ISPRS Journal of Photogrammetry and Remote Sensing 184 (2022) 96–115

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images

Hamidreza Hosseinpour [a,*], Farhad Samadzadegan [a], Farzaneh Dadrass Javan [a,b]

[a] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 NB Enschede, the Netherlands

ARTICLE INFO

Keywords:
Building extraction
VHR remote sensing image
Digital surface model
Gated fusion module
Cross-modal

ABSTRACT

The extraction of urban structures such as buildings from very high-resolution (VHR) remote sensing imagery has improved dramatically, thanks to recent developments in deep multimodal fusion models. However, Due to the variety of colour intensities with complex textures of building objects in VHR images and the low quality of the digital surface model (DSM), it is challenging to develop the optimal cross-modal fusion network that takes advantage of these two modalities. This research presents an end-to-end cross-modal gated fusion network (CMGFNet) for extracting building footprints from VHR remote sensing images and DSMs data. The CMGFNet extracts multi-level features from RGB and DSM data by using two separate encoders. We offer two methods for fusing features in two modalities: Cross-modal and multi-level feature fusion. For cross-modal feature fusion, a gated fusion module (GFM) is proposed to combine two modalities efficiently. The multi-level feature fusion fuses the high-level features from deep layers with shallower low-level features through a top-down strategy. Furthermore, a residual-like depth-wise separable convolution (R-DSC) is introduced to enhance the performance of the up-sampling process and decrease the parameters and time complexity in the decoder section. Experimental results from challenging datasets show that the CMGFNet outperforms other state-of-the-art models. The efficacy of all significant elements is also confirmed by the extensive ablation study.

1. Introduction

Accurate extraction and identification of manufactured structures in urban environments, especially buildings, from very high-resolution (VHR) images obtained by aerial or satellite sensors are essential for a variety of applications, like 3D modeling, infrastructure planning, and urban expansion analysis (Freire et al., 2014; Hoeser and Kuenzer, 2020; Wu et al., 2018; Xu et al., 2019). Due to the easy access and cost-effectiveness of VHR images in recent years, the extraction and segmentation of building objects from these images have been considered by researchers (Osco et al., 2021).

Before applying deep learning methods in the community of photogrammetry and remote sensing, various traditional methods are classified into two groups depending on the dimensions and availability of the dataset: 3D point cloud-based methods and 2D image-based methods. The first category of these algorithms exclusively employs 3D Lidar point clouds or DSM data to extract building objects. Methods that use these types of data include threshold height information (Weidner, 1997), edge detection for 2D line extraction (Hermosilla et al., 2011), planes analyses in the 3D space (Hu et al., 2004), and using 3D templates (Hammoudi and Dornaika, 2010). Nevertheless, the disadvantage of these algorithms is that these data can have the limitation of texture and inaccurate boundary information, introducing errors to the building extraction task. In the second category, various methods for extracting buildings through 2D VHR images are presented. For instance, the method based on energy minimization function like active contour model-based (Ahmadi et al., 2010), the process based on image segmentation like graph theory (Sirmacek and Unsalan, 2009), the method based on graphical models like Markov Random Fields (MRFs) (Ngo et al., 2017), object-based method (Mohammadi and Samadzadegan, 2020; Tomljenovic et al., 2016), morphological index-based method (Huang and Zhang, 2012) and traditional machine learning methods (Jiang et al., 2018; Ozdarici-Ok et al., 2015; Pacifici et al., 2009; Zhang et al., 2015). Although these methods have made significant advances in the extraction of building objects from 2D VHR images, achieving accurate results in the extraction of buildings objects using only RGB spectral channels is challenging for various reasons: 1- Unlike the images used in computer vision applications, VHR images include vast areas

* Corresponding author at: North Kargar Street, P.O Box: 11365-4563, Tehran 11365-4563, Iran.
E-mail addresses: hosseinpoor@ut.ac.ir (H. Hosseinpour), samadz@ut.ac.ir (F. Samadzadegan), fdadrasjavan@ut.ac.ir, f.dadrassjavan@utwente.nl (F.D. Javan).

https://doi.org/10.1016/j.isprsjprs.2021.12.007
Received 26 May 2021; Received in revised form 15 August 2021; Accepted 17 December 2021
Available online 29 December 2021

that cause variation scale and dimensions of buildings. 2- The limited spectral properties of VHR images cause a great variety within the low inter-class, and extensive intra-class features in buildings object. 3- Existence of shadows, noises, obstructions, geometric deformation, and height displacement of high buildings. Accordingly, to overcome the problems mentioned, due to the development of airborne light detection and ranging (LiDAR) technology and the advancement of Image-Dense-Matching (IDM) (Salach et al., 2018), traditional fusion methods in building extraction from VHR images have been developed to combine spectral and height information (Guan et al., 2013; Vetrivel et al., 2015), that have shown better building extraction results rather than using a single modality (Zhang et al., 2020a).

Recently, most segmentation methods of VHR remote sensing images use deep convolutional neural networks (DCNNs), which are generally superior in performance to traditional methods and have attained state-of-the-art results in challenging datasets. Building extraction from VHR remote sensing images is similar to semantic segmentation task developed in the computer vision community, and it is aimed to label the entire pixels of an image as building or non-building classes. Several recent studies have also applied DCNN-based techniques specifically to extract buildings automatically from remote sensing images (Feng et al., 2020; Hosseinpoor and Samadzadegan, 2020; Ji et al., 2019; Ma et al., 2020; Maggiori et al., 2017; Maltezos et al., 2019; Pan et al., 2019; Shao et al., 2020; Wu et al., 2018; Xu et al., 2018; Zhang et al., 2020b; Zhang and Wang, 2019). Most of the above research used the idea of fully convolutional networks (FCN) (Long et al., 2014) for segmentation. In FCN, fully connected layers of DCNNs are replaced by standard convolution layers. To reduce the trade-off between recognition and correct localization, (Maggiori et al., 2017) created a two-scale neuron module in an FCN. (Xu et al., 2018) merged the new DCNN based on FCN and guided filtering to further refine the extraction results of buildings. (Wu et al., 2018) proposed a multi-constraint FCN to perform end-to-end extraction of building from aerial images. (Zhang and Wang, 2019) proposed a network based on the combination of atrous convolution and dense connectivity to increase the receptive field. (Ma et al., 2020) proposed GMEDN, which uses a local and global encoder based on the VGG-16 network. In (Shao et al., 2020), The prediction and the residual refinement module are introduced. The prediction module introduces atrous convolution of different dilation rates to extract more global features. These advanced DCNNs, on the other hand, are typically limited to three-channel RGB images, which cannot be easily used for multimodal data. For more accurate and reliable building extraction, a comprehensive deep model integrating VHR and DSM data is required. As indicated in Fig. 1, most DCNN-based fusion algorithms may be categorized into three types, depending on where multiple modalities are fused (Zhang et al., 2021).

Early or data-level fusion approaches (Fig. 1a) combine spectral information such as red, green, blue, and near-infrared spectrum with structural information such as DSM as the input to a conventional unimodal or multimodal network. In (Nahhas et al., 2018) utilized object-based analysis with an autoencoder-based dimensionality reduction. The output features transform into high-level features by a DCNN, used to classify objects into buildings and background. Liu et al. (2020) developed a trainable enhanced U-Net model (Ronneberger et al., 2015) for building extraction that combines high spatial resolution unmanned aerial vehicle (UAV) Images with DSM. In (Huang et al., 2019), the author introduced a gated residual refinement network (GRRNet), which concatenates raw data from different modalities into several channels. The encoder element of the GRRNet is composed of this modified residual network, and gated feature labeling (GFL) is employed to improve segmentation results. Data-level fusion methods produce incorrect or irrelevant features in network training because such methods may not efficiently utilize the complementary nature of the modalities. Also, it is impossible to initialize the proposed models with pre-trained DCNNs models in this method.

