segmentation in this article. Specifically, we propose a CFF module to improve feature fusion, and a CEL module strengthens model learning capabilities. The results on two benchmark multimodal remote sensing datasets prove the superiority of our method over other state-of-the-art semantic segmentation methods. The main contributions in this article are as follows.

1) We propose a CaFE framework for multimodal remote sensing image semantic segmentation. It effectively solves the problems of feature fusion and sample imbalance in multimodal remote sensing image semantic segmentation from a new perspective.
2) We propose a CFF module, which uses the category mask in the label map to achieve the category-wise fusion while suppressing redundant feature.
3) We propose a CEL module to overcome the sample imbalance problem, which uses prior knowledge of the distribution of different categories in the input image to guide the network to learn more effective category features.

II. Related Work

A. Semantic Segmentation of Single-Modal Remote Sensing Images

Several earlier works [23], [24], [25] focus on using multilevel semantic features on local patterns of images.

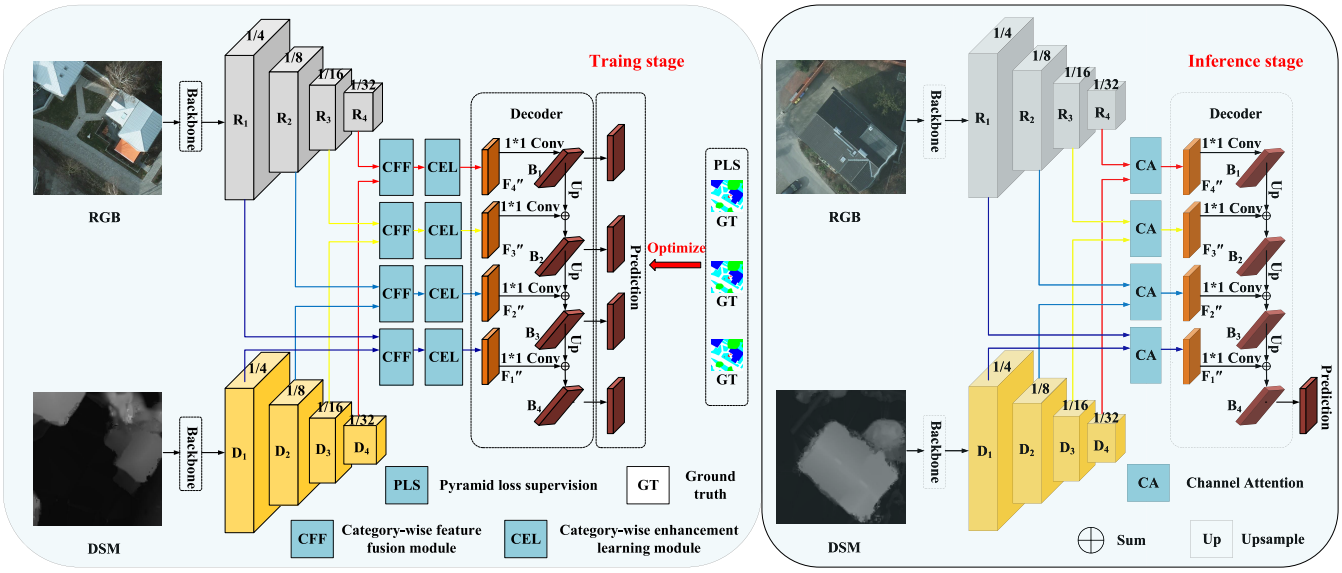

Fig. 3. Overview of the proposed CaFE framework, which includes training phase and inference phase. The training stage contains four main parts: CFF, CEL, decoder, PLS. The inference stage mainly contains the CA module.

Meanwhile, earlier studies mainly rely on low-resolution and medium-resolution datasets, such as [26], [27]. With the development of remote sensing technology and the advancement of measurement technology, a large number of high-resolution remote sensing images can be obtained from airborne and spaceborne platforms. Such as Gaofen image dataset (GID) [28], DeepGlobe [29], iSAID [18], unmanned aerial vehicle video dataset (UAVid) [30], Zurich Summer [31], SpaceNet [32], and WHU building change detection dataset (WHU) [6]. For the challenges of different datasets, different researchers propose different solutions. On the iSAID dataset, foreground objects only occupy 1.34% of the pixels. This increases the difficulty of extracting foreground object features and brings the sample imbalance problem. Some works [11], [12] address such challenges by modeling foregrounds and optimizing network training using loss functions. On the GID dataset, Li et al. [33] design multiscale skip connections and CA blocks to fuse multiscale features. On the WHU Building dataset [6], Xiang et al. [34] propose adaptive feature selection (AFS), which uses CA to select important feature. Recently the transformer has made great achievement in the field of remote sensing. Wang et al. [35] use Swin Transformer [36] as the backbone network to extract image feature, and then propose densely connected feature aggregation module (DCFAM) acts as a decoder to restore the resolution of the image. He et al. [37] embed the Swin transformer into the classical convolutional neural network (CNN)-based UNet. Meanwhile they propose a spatial interaction module (SIM) to focus on the pixel-level feature correlation in the spatial dimension, a feature compression module (FCM) to alleviate the omission of small-scale features during patch token downsampling, and a relational aggregation module (RAM) to integrate global dependencies from the Swin transformer into the features from CNN hierarchically.

Attention mechanism has also achieved certain results in the field of remote sensing. Li et al. [13] propose a novel and efficient module for point-wised affinity learning to handle dense affinity learning forces small objects to absorb noisy context. Mou et al. [38] propose a spatial relation module and a channel relation module to capture long-range spatial relationships between entities. Li et al. [39] propose a multihead attention module to capture the global context information thoroughly. Li et al. [40] employs a new kernel attention mechanism with linear complexity to solve the traditional computationally demanding problems of attention. To relieve the sample imbalance problem in remote sensing images, Zhou et al. [41] use dice loss to improve road regions weight in road extraction. Kellenberger et al. [42] utilize the online hard example mining (OHEM) strategy in the task of animal detection in remote sensing images. However, the ability of identifying objects remains limited due to the lack of rich and diverse information in single-modal HSR remote sensing images, particularly in challenging scenes where the spectral characteristics of foreground objects are similar.