In late or decision-level fusion methods (Fig. 1b), spectral and structural data are sent to the two different encoder-decoder network streams and predicted classes fused in the final stage (Marmanis et al., 2018; Piramanayagam et al., 2018). This fusion method may provide more scalability and flexibility than the early fusion method. However, there is an inadequate cross-modal correlation between the corresponding features in the two streams. Middle or feature-level fusion methods (Fig. 1c) that spectral and structural information are sent to separate identical encoders for each modality, and the lateral features of two encoders are merged in the cross-modal using concatenation operation or element-wise summation at different scales (Audebert et al.,

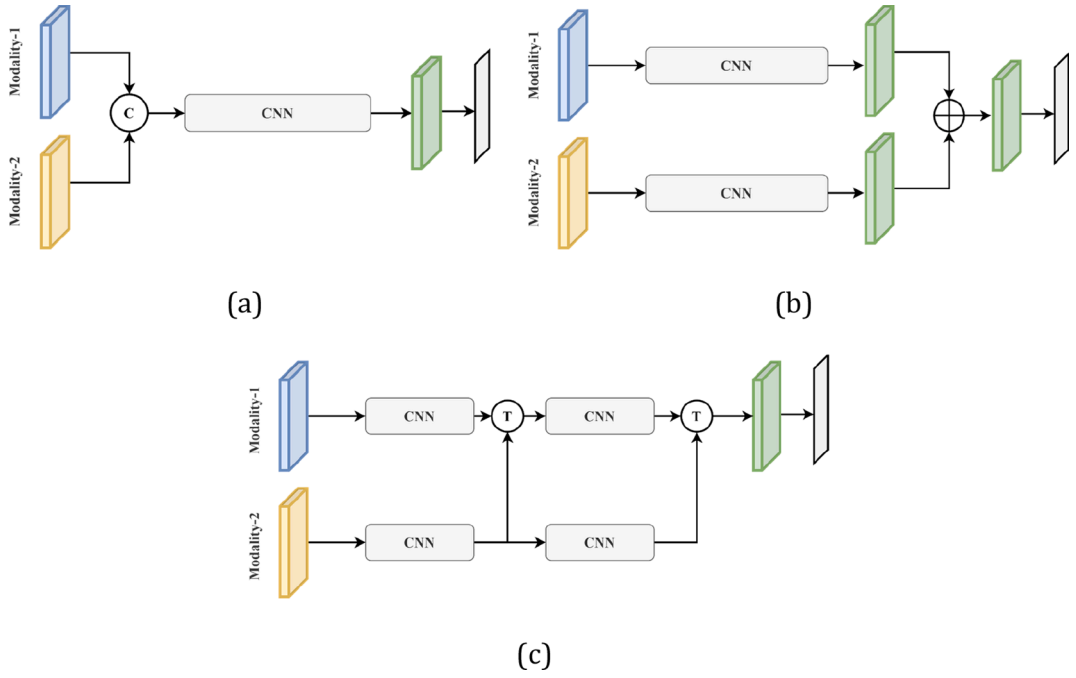

Fig. 1. Different methods for multimodal deep learning semantic segmentation. (a) Early fusion, (b) Late fusion, (c) Middle fusion. Where '+' and 'C' represents the element-wise summation and concatenation operation, respectively. 'T' can denote each of 'C' or '+'.

2018; Piramanayagam et al., 2018; Sun et al., 2018; Xu et al., 2019; Zhang et al., 2020a; Zhang et al., 2017a). For instance, FuseNet (Hazirbas et al., 2017) uses the SegNet (Badrinarayanan et al., 2017) architecture to meaningfully segment RGB-D data by incorporating a cross fusion algorithm into the encoder section. (Piramanayagam et al., 2018) introduced a DCNN based on FCN-32 to merge features from multi-sensor for semantic segmentation tasks. The author combines two modalities in the different convolutional layers. The results show that the early layer fusion, specifically after three layers achieves better results than a decision level fusion. (Zhang et al., 2017a) conducted a thorough analysis of the sensitivities and contributions of each layer in FCN, resulting in the creation of an optimal layer fusion architecture. (Audebert et al., 2018) applied two techniques for fusing multi-sensor features. In cross-modal fusion, they use the FuseNet model, and for late fusion, they use residual correction. (Zhang et al., 2020a) proposed a feature-level fusion network based on a hybrid attention-aware mechanism (HAFNet). They train both individual and cross-modal features using RGB image and DSM data. They also develop a multimodal attention-aware fusion block to solve the fusion problem of multiple modalities in the cross-modal stream (Att-MFBlock). In the field of computer vision, (Liu et al., 2020) presented an adaptive gated fusion generative adversarial network (GAN). The generator part of the network adopts two RGB and depth data encoder-decoder networks, and the RGB stream feature guides depth stream to achieve cross-modal fusion. The results of these previous studies show that the combination of height information with DCNN models has improved the building extraction accuracy (Zhang et al., 2021). However, most of the previous feature-level fusion methods contain complex structures, and they are based primarily on the same and straightforward techniques of weighting the corresponding features from the RGB and DSM stream. Therefore, useful information in the DSM data is not thoroughly utilized. Furthermore, as an influential factor in extracting building objects, little attention has been paid to improving the encoder and decoder sections in recent fusion models.

To overcome the problems mentioned in previous researches to extract complex buildings from VHR images, this research proposes a novel cross-modal gated fusion network (CMGFNet). The primary task of the proposed network is how to fuse RGB and DSM features in cross-modal conduction. Rather than a straightforward concatenation or a summation fusion with equivalent weights for different modalities, a gated fusion module (GFM) is introduced to adaptively learn the discriminative features by weighting each modality and removing irrelevant parts. On the other hand, many DCNNs offered for automatic building extraction are built on the encoder-decoder architecture. Due to the unique structure of this type of architecture, the features extracted in the deep layers (i.e., high-level features) have high semantic information but instead low spatial resolution. The features in the shallower layers (i.e., low-level features) have high spatial information and contain low semantic information of building features. The information on these features has been lost, owing to the up and down sampling process in the encoder-decoder networks. In this regard, the multi-level feature fusion is proposed, in which high-level features with high semantic definitions introduce into low-level features. Therefore, how to preserve semantic information in the primary layers for proper up-sampling is the second goal of this study. As a third goal, a new decoder block based on residual pyramidal blocks and depth-wise separable convolution are proposed. This block is called residual-like depth-wise separable convolution (R-DSC). R-DSC is used to up-sample the high-level semantic features of building objects in both RGB and cross-modal streams. The R-DSC architecture is critical for retaining, distributing, and decreasing the number of decoder parameters, so it is advantageous to improve the efficiency of building extraction from VHR remote sensing images and DSM data. The following are the main contributions of this research:

- In this paper, the cross-modal gated fusion network (CMGFNet) is presented as a method for end-to-end building extraction from VHR remote sensing images and DSM data.
- Analysing how the proposed GFM, multi-level feature fusion, and R-DSC affect the refinement of multimodal data fusion and building extraction results.
- Comparing the proposed CMGFNet with other state-of-the-art models in three public urban scenes.

The remainder of this paper is organized as follows. The CMGFNet method is fully described in Section 2. Then the detail of the implementation and ablation experiments are dedicated to Section 3. Finally, the discussion and conclusion of this paper are in Sections 4 and 5.

2. Proposed method

Fig. 2 shows the workflow for extracting buildings from VHR remote sensing images and LiDAR-derived DSM data using the cross-modal gated fusion network (CMGFNet) architecture. The proposed network is based on a gated fusion module (GFM) to fully use each modal feature from multimodal data. First, red (R), green (G), and blue (B) bands of VHR image and single-band DSM data are fed into the two separate network streams. Each stream takes the ResNet-34 network (He et al., 2016) as the backbone network in the encoder part. Then, the lateral features of each ResNet-34 block are combined into the decoder stream and at least generate the RGB prediction map (P_{rgb}), DSM prediction map (P_{dsm}), and fused prediction map (P_{fusion}). The decoder section is based on depth-wise separable convolution (DSC) and residual-like convolutional module to process deep prediction information. This decoder module is called R-DSC in the paper. First of all, the general architecture of the CMGFNet is presented, then the component shown in this architecture describes individually.

2.1. Network architecture

The CMGFNet adopts two separate streams of the encoder-decoder network. The architecture of ResNet-34 (He et al., 2016) is chosen for extracting deep features of the building in both encoder branches. The last fully connected layers of this architecture are discarded for dense pixel prediction. One of the advantages of the proposed model is using other networks similar to the ResNet-34 in the encoder section. Deep-residual networks have been demonstrated to reduce gradient degradation problems in model training (He et al., 2016) and are accurate for the localization of building features (Audebert et al., 2018). In the CMGFNet, RGB and DSM streams have the same network settings, except that the first convolutional block on the DSM branch has only one channel because the DSM input is presented as one channel. For the RGB stream, at the end of the encoder part, after the pooling layer, the feature is sent to the R-DSC module. The lateral output of the Resnet-34 model from the RGB stream is cascaded with the up-sampled result. Once again, these cascaded features are sent to the R-DSC module, and in the same way, the process continues.

This process differs from the RGB decoder stream for DSM decoder streams. Since the features derived from the DSM stream are noisy and not clear enough, the feature of the RGB encoder is employed to take care of the DSM feature disadvantage. This is done by fusing lateral features from the RGB and DSM streams using the proposed GFM. The aim of the GFM is that it uses the representation of RGB and DSM features to understand which of the modalities should affect the prediction. In practice, the lateral features of RGB and DSM stream are used by the multi-level features fusion method. This method produces proper high-resolution semantic instruction by fusion of deeper layers with high-level features and shallower low-level features derived from the encoder part of the model.