B. Semantic Segmentation of Multimodal Remote Sensing Images

Due to its richer scene characterization, recent progress has been made in the semantic segmentation of multimodal remote sensing images. Hong et al. [43] propose joint exploitation of multiple modalities which characterizes the scene at a more detailed and precise level, meanwhile investigate five plug-and-play fusion modules. Recently, multimodal remote sensing image classification models have been roughly divided into two categories. One category inputs data as a local patch around its center pixel [44], [45]. However, the block operation will lose the spatial and context information. The other category aims to assign a semantic category to each pixel

based on CNN. Audebert et al. [46] introduce a multikernel convolutional layer that operates several parallel convolutions with different kernel sizes to aggregate predictions at multiple scales to segment RGB and DSM images. Volpi and Tuia [47] proves the use of spatial regularization can help simplify class structures spatially for semantic segmentation in aerial images. Srivastava et al. [48] propose a multimodal deep learning solution to enhance the understanding of urban land use from both overhead and ground images. Another interesting work [49] utilizes an optical branch to learn global relationships of any two regions and an elevation feature branch to learn boundary information. However, the above methods are difficult to segment objects with similar spectral characteristics. The main reason is that most of the above methods conduct global fusion at the feature level, which will bring redundant features.

III. CaFE Framework

The proposed CaFE framework uses an end-to-end encoder and decoder structure, consisting of training and inference parts. As shown in Fig. 3, CaFE mainly includes a feature encoder, feature decoder, CFF module, CEL module, and PLS during the training stage. First, we use a feature encoder to extract the features of the RGB and DSM images. Then, the extracted features will flow into the CFF, which uses the category mask to aggregate features of the same category. The feature maps after CA output strengthens the feature fusion and obtains a more salient feature representation. Considering the sample imbalance problem, we use CEL to guide the network training and let the network learn more effective category features. At the same time, we optimize the predictions for each decoder using a PLS based on the focal loss [19] to improve the feature extraction ability of objects at different scales. In the inference stage, only the CA mechanism in the CFF is retained.

A. CFF Module

The traditional feature fusion method ignores the modality differences and cannot ensure operational effectiveness when bringing additional parameters. Note that the label map of semantic segmentation contains the location information of objects. Therefore, we utilize the masks of different categories in the label map to fuse the feature of the same category on different channels from two modalities, thus enhancing the feature representation of each category and reducing the interclass differences. The category masks are decomposed from the label map of each input network, which using the prior knowledge of the labels without additional parameters. The basic idea is shown in Fig. 4. The CFF consists of two parts: (a) category decoupling and (b) CA. The task of the category decoupling stage is to separate the features of different categories in the channel direction of the feature map according to the category mask. The main reason is that the same category region tends to have similar characteristics. CA further improves the feature fusion ability of the same category to obtain more salient feature representations. We disentangle the feature fusion via the category mask to achieve the category-wise fusion, which improves the feature fusion capability of the two modalities.

As shown in Fig. 3, first, two independent residual modules process a multispectral RGB image and a corresponding DSM image. $R_i, D_i (i \in \{1, 2, 3, 4\})$ represent the RGB features and DSM features of the layer ith encoder, respectively, with the size of 1/4, 1/8, 1/16, 1/32 of the original image. The detailed CFF is shown in Fig. 4. $M_j (j \in \{0, 1, \ldots, 5\})$, represents the category masks of different scales of label maps in different layers of encoder. For the given R_i and D_i, we perform the dot product operation with M_0 to M_5, respectively, which use the category mask of the label map to aggregate the object features that match the category mask on the channel, the corresponding position "1" retains the feature, and "0" removes redundant feature. Then, the feature maps aggregated by each category mask are concatenated in the channel direction. Finally, we perform an element-wise sum operation between processed feature R_i and the features D_i. The main idea is to selectively fuse feature of the same category on different channels in two modalities using different category masks. We formulate this procedure as follows:

$$F_{rd} = C_{\text{cat}}\left(R_i \odot \sum_{j=0}^{5} M_j\right) + C_{\text{cat}}\left(D_i \odot \sum_{j=0}^{5} M_j\right) \quad (1)$$

where $i = 1, \ldots, 4$ and C_{cat} represents for concatenation on the channel. The feature maps F_{rd} effectively integrate the same category's feature and suppress the redundant feature. Then we use CA to improve the feature interaction ability of the same region of the feature map F_{rd}.

As shown in Fig. 4(b), we apply 3×3 convolutional layers on feature F_{rd} to reduce the number of channels. The main purpose is to ensure that the input channels are consistent when testing. A 3×3 convolution then follows the $\tilde{F}$ features for refining the information, obtaining F_i. The transformed features F_i are then reshaped into a 1-D vector through global pooling. The existing work [39] first reduces the dimensionality of the aggregated features obtained after global average pooling and generates the weights for each channel. Efficient channel attention (ECA)-Net [50] proposes that avoiding dimensionality reduction is essential for learning CA. Inspired by this, when calculating the weight of the channel ω, we generate channel weights by performing a 1×1 convolution of size k as

$$\omega = \varphi_k(F_i) \quad (2)$$

where φ represents 1×1 convolution, and ω is processed by the sigmoid function. Finally, we perform a matrix multiplication between the weight ω and the feature map F_i to obtain the result F_i' as

$$F_i' = \omega \otimes F_i. \quad (3)$$

The feature map F_i' obtained by the CFF contains the rich information of the same regions in the two modalities. Meanwhile, CFF requires only a small number of parameters.

B. CEL Module

The imbalanced samples can mislead the optimization direction of the network training. The label map contains not only the location information of objects, but also the distribution information of objects. From the perspective of image distribution, the CEL assigns each pixel a weight factor derived

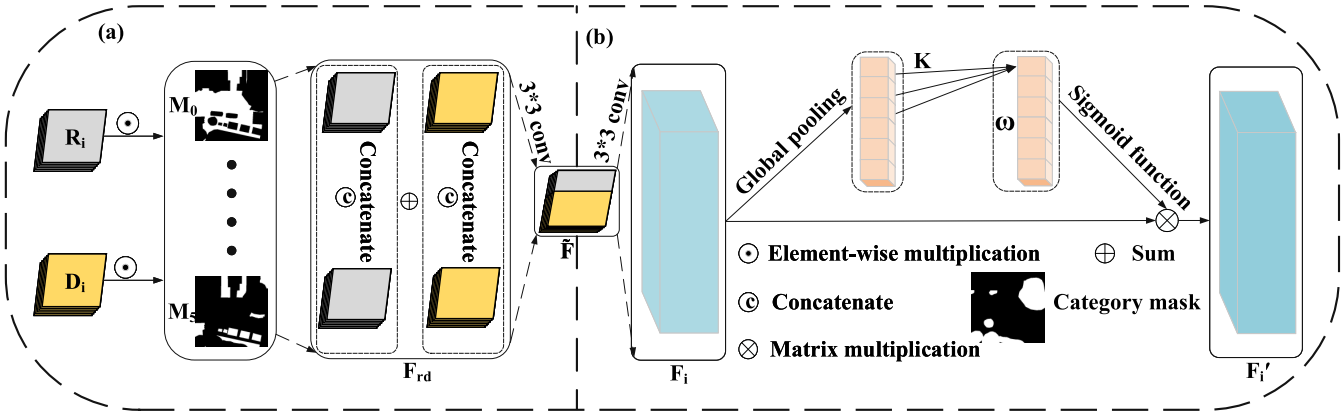

Fig. 4. Computational details of the CFF in the ith layer encoder, CFF consists of two parts: (a) category decoupling and (b) CA mechanism.