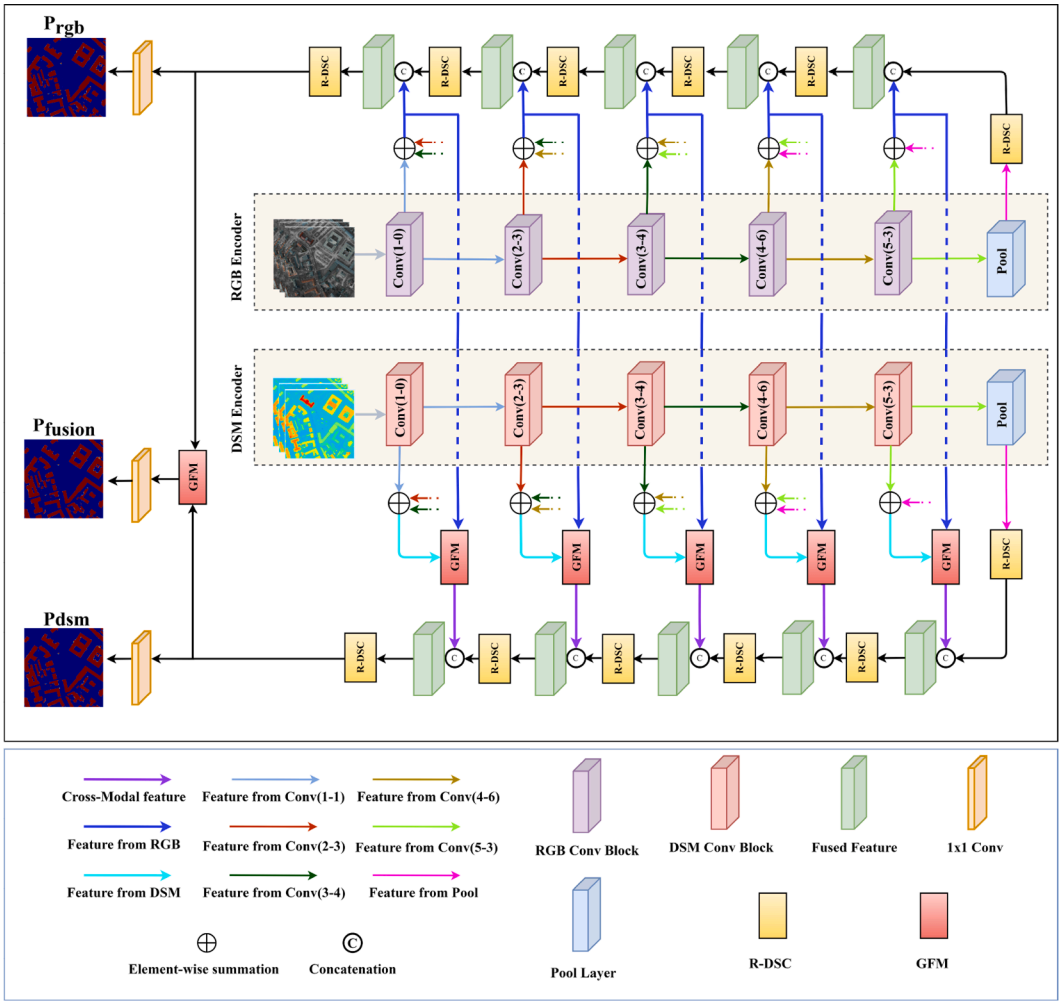

Fig. 2. Illustration of the CMGFNet architecture.

2.2. Two-Stream feature extraction

The CMGFNet utilizes two encoder-decoder networks to learn the deep representation of RGB and DSM features. In particular, the ResNet-34 structure is employed in the encoder parts. The ResNet-34 model first uses a convolution block with 7 × 7 kernels, followed by the four residual blocks. In Fig. 2, the blocks of ResNet-34 are named Conv(p-l). P denotes the number of blocks, and l is equal to the number of residual convolution layers. Each residual convolution layer is composed of two convolutions with 3 × 3 kernels, and by using a skip connection, the input of the block is summed into the output feature. A pre-trained model on the ImageNet dataset is used to initialize the weights of the encoder component of the convolution operation. Let E_{rgb} and E_{dsm} denote side-output encoder features for both RGB and DSM streams. If the input RGB and DSM Image are all cropped and resized into W × H, the dimension for both E^{i}_{rgb} and E^{i}_{dsm} from the first to fifth convolution block (i ∈ {1, 2, 3, 4, 5}) are (w/2 × h/2), (w/4 × h/4), (w/8 × h/8), (w/16 × h/16), and (w/32 × h/32) respectively. The E^{6}_{rgb} and E^{6}_{dsm} are related to the pooling layer and have (w/64 × h/64) dimensions. For decoder stream, inspired by (Zhang et al., 2018), used to up-sample every input features map by cascading them with the fusion of high-level lateral encoder features through the skip network connection. In other words, the fusion of high-level lateral features could be enhanced by incorporating more concepts of building objects into low-level features (Liu et al., 2020). The following formula is provided with the multi-level feature fusion for the RGB encoder of the network. Let D^{m}_{rgb} denote the decoder of RGB stream, then we have:

$$D^{m}_{rgb} = \begin{cases} \left(r^{m}_{m+1}E^{m+1}_{rgb} + r^{m}_{m+2}E^{m+2}_{rgb}\right) \copyright R - DSC\left(D^{m+1}_{rgb}\right), & m = 1, \cdots 4 \\ \left(r^{m}_{m+1}E^{m+1}_{rgb}\right) \copyright R - DSC\left(E^{m+1}_{rgb}\right), & m = 5 \end{cases} \quad (1)$$

where $r^{n'}_{n}$, (n > n'), denote the weight of short connection from RGB side-output n to side-output n', R-DSC(.) represents the processed feature by R-DSC module, and © indicates the feature concatenation operation. Concatenation operation combines high-level and low-level features and expands the feature channel space by allowing the subsequent R-DSC module to learn additional features dependent on both features. Because the features produced from DSM encoders have little edge information, this work designs the straightforward method based on the GFM for fusing the lateral feature from the RGB encoder to the corresponding lateral feature of the DSM encoder. Similar to the RGB decoder, the multi-level feature fusion is used for both the RGB and DSM encoder feature before applying cross-modal fusion. Let D^{m}_{dsm} represent the decoder output of the cross-modal fusion, then the decoder feature of the DSM stream can be expressed mathematically as:

$$D_{dsm}^{m}=\begin{cases}GFM\left(r_{m+1}^{m}E_{rgb}^{m+1}+r_{m+2}^{m}E_{rgb}^{m+2},l_{m+1}^{m}E_{dsm}^{m+1}+l_{m+2}^{m}E_{dsm}^{m+2}\right)\copyright R-DSC\left(D_{dsm}^{m+1}\right), & m=1,\cdots,4\\ GFM\left(r_{m+1}^{m}E_{rgb}^{m+1},l_{m+1}^{m}E_{dsm}^{m+1}\right)\copyright R-DSC\left(E_{dsm}^{m+1}\right), & m=5\end{cases} \tag{2}$$

where $l_{n}^{n'}$, (n > n'), denote the weight of short connection from DSM side-output n to side-output n', GFM (.,.) represent the gated fusion module which takes two side-output feature and fuses them to obtain better feature representation.

Finally, the prediction map of building extraction from the RGB and DSM streams is calculated according to the following equation:

$$P_{rgb}=sig\left(conv_{1\times1}^{2}D_{rgb}^{1}\right) \tag{3}$$

$$P_{dsm}=sig\left(conv_{1\times1}^{2}D_{dsm}^{1}\right) \tag{4}$$

where the P_{rgb}, and the P_{dsm}, represent the RGB and DSM prediction map respectively, furthermore, the fused prediction map can be express as:

$$P_{fusion}=sig\left(conv_{1\times1}^{2}(GFM(D_{rgb}^{1},D_{dsm}^{1}))\right) \tag{5}$$

$conv_{k\times k}^{c}(.)$ is applied to reduce the dimensions of the features channels. 'c' represents the number of output channels, and convolution operation uses k × k kernel to obtain c-channel features. The sigmoid function for constructing a prediction map is denoted by sig (.).

2.3. Gated fusion module (GFM)

Unlike conventional methods for feature fusion in the cross-modal stream, which primarily are based on elementwise summation and concatenation operation, the proposed fusion module in this work is based on a gated fusion module (GFM). This module is influenced by (Arevalo et al., 2020) to calculate the utility of each corresponding lateral feature from RGB and DSM encoder and accumulates information accordingly.

Fig. 3 illustrates the structure of the GFM; consider F_{rgb} and F_{dsm} represent the output of the feature map on the RGB and DSM decoder, respectively. The number of feature channels for each modality is the same. First, the features are concatenated and produced fused feature map F_{fusion}. Then the convolution operation with 1 × 1 kernel, W_z, is used to calculate how modalities correlate with each other and decrease the dimension of the feature channels. At least, the sigmoid function is employed to obtain the weighed probability matrix, G.

$$F_{fusion}=W_z\left(F_{rgb}\copyright F_{dsm}\right) \tag{6}$$

$$G=sig\left(F_{fusion}\right) \tag{7}$$

In Eq. (6), © denotes the feature concatenation operator, and in Eq. (7) sig (.) represents the sigmoid function. Let $G^{rgb}=G$ and $G^{dsm}=1\text{-}G$ represent the weighted gates of the RGB and DSM modalities. So the gate fusion map, $F_{gate\text{-}fusion}$, of the GFM is given by:

$$\begin{aligned}F_{gate-fusion}&=\left(f_{rgb}\otimes G^{rgb}\right)\copyright\left(f_{dsm}\otimes G^{dsm}\right)\\&=\left(\left(f_{rgb}\otimes G\right)\copyright\left(f_{dsm}\otimes\left(1-G\right)\right)\right)\end{aligned} \tag{8}$$

where ⊗ denotes Hadamard product. the Eq. (8) demonstrates in a convex combination and each modality can have different weights.

2.4. Residual depth-wise separable convolutional (R-DSC)

This work proposes a residual depth-wise separable convolutional (R-DSC) module, which is based on residual connection unit (RCU) (He et al., 2016) and depth-wise separable convolutional (DSC) operation (Chollet, 2016) for the decoder section of the CMGFNet. Combining original and residual features be beneficial for the various networks in deep learning tasks. Different versions of the residual connection unit consisting of different configurations of convolution layers, rectified linear unit (ReLU) (Nair and Hinton, 2010), and batch normalization (BN) (Ioffe and Szegedy, 2015) have been proposed in recent years. Contrary to the original RCU presented in (He et al., 2016), this work uses the idea of pyramidal RCU (Han et al., 2016). The modification in the location of ReLU and BN is shown to improve the training models and achieve better results in the computer vision tasks. It performs the BN before the first standard 3 × 3 convolution layer, as shown in Fig. 4. After applying the up-sampling and residual function, the output feature has a higher spatial resolution, but the number of feature channels is lower than the input feature channel. As a result, it is not possible to use the residual connections directly for up-sampling features in the decoder. Therefore, the convolution operation with 1 × 1 kernel is applied to convert the input features to the desired number of channels and fuses its result by element-wise summation with the residual unit.

In addition, the original standard convolutional layer is changed with depth-wise separable convolution to minimize the computation cost and parameters of the model. There are two types of convolutions in this process: depth-wise and point-wise convolutions, as shown in Fig. 5. The standard convolution is used for each channel of the input features in depth-wise convolution individually. In point-wise convolutions, a standard 1 × 1 convolution operation performs on the output feature map from the depth-wise convolution.

The presented R-DSC is important from two points of view. From one side, by retaining more features in the pyramidal RCU, R-DSC enhances and keeps more original features. R-DSC is used to each step of the decoder network of both RGB and DSM stream to rebuttal information. The high-level feature of the encoder parts of the CMGFNet is processed by R-DSC, and the information is more maintains and passed to primary layers. Additionally, improved features at various scales improve the building extraction performance by increasing the diversity of original features. On the other hand, in pyramidal RCU, the standard convolution module is replaced with depth-wise separable convolution, which reduces the number of parameters in our network. According to (Kaiser et al., 2017), the parameters and calculation cost of the depth-wise separable convolution perform $(1/n+1/k^2)$ times lesser than a standard convolution. Where k denotes the kernel size of the depth-wise convolutions, and the number of output channels from the point-wise convolutions is given by n.

2.5. Loss function

In this research, the loss function for the CMGFNet model is divided into two components: the binary cross-entropy (BCE) loss (L_{bce}), and the dice loss (L_{dice}). It can be expressed as:

$$Loss(p,\hat{p})=\alpha L_{bce}(p,\hat{p})+\beta L_{dice}(p,\hat{p}) \tag{9}$$

p and $\hat{p}$ denote the ground-truth map and prediction map, and the parameters α and β represent the weight coefficients of L_{bce} and L_{dice},

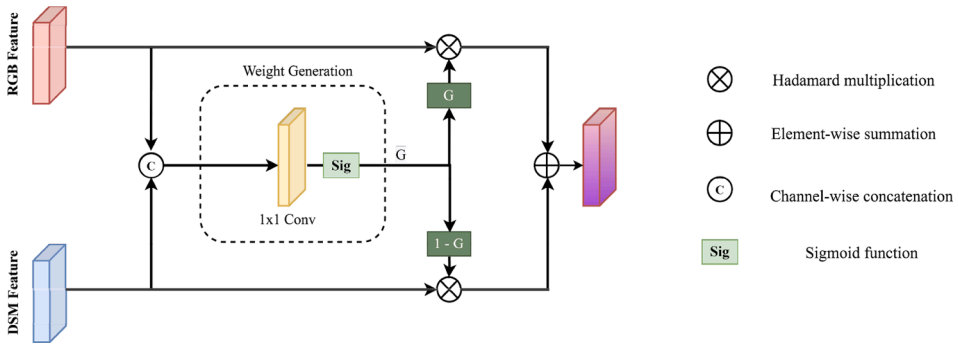

Fig. 3. The structure of the Gated Fusion Module.

respectively, to show their importance. These parameters are set to 1 in this study. The purpose of this weighting is to ensure the equal role of each loss function in the process of network training. Now, each of these cost functions, L_{bce} and L_{dice}, will be examined. The difference between two probability distributions for a random variable is measured by the cross-entropy loss function (Ma, 2020). This loss function is commonly utilized in pixel-level segmentation and classification objectives. The RGB, DSM, and fused prediction maps are all monitored simultaneously to completely leverage the information of various modalities. Accordingly, L_{bce} consists of three separate parts. The BCE loss for RGB stream, The BCE loss for DSM stream, and The BCE loss for fused stream.

$$L_{bce}(p,\hat{p}) = L_{rgb}(p,\hat{p}) + L_{dsm}(p,\hat{p}) + L_{fused}(p,\hat{p}) \tag{10}$$

For each loss function the BCE loss is defined as:

$$L_i(p,\hat{p}) = -(p\log(\hat{p}_i) + (1-p)\log(1-\hat{p}_i)) \tag{11}$$

The subscript 'i' indicates a modality, which could be RGB, DSM, or final fused features. The number of background pixels in VHR remote sensing images is usually more than the number of building objects. As a result, while employing simply BCE loss in the training phase, there is a class-imbalanced problem in the building extraction task. To overcome this problem, another loss function based on dice coefficient in addition to the BCE loss function is used in equation 8. The well-known dice overlap coefficient is a metric used to determine the similarity between two images (Milletari et al., 2016). This coefficient was adopted as a regional loss function and outperform BCE loss in class-imbalanced problems. The calculation of L_{dice} is defined as follows:

$$L_{dice} = 1 - \frac{2|p \cap \hat{p}_{fused}| + \varepsilon}{|p| + |\hat{p}_{fused}| + \varepsilon} \tag{12}$$

where $\hat{y}_{fused}$ denotes fused prediction map. Furthermore, ε is summed to the denominator and numerator to ensure that the loss function is not undefined in the edge case.

3. Experiment

3.1. Dataset

Three datasets with high spatial resolution have been used to evaluate the CMGFNet model in this research. An essential feature of these datasets is the availability of DSM data along with VHR images. Potsdam and Vaihingen datasets, which are provided by Commission II/4 of the ISPRS[1]. The third dataset[2] is published by United States Geological Survey (USGS) and named the USGS dataset in this paper. Fig. 6 shows examples of images related to the training set of these datasets.

3.1.1. Potsdam dataset

The Potsdam dataset consists of 38 true orthophotos (TOP) with corresponding DSM patches. Both the TOP and the DSM have a spatial resolution of 5 cm and 6000 × 6000 pixels. These patches are presented in two modes, including red (R), green (G), blue (B) bands (Potsdam-RGB), and near-infrared (NIR), red (R), green (G) bands (Potsdam-IRRG). The original ground-truth contains six major land cover classes: impervious surface, buildings, low vegetation, trees, cars, and clutter. In this paper, the foreground and background classes are used to classify the building and the rest of the objects, respectively. According to the data provider, the training set consists of twenty-four patches, with the remaining patches serving as the test set.

3.1.2. Vaihingen dataset

In the Vaihingen dataset, each patch is cut from a sizeable TOP associated with the town of Vaihingen (Germany). The dimensions of the patches are not equal to each other and are approximately 2000 × 2500 pixels in size, with a spatial resolution of 9 cm. Each patch contains only three bands (Vaihingen-IRRG): near-infrared (NIR), red (R), and green (G). In addition, a corresponding digital surface model is provided for each image patch. The classes in the ground-truth are the same as the Potsdam dataset. According to the data provider, sixteen patches from all patches are used as the training set and the remaining patches for the test.

3.1.3. USGS dataset

The USGS dataset includes high-resolution orthophotos with spatial resolutions ranging from 0.15 m to 0.3 m, LiDAR point clouds, and ground-truth building masks. Like research (Huang et al., 2019), fourteen orthophoto patches related to five cities are selected in the United States. The dimensions of the patches are 5000 × 5000 pixels, and it has near-infrared (NIR), red (R), green (G), and blue (B) bands. In this paper, only RGB bands (USGS-RGB) are used. DSM data is prepared from LiDAR points cloud with LasTools software in the '.CSV' format by the data provider. Then, this height information in the '.CSV' format is rasterized to the form of patches in the '.tiff' format. Ground-truth data related to buildings are extracted and processed from Open Street Maps (OSM) service and are resampled to their original resolution. One patch from each city is selected as test data, and the remaining patches are used in network training. According to the report (Bradbury et al., 2016), the selected data from different cities have a spatial resolution of 30 cm with various changes in altitude information. This diversity is evident in the dimensions of buildings and their density in urban and non-urban environments. Therefore, the USGS data are ideal for evaluating and understanding the power of the proposed method.