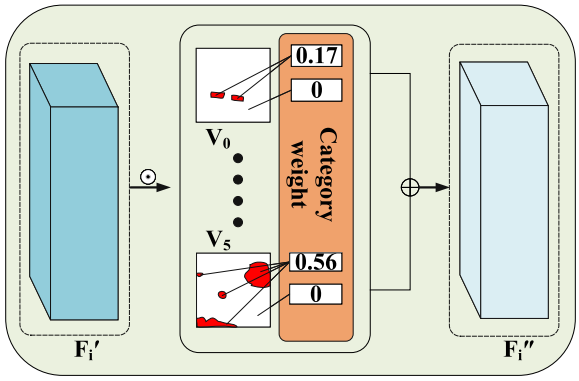

Fig. 5. Structure of CEL. The nonzero value in $V_j, j \in \{0, 1, \ldots, 5\}$ represents the proportion of the corresponding category in the image.

from the label map while helping the network learn the image distribution information to maintain the stability of network training, which to implement CEL, focusing on the categories with a small proportion.

The basic structure is shown in Fig. 5. First, we get the category masks from the label map used by the encoder at each layer and then calculate the weights of different categories in the different category masks, obtaining the weight map $\{V_0, V_1, V_2, V_3, V_4, V_5\}$. The nonzero value in the weight map V_j represents the proportion of the corresponding category on the image. Then, we perform a dot product operation between feature $F_i{}'$ and the feature V_j, obtaining features $F_i{}''$. We formulate this procedure as follows:

$$F_i{}'' = F_i{}' \odot \sum_{j=0}^{5} V_j. \tag{4}$$

When the weight of the categories with a small proportion in the label map is low, after assigning the weight to the same region on the feature map, The output corresponding to the same region on the feature map is suppressed to a certain extent, making the prediction of the corresponding region output more challenging when the model is trained. With the constrain of the focal loss [19] and the prior knowledge of images, the model will gradually learn the knowledge of the sample distribution, then dynamically change the training director of the network, thus paying more attention to this part of the region. Each pixel in the feature map F_i'' not only contains weight information but also learns the category distribution of images, effectively helping the network to accurately segment the categories with a small proportion during testing.

C. Pyramid Loss Supervision

The network utilize top-down pathway and skip connections to yield the pyramidal features maps $\{B_1, B_2, B_3, B_4\}$. The detailed architecture of the decoder is illustrated in Fig. 3.

Specifically, the deepest-layer features $F_4{}''$ is first reduced (256 by default) with a 1×1 convolution, obtaining the B_1 features. The B_1 features are then by two times upsampling. The features $F_3{}''$ from the encoder is then gathered to sum with B_1, obtaining B_2. Generally speaking, we formulate this procedure as follows:

$$B_{(h+1)} = U_{p\text{sample}_{\times 2}}(B_h) + \varphi\left(F''_{(4-h)}\right)(h = 1, 2, 3) \tag{5}$$

where φ represents 1×1 convolution. The decoder is designed to recover the spatial resolution from the encoder feature maps. We adopt simple yet effective form to capture multiscale feature. Finally, the feature maps $\{B_1, B_2, B_3, B_4\}$ are acquired.

To preserve the detailed information of objects in the output of the multilayer decoder, we use the PLS to optimize the output of each layer of the decoder, focusing on the feature changes of objects at different scales.

The feature map F_i'' outputted by the CEL is only assigns weight information to different regions of the feature map. How to use prior knowledge to guide network learning to overcome the sample imbalance problem is still an important issue. Based on the idea of focal loss [19], we use $(1 - p)^r$ to represent the weight of categories with a small proportion, r is the weighting factor, and p is the probability predicted by the model. The introduction of weight estimation allows the network to pay more attention to categories with a small proportion during training. At the same time, same as FarSeg [11], a constant Z is introduced, where Z represents the number of all pixels in the current prediction map, which avoids gradient

vanishing. The loss function of the last layer of the decoder is

$$\mathcal{L}_4 = \frac{1}{Z}\sum (1 - p_i)^r C(p_i, y_i) \tag{6}$$

where p_i represents the predicted probability, y_i represents the ground truth, and $C(p_i, y_i)$ represents the cross-entropy loss of each pixel. At the same time, to preserve the corresponding information of each layer in the decoder, the loss function optimizes network training in the decoder's second, third, and last layers. The total loss function is

$$\mathcal{L}_{\text{total}} = \sum_{q=2}^{4} \mathcal{L}_q \tag{7}$$

where $\mathcal{L}_q$ represents the loss function of the qth layer decoder.

The PLS can pay more attention to the multiscale changes of the object and addressing the sample imbalance problem. At the same time, it can also avoid the problem of gradient disappearance.

IV. Experiments

First, we will introduce the dataset and experimental details in detail. Second, we will perform extensive experiments on Potsdam and Vaihingen datasets and give a detailed analysis to verify the effectiveness of the proposed CaFE. Finally, we add the CFF and CEL to various baselines to verify the effectiveness.

A. Data Description

Potsdam: Potsdam[1] contain 38 fine-resolution images of size 6000 × 6000 pixels with a ground sampling distance (GSD) of 5 cm. The dataset provides near-infrared, RGB channels as well as DSM and normalized DSM. There are 24 images in the training set and 16 in the test set. Specifically. We utilize ID: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13 for testing, and the remaining 24 images for training. We use only the RGB channels and corresponding DSM in our experiments.

Vaihingen: Vaihingen[2] semantic labeling dataset is composed of 33 orthorectified image files mosaic with three spectral bands (red, green, near-infrared), plus a normalized DSM of the same resolution. The dataset has a spatial resolution of 9 cm, with an average size of 2494× 2494 pixels and a GSD of 5 cm. There are 16 images in the training set and 17 in the test set. We exploit ID: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38 for testing and the remaining for training. We also use the corresponding DSM in our experiments.