3.2. Evaluation metrics

Deep learning models are typically evaluated for quality and performance by analyzing their performance in test data. In this paper, the

[1] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[2] https://doi.org/10.6084/m9.figshare.3504413.

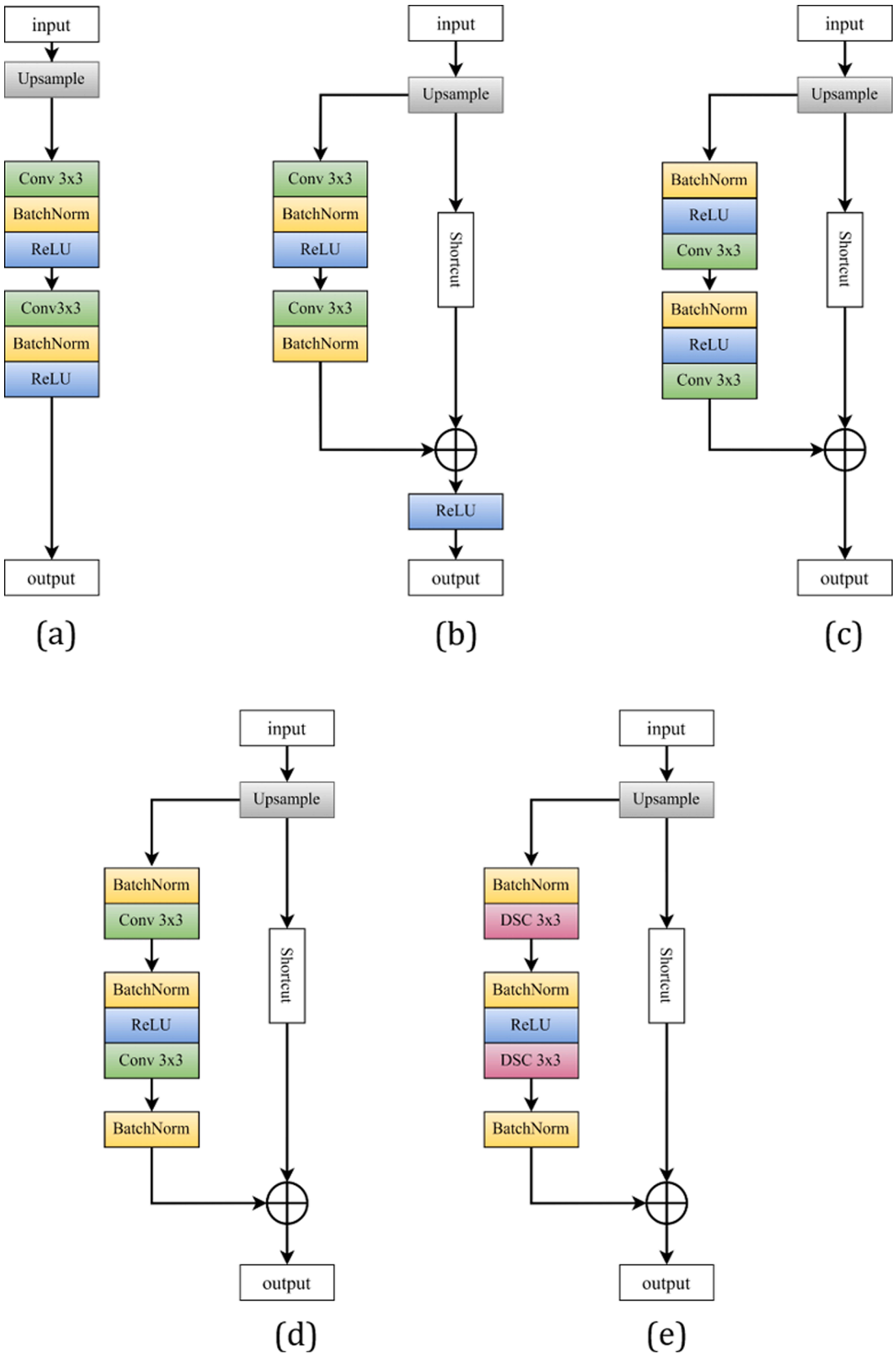

Fig. 4. Various types of decoder blocks. 'BatchNorm' denotes a BN layer, and The term 'shortcut' refers to a 1 × 1 convolution layer that converts the number of feature channels in the input to the desired number of channels. (a) Conventional decoder block without any RCU (CD). (b) Decoder block based on the original residual unit (RCUD), (c) Decoder block based on the original pre-activation RCU (pre-RCUD), (d) Decoder block based on the pre-activation RCU with a BN layer after the last convolution operation and removing the first ReLU (modified pre-RCUD), and (e) The proposed decoder block based on (d) witch traditional convolution layer replaced with depth-wise separable convolutional (R-DSC).

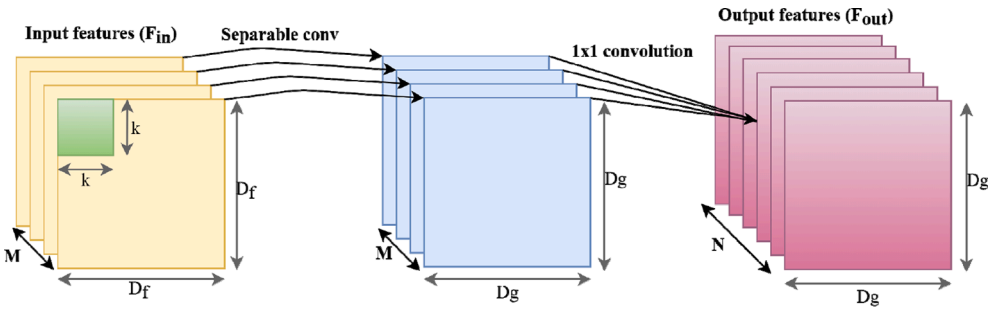

Fig. 5. The architecture of depth-wise separable convolution. There are two phases to this procedure: separable and 1 × 1 convolution.

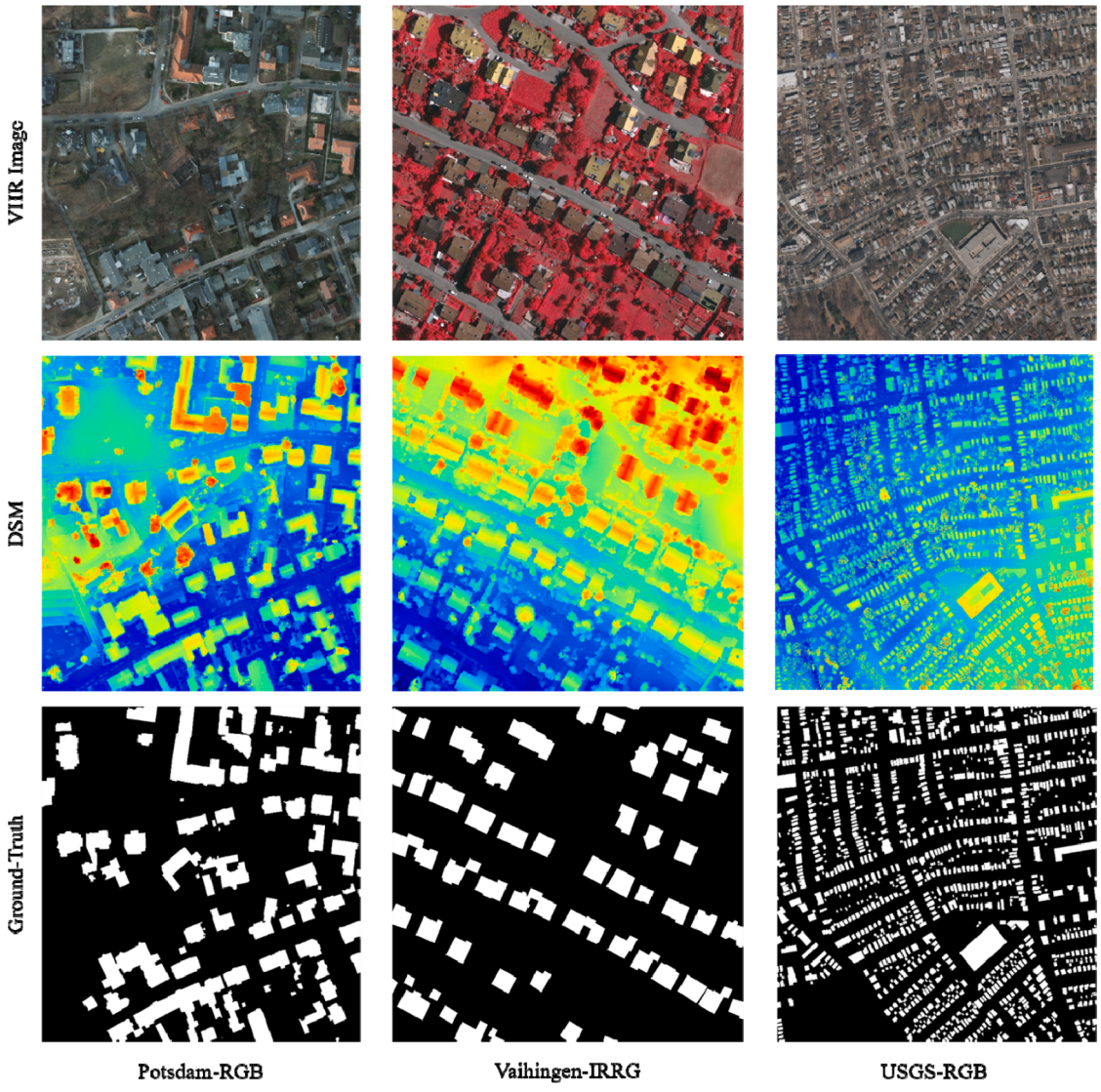

Fig. 6. Examples for the training image sets, the corresponding DSM, and the ground-truth, respectively. The ground-truth of the Potsdam and Vaihingen challenge includes six categories, in which only the building objects are used in this research, and the rest are considered the background.