B. Implementation Details

The backbone used in CaFE is ResNet-50 [51] for all the experiments, which is pretrained on ImageNet [52]. The single-modal method only uses RGB data for training in the experiment. The multimodal method uses RGB and the corresponding DSM dataset. To maintain the fairness of the experiment, all methods are carried out under the same experimental conditions. Same as FactSeg [12]. The stochastic gradient descent (SGD) optimizer is adopted during the training with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate is set to 0.0007 and a "poly" schedule with a power 0.9. All models are trained for 150 epochs. We randomly crop the image into a fixed size (512, 512) from the original image. The batch size is set to 8. The k in the CA module and r in the PLS module were 4 and 2, respectively. For data augmentation, random horizontal and vertical flips are used during training. We normalize the depth values in the DSM images to between 0 and 255 during network training, and the DSM images satisfy the rule that objects close to the ground present low depth values and objects high above the ground present high depth values. When the model is tested, since the CFF and CEL have the label information, only the CA in CFF and the loss function of the last layer in the decoder are retained during the test. All methods are tested on NVIDIA GeForce RTX3090 graphics processing unit (GPU) with 24 GB RAM. To evaluate the performance of the proposed methods, we use three metrics to evaluate the performance of our method, overall accuracy (OA), mean F_1-score (mean-F_1), and mean intersection over union (mIoU).

C. Comparison Results

Results on Potsdam: As shown in Table I, after fusing the elevation information in DSM image, the state-of-the-art multimodal methods V-FuseNet [17], digital surface model fusion network (DSMFNet) [16] perform even overshadowed by the single-modal methods such as FarSeg [11] and FactSeg [12]. The main reason is that simple concatenating or summing in the channel direction ignores the distribution between modalities and may suppress another modality. By contrast, the proposed CFF overcomes the problem of modal distribution differences by selectively guiding the feature fusion of two modalities of the same category through a category mask. Therefore, our method (CaFE) significantly beats all the compared state-of-the-art methods in both the RGB and RGB + DSM scenarios in OA, mean-F_1, and mIoU metrics, especially in "Building," "Car," and "Tree." The DSM image provides the elevation information of ground objects. The categories of "Building," "Car," and "Tree" have a certain height from the ground, which present more discriminative features in the DSM image. It can be seen that our method fully excavates the feature of these categories in DSM images. Although the features of the categories "Imp. surf." and "Low veg." are not obvious in DSM images, our method learns complementary features in the two modalities and achieves promising results. Fig. 6 provides more intuitive segmentation results of each method on the Potsdam datasets. Most of the red boxes in Fig. 6 are the shadow regions and the regions with similar spectral features in the foreground. As shown in the seventh row of Fig. 6, the categories marked in the red box are "Low veg." and "Tree," respectively. In the RGB optical image, the spectral features of these two categories are not highly discriminative. However, in the DSM image, the elevation

[1] https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/

[2] https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/

TABLE I

PERFORMANCE ON THE REFERENCE METHODS AND THE PROPOSED CaFE FRAMEWORK ON THE POTSDAM DATASET

Method		F_1 per category (%)						OA (%)	mIoU (%)	mean-F_1 (%)
		Imp. surf.	Building	Low veg.	Tree	Car	Background			
RGB	FCN-8s [53]	90.29	92.45	81.66	82.83	88.65	54.80	84.45	69.49	81.78
	U-Net [54]	90.56	94.55	82.34	83.94	89.12	55.27	86.78	71.74	82.63
	Deeplabv3 [21]	90.54	92.78	84.13	86.32	89.76	56.77	87.51	74.41	83.38
	MANet [40]	90.03	93.56	**85.92**	85.34	89.15	54.29	90.34	74.28	83.04
	MACU-Net [33]	**92.34**	93.23	85.23	87.45	88.19	55.68	90.89	74.76	83.68
	RefineNet [55]	90.45	92.18	84.45	87.76	89.12	53.41	85.17	73.49	82.89
	PSPNet [22]	89.34	92.21	84.34	87.32	88.54	54.26	89.07	72.39	82.66
	FarSeg [11]	91.56	94.22	84.67	89.45	90.41	57.28	91.57	76.59	84.59
	HMANet [56]	90.23	92.89	83.87	88.12	90.08	56.84	90.32	74.66	83.67
	FactSeg [12]	90.68	94.17	85.37	89.01	91.10	58.10	91.21	77.14	84.73
	DANet [57]	90.67	92.45	83.52	84.78	88.09	57.26	88.32	72.68	82.79
RGB+DSM	FCN_MFS_DSMBackend [15]	90.34	92.35	83.39	85.34	88.14	56.24	89.34	71.70	82.63
	V-FuseNet [17]	90.91	93.23	84.30	86.23	90.56	56.64	88.49	74.74	83.64
	DSMFNet [16]	91.64	92.24	84.66	86.80	89.02	57.54	87.17	74.76	83.65
	Ours	91.23	**95.01**	85.12	**89.83**	**92.17**	**58.24**	**92.45**	**77.81**	**85.26**