proposed network is evaluated using three popular metrics: *overall accuracy (OA)*, *F-score*, and intersection over union (IoU). Typically, the overall accuracy, also known as pixel accuracy, is used to evaluate segmentation performance. This metric is the ratio of the number of successfully predicted pixels to the total pixels in all patches of the test datasets, and it is calculated as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

TP, TN, FP, and FN denote true-positive, true-negative, false-positive, and false-negative, respectively. The *Precision* measures the accuracy for the minority class (building object class) and focuses on the correct positive predictions out of all positive predictions. Unlike the *Precision*, the *Recall* gives an offer for missing positive predictions. *F-*

score offers the harmonic mean for both *Precision* and *Recall*:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F} - \text{score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{16}$$

The IoU, commonly known as the Jaccard index, is a metric that measures how much the ground-truth map and prediction result overlap. IoU is defined as follows:

$$IOU\left(\mathscr{P}_p, \mathscr{P}_{gt}\right) = \frac{\left|\mathscr{P}_p \cap \mathscr{P}_{gt}\right|}{\left|\mathscr{P}_p \cup \mathscr{P}_{gt}\right|} \tag{17}$$

$\mathscr{P}_{\text{p}}$ represents the set of prediction pixels, $\mathscr{P}_{\text{gt}}$ denotes the set of ground-truth pixels, and $|\cdot|$ represents the number of members in a set. '$\cup$' and '$\cap$' indicate the union and intersection of two sets, respectively.

The graph of Precision-Recall (PR) is presented to compare the output of binary segmentation in different state-of-the-art models. The PR curve depicts the trade-off between Precision and Recall rates, with varying thresholds of probability. The prediction map is binarized, with thresholds ranging from 0 to 1, and then compared to the associated ground-truth to determine Precision and Recall values. Better model performance is indicated by a larger area under the PR curve.

3.3. Method implementation

In this paper, The popular Pytorch framework is used in all experiments (Paszke et al., 2019) on a system with a single Tesla K80 GPU. Data augmentation techniques are instrumental in avoiding the overfitting problem and are used to increase the training samples artificially when reading them from memory in each epoch. These methods include rotating input image, DSM and ground-truth randomly in the step of 90° both horizontal and vertical directions at 90° from 0° to 270°, and then randomly flipped it vertically and horizontally, respectively. The training dataset is cropped to 640 × 640 pixels due to the original size of the input data and the limitation of GPU memory. The network training procedure uses about 80% of each dataset at random, while the network evaluation step uses the remaining 20%. AdaMax optimizer, known as a type of Adam optimizer (Kingma and Ba, 2014), with a weight decay of 0.0009, an initial learning rate of 0.001, is used to train the network. A '*poly*' policy is used to optimize the learning rate of the networks. In this method, the initialized learning rate is multiplied by $\left(1 - \frac{\text{iter}}{\text{max}-\text{iter}}\right)^{\text{power}}$. In this experiment, the '*power*' is set to 0.3, and the '*max-iter*' can be computed by multiplying the number of epochs with whole batches in each dataset. The parameters of the pre-trained ResNet-34 network are used to configure for both the RGB and DSM encoder in the CMGFNet. The remaining parameters in the model are initialized using the methods described in (He et al., 2016). The training continues until validation loss converged and the best result of the parameter in each iteration is stored. The patches with 50 percent overlap are generated for each test in the inference phase without data augmentation. In addition, for the segmentation results, no post-processing is implemented. The final result is acquired by integrating all probability maps, and the mean values are used to generate the final prediction values. The BCE loss value in each of the training and validation stages is shown in Fig. 7 for three datasets. In each epoch, the entire dataset is run, and the training speed is about 3.07 s per batch.

3.4. Ablation experiment

The efficiency of the CMGFNet for building extraction is investigated in this section. We perform ablation experiments by removing and changing each critical component of the network independently. The CMGFNet is named 'Model' to make it easier distinguishing between the models that are compared to it. At First, the structure of cross-modal fusion is examined. Then the effect of different fusion methods in

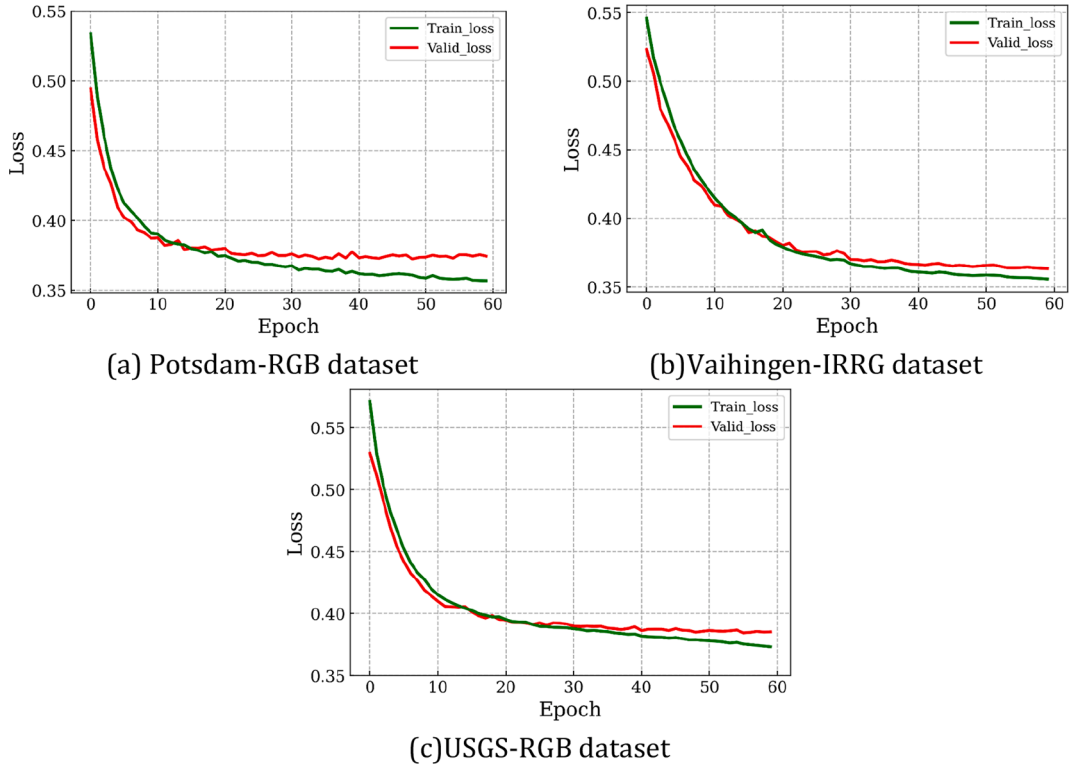

Fig. 7. Training and validation loss of the CMGFNet method.

cross-modal, multi-level feature fusion, the R-DSC module, and the combination of loss functions are investigated. Note that all settings and metadata for training and validation processes are fixed in all ablation experiments.

3.4.1. Cross-modal fusion

The CMGFNet is compared to two models with popular structures in this section to demonstrate the capability of cross-modal fusion. In the first comparison, The DSM stream is completely removed from the CMGFNet, and the new model is trained based only on RGB images. This model is denoted as 'Model–DSM'. The purpose of this proceeding is to investigate the role of using DSM data in the final results of building extraction. In the second comparison, the model is trained based on two separate RGB and DSM streams without any feature-level fusion in cross-modal. Then the output feature from the last layer of each RGB and DSM stream is fused at the decision level. This model is denoted as 'Model–CM', which is meant the CMGFNet model with no cross-modal fusion. The last model tested in this section is based on the original CMGFNet model (denoted as 'Model') that RGB and DSM features are fused in cross-modal with GFM block.