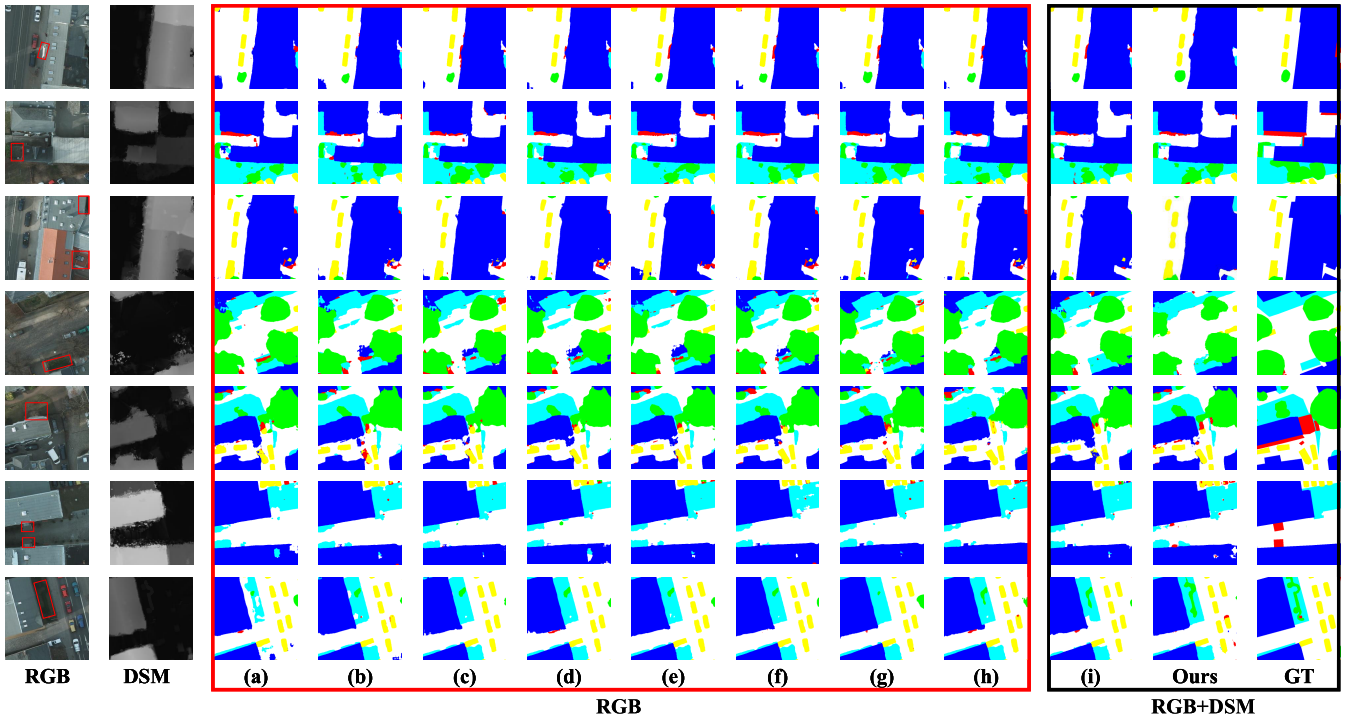

Fig. 6. Visual results of our proposed method compared with other state-of-the-art methods on the Potsdam Dataset. The white, blue, cyan, green, yellow, red, separately represent the categories of "Imp. surf.," "Building," "Low vegetation," "Tree," "Car," "Background." (a) Fully convolutional networks (FCN)-8s. (b) U-Net. (c) Deeplabv3. (d) RefineNet. (e) MANet. (f) Hybrid multiple attention network (HMANet). (g) FarSeg. (h) FactSeg. (i) V-FuseNet.

information of these two categories is more distinguishable. The single-modal method multi-scale skip connected architecture (MACU)-Net [33] and FarSeg [11] perform poorly, the main reason is that the single-modal condition lacks rich and diverse information. Their performance is limited when facing the challenge of similar spectral features of foreground objects. After using the elevation information in the DSM image, the multimodal method DSMFNet [16] has a certain improvement, but it still cannot accurately distinguish the semantic regions of these two categories. Our method decouples the process of feature fusion through category masks while enhancing features of the same category through CA to obtain more salient feature representations for each category. As shown in the shaded regions marked in red boxes in Fig. 6, our method achieves more accurate results.

Results on Vaihingen: As shown in Table II, in the Vaihingen dataset, our method (CaFE) consistently achieves the best results on OA, mIoU, and mean-F_1 metrics. Compared with the Potsdam dataset, the color information of the Vaihingen dataset is more abundant. However, the occlusion problem usually leads to semantic ambiguity. For example, the shadow area caused by the occlusion of the category of "Tree" brings certain challenges to recognizing the category of "Car" and "Imp. surf." Meanwhile, the existing multimodal methods still work poorly compared to some single-modal methods, such as FarSeg [11] and FactSeg [12], which indicates the importance of fusion strategy while integrating the DSM image information. As shown in Fig. 7, the texture difference between category "Tree" and category "Building" is relatively large. Both the single-modal methods and the multimodal methods

TABLE II
PERFORMANCE ON THE REFERENCE METHODS AND THE PROPOSED CAFE FRAMEWORK ON THE VAIHINGEN DATASET

Method		F_1 per category (%)						OA (%)	mIoU (%)	mean-F_1 (%)
		Imp. surf.	Building	Low veg.	Tree	Car	Background			
RGB	FCN-8s [53]	88.61	91.34	77.94	84.60	73.54	40.34	82.45	66.91	76.06
	U-Net [54]	87.32	90.48	78.45	85.42	73.49	42.18	82.91	67.14	76.22
	Deeplabv3 [21]	88.54	92.23	78.90	85.63	72.48	40.56	84.98	68.54	76.39
	MANet [40]	87.45	91.48	79.28	87.45	74.69	43.24	90.67	70.57	77.265
	MACU-Net [33]	**91.23**	92.12	77.89	87.78	75.26	45.48	90.94	71.21	78.29
	RefineNet [55]	87.29	90.18	78.84	86.28	75.12	47.17	88.89	70.66	77.48
	PSPNet [22]	88.40	91.74	79.20	85.90	73.5	41.06	86.09	69.48	76.63
	FarSeg [11] [11]	90.44	92.56	79.94	87.14	77.46	49.87	90.11	72.59	79.56
	FactSeg [12]	91.07	91.97	78.26	88.97	78.59	50.95	91.98	73.24	79.96
	HMANet [56]	90.48	92.3	**79.98**	88.34	78.15	51.47	90.99	73.88	80.12
	DANet [57]	88.61	91.18	78.19	86.23	77.19	49.37	87.45	71.47	78.46
RGB+DSM	FCN_MFS_DSMBackend [15]	88.87	90.48	78.34	87.14	76.82	50.64	86.21	71.62	78.71
	V-FuseNet [17]	89.14	92.14	79.18	86.34	77.47	51.67	88.89	72.01	79.32
	DSMFNet [16]	90.25	91.59	78.39	87.49	77.18	48.56	87.22	71.90	78.91
	Ours	91.19	**93.89**	79.10	**89.10**	**79.04**	**54.15**	**92.23**	**74.05**	**81.07**

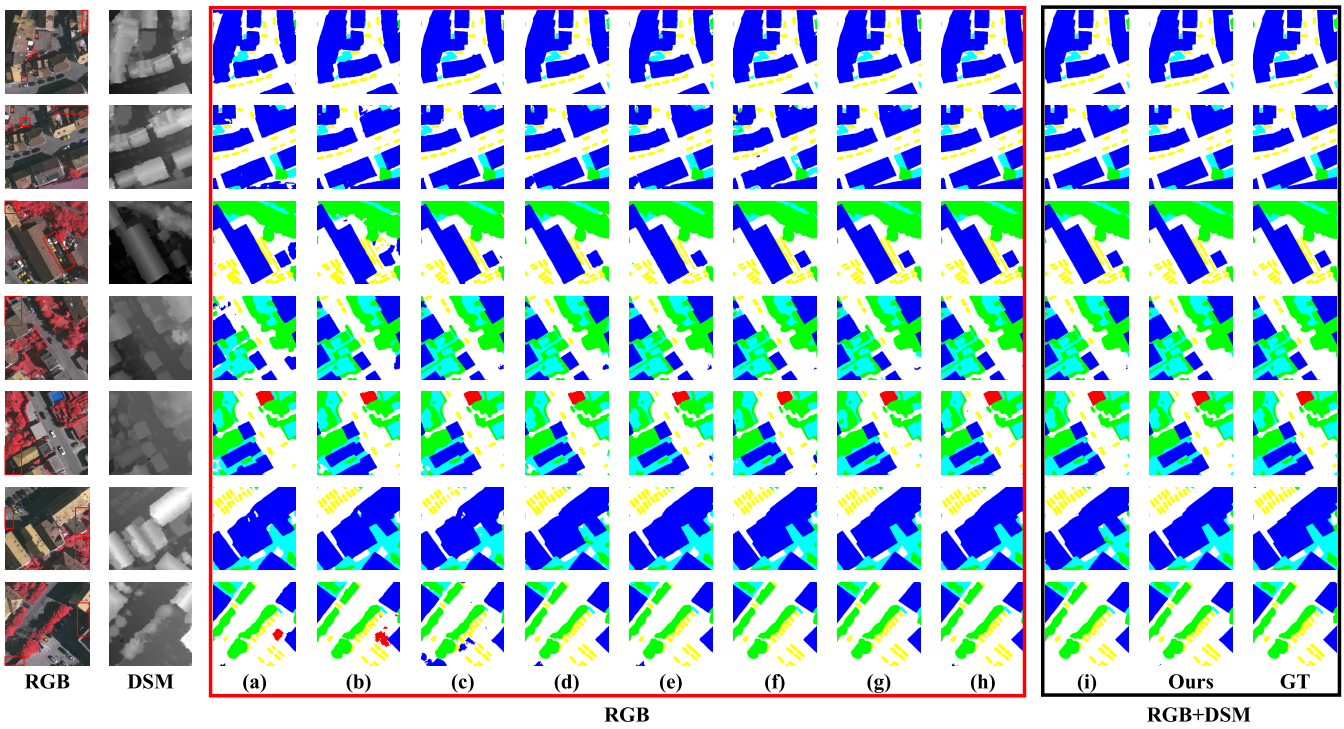

Fig. 7. Visual results of our proposed method compared with other state-of-the-art methods on the Vaihingen dataset. The white, blue, cyan, green, yellow, red, separately represent the categories of "Imp. surf.," "Building," "Low veg.," "Tree," "Car," "Background." (a) FCN-8s. (b) U-Net. (c) Deeplabv3. (d) RefineNet. (e) MANet. (f) HMANet. (g) FarSeg. (h) FactSeg. (i) V-FuseNet.

perform well. As shown in the third row of Fig. 7, due to the occlusion caused by the shadow of the category "Building," the edge of the category "Car" marked in the red box is not well segmented in the single-modal methods. By contrast, the multimodal method V-FuseNet [17] has a certain improvement after introducing the elevation information in the DSM image. Our proposed method is able to recognize the category of "Car" with finer boundaries. Furthermore, single-modal methods cannot well handle the challenging shadow area as shown in the red boxes marked in Fig. 7. After introducing the elevation information in the DSM image, our method (CaFE) segments more continuously at the object's boundary. It shows strong robustness despite the challenges of shadowed regions and similar spectral features of foreground objects.

D. Component Analysis

Ablation Study: Table III shows the ablation study on the key components, CFF, CEL, and PLS, in the proposed CaFE framework. First, by progressively introducing each component, the mIoU consistently improves, which validates the effectiveness of each component in our model. As shown in Table III(c) and (d), the mIoU has been distinctively improved after introducing CFF and CEL, respectively. This evidence the significant contribution of the proposed CFF, which uses category masks to achieve category-level feature fusion, so as

TABLE III

OBJECT SEGMENTATION mIoU (%) ON POTSDAM DATASET. STARTING FROM BASELINE, THE PROPOSED MODULES ARE GRADUALLY ADDED IN THE PROPOSED CAFE FOR THE MODULE ANALYSIS

Method	CFF	CEL	PLS	mIoU (%)
(a) Baseline	-	-	-	74.38
(b)	-	-	√	75.02
(c)	√	-	√	77.13
(d) CaFE	√	√	√	77.81

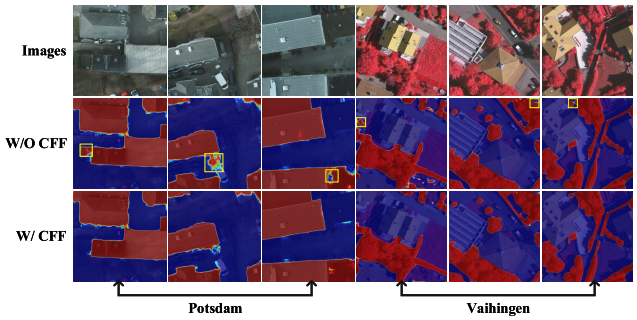

Fig. 8. Visual representation of the feature response maps on the category of "Building" and "Tree" with/without our proposed CFF in the Potsdam and Vaihingen dataset.

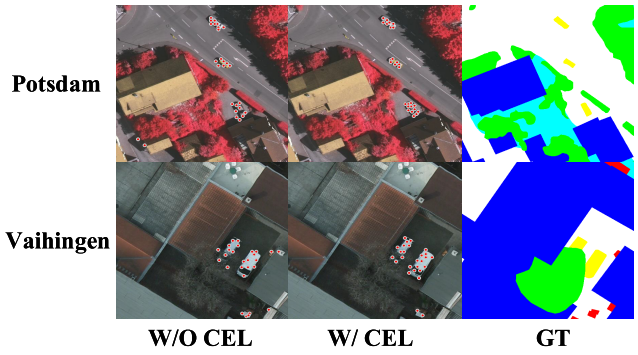

Fig. 9. Visualization of keypoint detection on the "Car" category in the Potsdam dataset.

to more effectively fuses features from two modalities. And the proposed CEL, which introducing weights of different categories in the label map and embedding network training, implements CEL to overcome the sample imbalance problem.

Effectiveness of CFF and CEL: Fig. 8 shows the class activation maps for "Building" and "Tree" on the Potsdam and Vaihingen dataset, respectively. Our model accurately segments the categories even under the shadow challenges marked by yellow boxes in the Potsdam dataset. Meanwhile, the yellow boxes marked in the Vaihingen dataset are not in the "Tree" category. While our model successfully avoids false prediction with the guidance of CFF.