The BCE loss is used in all experiments of the network structure to ensure equity. As shown in Table 1, the comparison between the third, fourth and fifth columns indicates that the 'Model' improve the result of F-score and IoU metrics in all three datasets significantly. The 'Model–DSM' does not perform as well as the other models. This model does not consider the DSM feature, which helps analyze the building object against a complex context. The improvement in the IoU score of the 'Model' compared with the 'Model–DSM' are 2.92% (Potsdam-RGB), 3.25% (Vaihingen-IRRG), and 6.84% (USGS-RGB). The results in the USGS dataset are better improved than the Potsdam and Vaihingen dataset by using the DSM data. The reason for this is that the USGS dataset has a lower spatial resolution than the others, and smaller buildings are extracted with lower accuracy in 'Model–DSM'. The comparison of 'Model–CM' and 'Model' reveals that the model employs a cross-modal fusion in the DSM decoder from the RGB encoder outperforms the model that does not utilize a cross-modal module. The improvement in the IoU score of 'Model' compared with 'Model–CM' are 1.41% (Potsdam-RGB), 1.47% (Vaihingen-IRRG), and 2.1% (USGS-RGB). These encouraging results show that guiding the side-output feature from the RGB stream is essential.

To better determine the practicality of the suggested strategy, the building extraction results of several methodologies are evaluated from qualitative perspectives. Fig. 8 shows the building extraction results of different methods for three datasets. The proposed 'Model' obtained the most visually consistent results when compared to the ground-truth building maps. The 'Model', compared with two other models, is robust for extracting buildings in various complex scenes, significantly when the height parameter of non-building objects influences the final results, as shown in Fig. 8. For instance, in the test image selected from the Vaihingen-IRRG dataset in the first row of Fig. 8, a sports field with a texture similar to the roof of buildings can be seen in the red square. In the 'Model–DSM' method, since the DSM data is not used, the model cannot distinguish the sports field from building objects only by relying on its texture information. Similar behaviours with the increase in the complexity and structure of buildings are observed in selected images from Potsdam and USGS datasets. This inability to detect the building from non-building objects is not seen in the other two methods in which DSM data are used. The proposed 'Model', on the other hand, outperforms the 'Model–CM' in terms of recognizing buildings and their boundary location. The reason for this is that in the 'Model–CM', due to the low quality of the DSM information, the generated features in the DSM stream without fusion of the side-output features from RGB stream do not provide convenient features for improving the extraction of buildings. Based on this experience, it can be understood the importance of using DSM information and fusion with an RGB feature in cross-modal to achieve a better model. Close-ups of the selected regions in the tested images are shown in Fig. 9. These highly detailed results indicate the high power of the CMGFNet method for extracting the building object from VHR remote sensing images.

3.4.2. Effects of GFM on cross-modal fusion

This section seeks to confirm the practicality of the proposed GFM in cross-modal fusion. Besides GFM, two traditional approaches are employed to investigate the fusion of RGB and DSM features in cross-modal based on element-wise summation and channel concatenation. The behaviour of these two operators is similar to each other. However, if the two modalities are not very closely related, concatenating might be appropriate. The sixth to eighth columns of Table 1 show the quantitative results of cross-modal fusion. The SUM and CAT subscriptions in the sixth and seventh column model names denote that the RGB and DSM feature fusion methods are based on element-wise summation and concatenation, respectively. The results show a significant boost in evaluation metrics if GFM is used in the proposed model. However, the results of the two models, '$Model_{SUM}$' and '$Model_{CAT}$', are slightly different from each other.

For a deeper look at GFM actions, the colour information of the square part of the RGB image is hidden. This part includes building area and background information (Fig. 10c). The prediction maps are shown in Fig. 10d to Fig. 10g. In Fig. 10d, the prediction map is related to using the original image in the CMGFNet model. Fig. 10e and g represent the prediction maps of the '$Model_{SUM}$' and '$Model_{CAT}$', respectively. Since RGB information is not available in the hidden part, the fusion of DSM data in this part based on an element-wise summation and concatenation methods has not helped to distinguish building areas from the background. Fig. 10g represents the prediction map of the proposed method. The visual result shows the power of GFM to determine the building from the background. This module produces features that contain more discriminative information about building objects. In addition, the G matrix, according to Equation 6, which is considered the weight matrix, is shown in Fig. 10h. The weight matrix is related to the final GFM in the proposed method, and each element of it is averaged over the entire weight tensor. It can be concluded from the results that the weights in the RGB features are low just in the hidden area and high in the remainder of the area. On the contrary, for the hidden area, the weights for the DSM features are increased. Eventually, these weights are multiplied by the RGB features to decrease the contribution of the

Table 1
Evaluation metrics score (%) of ablation experiment for different models and losses.

Datasets	Metrics	Model–DSM (l_{BCE})	Model–CM (l_{BCE})	Model–SC (l_{BCE})	$Model_{SUM}$ (l_{BCE})	$Model_{CAT}$ (l_{BCE})	Model (l_{BCE})	Model ($l_{BCE}+l_{Dice}$)
Potsdam-RGB	OA	97.48	97.89	98.22	98.18	98.21	98.24	**98.36**
	F-score	96.34	96.99	97.49	97.44	97.48	97.37	**97.50**
	IoU	89.76	91.27	92.66	92.51	92.63	92.68	**92.80**
Vaihingen-IRRG	OA	96.03	96.55	96.91	96.81	96.90	96.96	**96.91**
	F-score	94.57	95.26	95.78	95.64	95.71	95.84	**95.96**
	IoU	85.28	87.06	88.38	88.01	88.15	88.53	**88.84**
USGS-RGB	OA	94.97	95.86	96.26	95.97	96.07	96.27	**96.32**
	F-score	90.61	92.41	93.12	92.64	92.76	93.18	**93.24**
	IoU	72.95	77.69	79.55	78.55	78.67	79.79	**79.96**

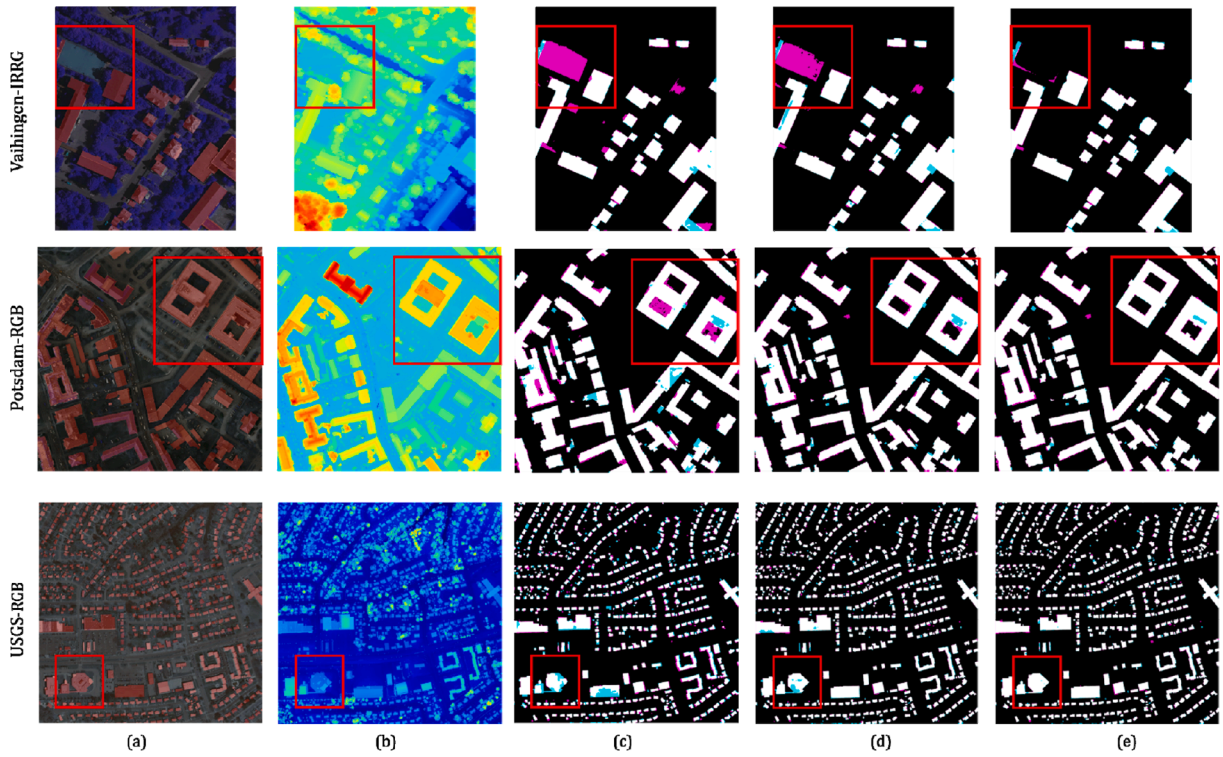

Fig. 8. Selected images from each dataset and the results of building extraction produced by different models. (a) Original image with the corresponding label. (b) DSM. (c) Model–DSM. (d) Model–CM. (e) CMGFNet (proposed). TP, FP, and FN are marked in white, cyan, and pink, respectively. The red rectangles represent the selected regions for close analysis in Fig. 9. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