Fig. 9 shows the visual examples of the key points in the "Car" category on the original image. The category of the "Car" belongs to hard examples (with a small proportion), and it is challenging to identify them accurately. As shown in Fig. 9, CEL uses the prior knowledge of category weights to guide the network to learn more effective category features, and the key points are more densely distributed in the "Car" category area.

TABLE IV

mIoU (%) ON POTSDAM DATASET USING LOSS FUNCTION IN DIFFERENT DECODERS

$\mathcal{L}oss$	mIoU (%)
(a) $\mathcal{L}_4$	76.88
(b) $\mathcal{L}_4$ +$\mathcal{L}_3$	77.07
(c) $\mathcal{L}_4$ +$\mathcal{L}_3$ +$\mathcal{L}_2$	77.81
(d) $\mathcal{L}_4$ +$\mathcal{L}_3$ +$\mathcal{L}_2$ +$\mathcal{L}_1$	77.64

Evaluation on PLS: Adding the focal loss [19] function to each layer of the pyramid loss can supervise network training and focus on the features of objects in feature maps at different scales. To investigate the impact of different loss combinations, we evaluate the performance of our method by progressively adding the focal loss into each layer in the decoder. As shown in Table IV, in general, after progressively adding the loss function in the second to the fourth layers of the decoder, the mIoU consistently increases. This implies that PLS can effectively learn the object features in different layer decoders. However, as shown in Table IV(d), the mIoU declines after further adding the loss at the first layer of the decoder. The main reason is that the output resolution of the first layer is shallow. Meanwhile, the label map downsampling may lose detailed information, which brings training fluctuations caused by noisy labels. Therefore, in the PLS, we enforce the loss into the last three layers as shown in Fig. 3.

E. Baseline and Backbone Analysis

Different Baselines Plugin: To verify the generality of the CFF and CEL, we integrate them together into four latest single-modal remote sensing image semantic segmentation method FarSeg [11], MACU-Net [33], and FactSeg [12], MANet [40] and two multimodal latest method DSMFNet [16] and V-FuseNet [17]. As shown in Table V, CFF + CEL further boost the mean-F_1 on all the six methods, which verifies the generality of the proposed CFF and CEL. The improvements in the categories of "Low veg." and "Imp. surf." are slightly overshadowed comparing with the other three categories after introducing CFF and CEL. The main reason is that "Low veg." and "Imp. surf." encounter occlusion problems that lead to semantic ambiguity. However, we can still improve the segmentation results in these two categories. Note that introducing CFF and CEL may bring negative contribution to the segmentation, such as the "Low veg." category in multimodal methods DSMFNet [16] and V-FuseNet [17]. The main reason is that the features of the "Low veg." categories are not distinct in DSM images, which is challenging to be extracted. The simple feature extractor of the DSM images in DSMFNet [16] and V-FuseNet [17] cannot sufficiently learn the challenging "Low veg." category.

Backbones Analysis: For the more comprehensive comparison, we further evaluate our modules with two more complex backbones ResNet-101 [51] and vision transformer (ViT) [58] as shown in Table VI. First, it can be seen that when adding our proposed module to each backbone, there is a significant improvement on mIoU, which verifies the generality of the proposed modules. Second, using the more complex backbones ResNet-101 [51] and ViT [58] do not achieve better results comparing to the widely used backbone

TABLE V
EVALUATION OF CFF AND CEL ON OTHER ADVANCED METHODS

Method	Imp. surf.	Building	Low veg.	Tree	Car	mean-F_1 (%)
FarSeg [11]	91.56	94.22	84.67	89.45	90.41	90.06
+CFF+CEL	+0.08	+0.98	+0.32	+0.68	+0.52	+0.51
FactSeg [12]	90.68	94.17	85.37	89.01	91.10	90.06
+CFF+CEL	+0.20	+0.11	+0.18	+0.76	+0.41	+0.33
MACU-Net [33]	92.34	93.23	85.23	87.45	88.19	89.28
+CFF+CEL	-0.22	+0.79	+0.19	+0.52	+0.20	+0.29
MANet [40]	90.03	93.56	85.92	85.34	89.15	88.80
+CFF+CEL	+0.12	+1.20	+0.43	+1.37	+0.82	+0.78
DSMFNet [16]	91.64	92.24	84.66	86.80	89.02	88.87
+CFF+CEL	+0.22	+0.06	-0.37	+0.83	+0.11	+0.17
V-FuseNet [17]	90.91	93.23	84.30	86.23	90.56	89.04
+CFF+CEL	+0.33	+1.37	-0.37	+0.69	+0.49	+0.50

TABLE VI
MIOU (%) ON POTSDAM DATASET USING VARIOUS BACKBONES

Method	mIoU (%)
ResNet-101 [51]	73.89
+ CaFE (CFF+CEL+PLS)	77.62
ViT-B/16 [58]	74.32
+ CaFE (CFF+CEL+PLS)	77.39
ResNet-50 [51]	74.38
+ CaFE (CFF+CEL+PLS)	77.81

TABLE VII
MIOU (%) ON POTSDAM DATASET USING IMAGES WITH DIFFERENT CATEGORY DISTRIBUTIONS

Method	Balanced	Imbalanced	mIoU (%)
FactSeg [12]	√		78.86
		√	75.42
MACU-Net [33]	√		76.19
		√	73.33
Ours (CaFE)	√		78.94
		√	76.68

ResNet-50 [51]. This is because DSM images do not contain much texture and color information, and a more complex backbone may affect the model training. As shown in Table VI, our method (CaFE) achieves the best results when adopting ResNet-50 [51].

Other Analysis: To evaluate our method when handling the images with different category distributions, we divide the test images in the Potsdam dataset into two parts: balanced (images with all six categories) and imbalanced (the rest of images with less categories) for evaluation. As shown in Table VII, both the state-of-the-art methods FactSeg [12] and MACU-Net [33] and our CaFE decline in the imbalanced case comparing to the balanced case. The main reason is that when there are only a few categories in the image, the help of contextual information of other categories is lacking. Comparing with FactSeg [12] and MACU-Net [33], the proposed CaFE achieves the best performance in the imbalanced case as well as the balanced case. This evidences that under the guidance of the CEL, our model can better learn the image distribution while overcoming the category imbalance problem.

F. Parameter Analysis

There are two critical parameters in our methods, k in (2) indicating the number of interactions among the channels, and r in (6) representing the weight of samples. Table VIII evaluates the influence of these two parameters in our method. Generally speaking, our method is not sensitive to these two parameters. As the value of k increases, the mIoU slightly decreases, which indicates that under the condition of sample imbalance, the higher the similarity of category features between adjacent channels. When r is too large, the mIoU tends to decline. The main reason is that too much supervision of hard samples (with a small proportion) slows down the network convergence. Therefore, we set k to 4 and r to 2 for the best performance.

TABLE VIII
MIOU(%) ON POTSDAM DATASET USING VARYING k FOR CA AND r FOR PLS

k	3	4	6	8	12
mIoU (%)	77.68	77.81	77.51	77.02	76.84
r	0.6	1	2	4	6
mIoU (%)	76.47	77.28	77.81	77.62	77.01

V. CONCLUSION

In this article, to solve the problem of similar spectral features of foreground objects and larger intraclass variance of background and the sample imbalance problem. We propose a multimodal remote sensing image semantic segmentation framework called CaFE based on sufficient exploitation of structural information in the label map. Specifically, we propose the CFF module which uses category mask to aggregates the feature of the same category in different channels to achieve the category-wise level feature fusion. Then we propose the CEL module by learning weight factors in the label map to overcome the sample imbalance problem. Meanwhile, we use the PLS to focus on the multiscale features of objects in HSR remote sensing images. Extensive experiments have shown the superior performance of the proposed CaFE on two public aerial benchmarks.

REFERENCES

[1] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.
[2] C. Zhang et al., "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, Feb. 2019.
[3] M. Dickenson and L. Gueguen, "Rotated rectangles for symbolized building footprint extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 225–228.

[4] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, 2018.

[5] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[6] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[7] J. Liang, N. Homayounfar, W.-C. Ma, S. Wang, and R. Urtasun, "Convolutional recurrent network for road boundary extraction," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9512–9521.

[8] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[9] M. T. Chiu et al., "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2828–2838.

[10] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[11] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4096–4105.

[12] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[13] X. Li et al., "PointFlow: Flowing semantics through points for aerial image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4217–4226.

[14] A. Zheng, M. Wang, C. Li, J. Tang, and B. Luo, "Entropy guided adversarial domain adaptation for aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[15] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.

[16] Z. Cao et al., "End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Oct. 2019.

[17] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[18] S. Waqas Zamir et al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 28–37.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[20] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder–decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[23] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[24] L. Zhang et al., "Learning deep structured active contours end-to-end," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8877–8885.

[25] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2016.

[26] D. Sulla-Menashe and M. A. Friedl, "User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product," USGS, Reston, VA, USA, 2018, pp. 1–18.

[27] H. Alemohammad and K. Booth, "LandCoverNet: A global benchmark land cover classification training dataset," 2020, *arXiv:2012.03111*.

[28] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[29] I. Demir et al., "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[30] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020.

[31] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–9.

[32] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*.

[33] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for semantic segmentation of fine-resolution remotely sensed images," 2020, *arXiv:2007.13083*.

[34] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[36] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[37] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[38] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[39] R. Li, S. Zheng, C. Zhang, C. Duan, and L. Wang, "A2-FPN for semantic segmentation of fine-resolution remotely sensed images," 2021, *arXiv:2102.07997*.

[40] R. Li, S. Zheng, C. Duan, C. Zhang, J. Su, and P. M. Atkinson, "Multi-attention-network for semantic segmentation of fine resolution remote sensing images," 2020, *arXiv:2009.02130*.

[41] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[42] B. Kellenberger, D. Marcos, S. Lobry, and D. Tuia, "Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep CNNs and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9524–9533, Dec. 2019.

[43] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, Aug. 2020.

[44] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[45] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.

[46] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 180–196.

[47] M. Volpi and D. Tuia, "Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 48–60, Oct. 2018.

[48] S. Srivastava, J. E. Vargas-Muñoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote Sens. Environ.*, vol. 228, pp. 129–143, Jul. 2019.

[49] C. Zhang, "Based on multi-feature information attention fusion for multi-modal remote sensing image semantic segmentation," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2021, pp. 71–76.

[50] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[53] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[54] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[55] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.
[56] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
[57] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[58] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

Aihua Zheng received the B.Eng. degree in computer science and technology from Anhui University, Hefei, China, in 2006, and the Ph.D. degree in computer science from the University of Greenwich, London, U.K., in 2012.

She visited the University of Stirling, Stirling, U.K., and Texas State University, San Marcos, TX, USA, from June to September in 2013 and from September 2019 to August 2020, respectively. She is currently an Associate Professor and the Ph.D. Supervision with the School of Artificial Intelligence, Anhui University. Her main research interests include vision-based artificial intelligence and pattern recognition. As the first author or corresponding author, she has published more than 40 academic articles including top conferences papers in the Association for the Advancement of Artificial Intelligence (AAAI) and the International Joint Conference on Artificial Intelligence (IJCAI), and authoritative journals in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), the IEEE TRANSACTIONS ON SYSTEMS, MULTI-ATTENTION-NETWORK (MAN) AND CYBERNETICS: SYSTEMS (TSMCS), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), *Pattern Recognition* (PR), *Cognitive Computation* (CogCom), and so on.

Dr. Zheng is a member of the China Computer Federation (CCF) and the China Society of Image and Graphics (CSIG). She is also serving as reviewers for representative conferences and journals, including AAAI, IJCAI, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, PR, and so on. She has been awarded the Best Paper in SERA 2017 and the Best Student Paper in the workshop in International Conference on Multimedia and Expo (ICME) 2019.

Jinbo He received the B.Eng. degree in software engineering from Bengbu University, Bengbu, China, in 2020. He is currently pursuing the M.Eng. degree in computer science and technology with Anhui University, Hefei, China.

His research interests include computer vision and semantic segmentation of remote sensing images.

Ming Wang received the B.Eng. degree in computer science and technology from the Anhui University of Technology, Maanshan, China, in 2019. He is currently pursuing the M.Eng. degree in computer science and technology from Anhui University, Hefei, China.

His research interests include computer vision and optical remote sense image semantic segmentation.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively.

From 2014 to 2015, he worked as a Visiting Student with the School of Artificial Intelligence, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently an Associate Professor and the Ph.D. Supervisor with the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning.

Dr. Li was a recipient of the Association for Computing Machinery (ACM) Hefei Doctoral Dissertation Award in 2016.

Bin Luo (Senior Member, IEEE) received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002.

He is currently a Professor with Anhui University. He has authored over 200 articles in journals, edited books, and refereed conferences. His research interests include random graph-based pattern recognition, image and graph matching, graph spectral analysis, and video analysis.

Dr. Luo is also the Chair of the IEEE Hefei Subsection. He has served as a Peer Reviewer for international academic journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition* (PR), *Pattern Recognition Letters* (PRL), the *International Journal of Pattern Recognition and Artificial Intelligence*, *Knowledge and Information Systems*, and *Neurocomputing* (NeuCom).