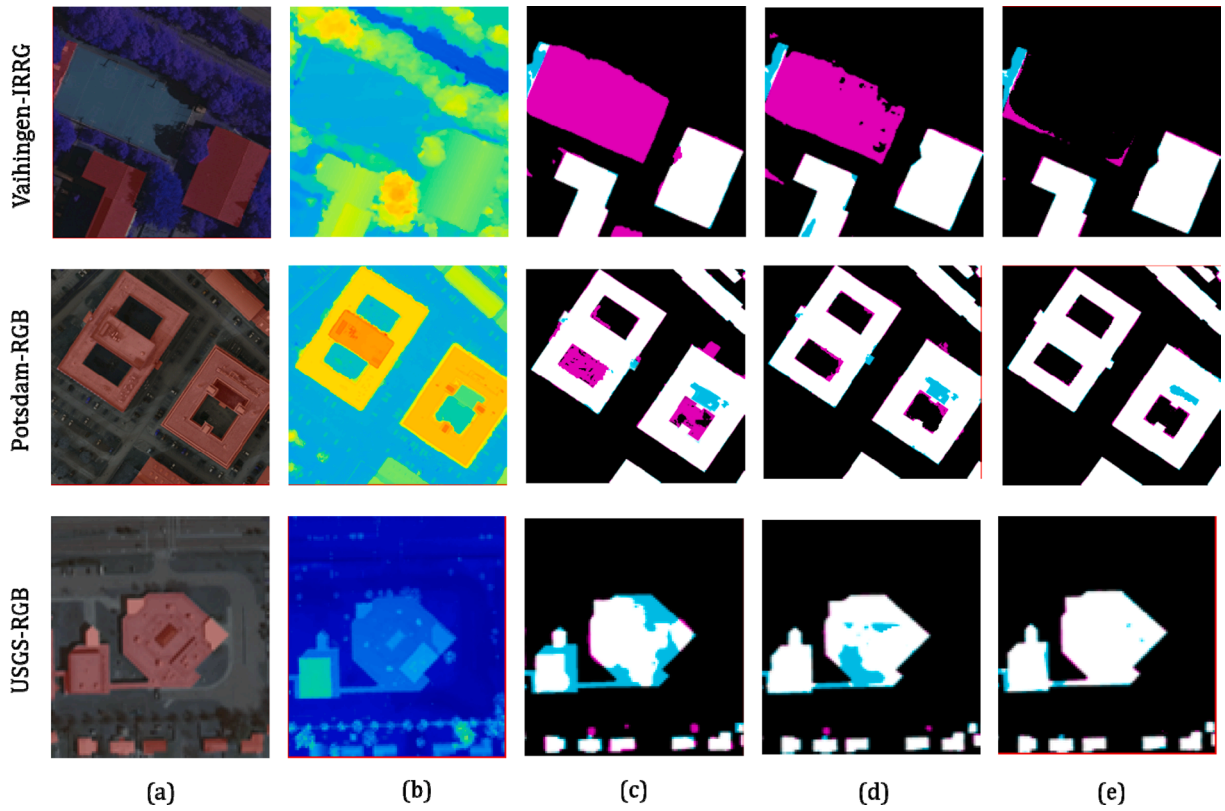

Fig. 9. Close-up views of results obtained using various models. The images and results in (a–e) are a subset of the regions highlighted in Fig. 8. (a) Original image with the corresponding label. (b) DSM. (c) Model–DSM. (d) Model–CM. (e) CMGFNet (proposed). TP, FP, and FN are marked in white, cyan, and pink, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

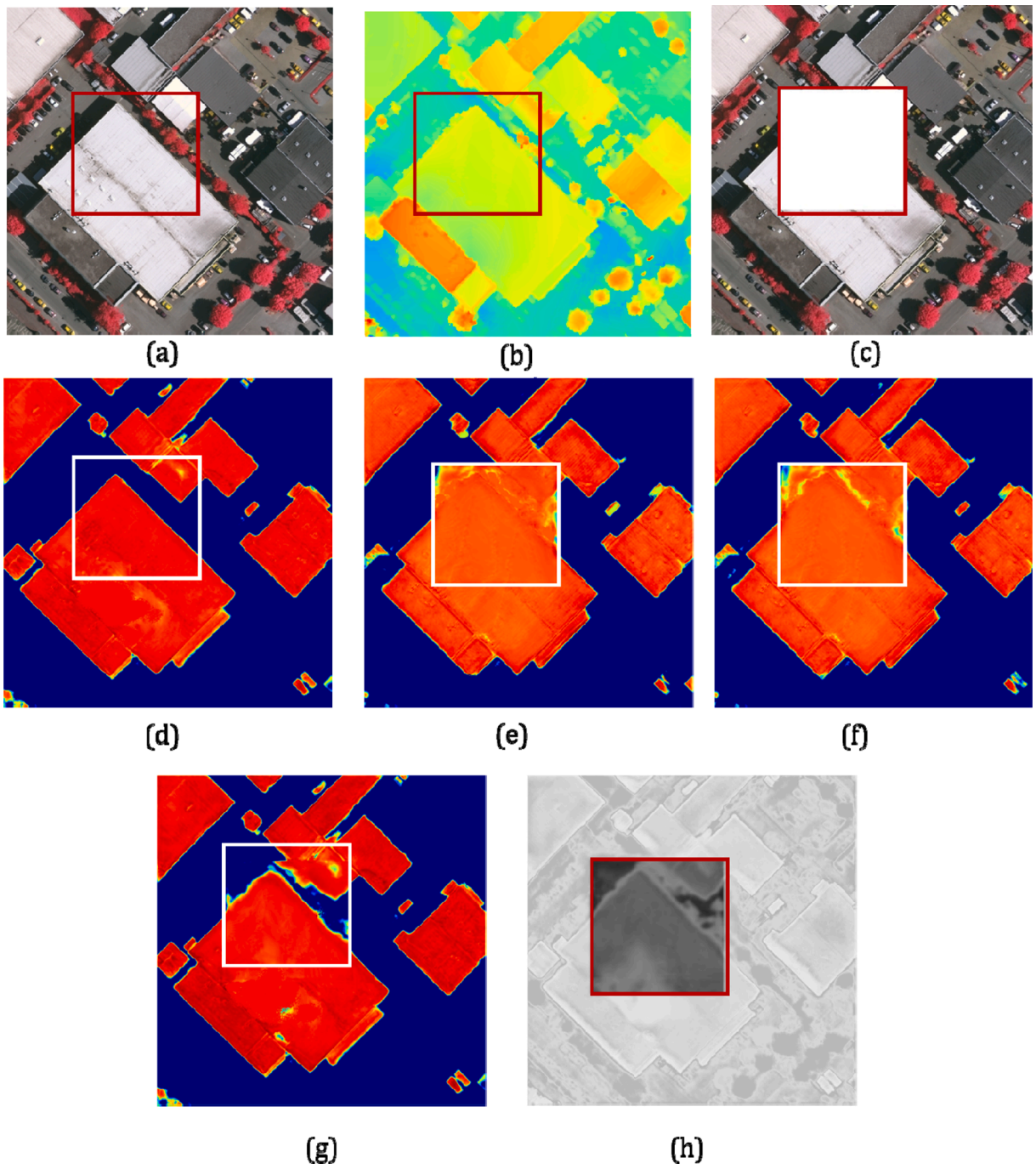

Fig. 10. Evaluation of the performance of GFM in the proposed method. (a) Original image; (b) DSM; (c) Image with the blank area. (d) Prediction map of the CMGFNet model on the original image. (e) Prediction map of $Model_{SUM}$ when using the image with the blank area. (f) Prediction map of $Model_{CAT}$ when using the image with the blank area. (g) Prediction map of the CMGFNet when using the image with the blank area. (h) The visualization of the GFM weight maps (G matrix) at the last layer.

blank area. The proposed GFM has merged information from two modalities according to the quality of functionality for each interested region.

3.4.3. *Effect of the multi-level feature fusion*

In this section, the effect of multi-level feature fusion in the proposed network is examined, and the improvement of the results is discussed. As mentioned in Section 3.2, the CMGFNet contains multi-level feature fusion, which high-level feature with low spatial resolution fused to the low-level feature with high spatial resolution in both RGB and DSM stream using the short connection (SC). To evaluate the performance of the multi-level feature fusion, short connections were removed from the original CMGFNet model. We denote this model as 'Model–SC'. In addition, the prediction map is calculated at each stage of the decoder for the DSM stream. In Table 2, the comparison of evaluation metrics of the seventh and eighth columns shows the benefit of using these extra feature fusion in the CMGFNet. To investigate the effect of multi-level feature fusion, the prediction map of each DSM decoder is determined. In Fig. 11, two test images from the Vaihingen-IRRG dataset are shown. The prediction map is expressed by P_i, which i is related to the decoder number. The final prediction map denotes P_{fusion}. Columns (a) and (b) represents the prediction maps of the 'Model–SC' and the CMGFNet

Table 2
Evaluation metrics score (%) of ablation experiment for different decoders.

Datasets	Metrics	Model + CD	Model + RCUD	Model + pre-RCUD	Model + Modified pre-RCUD	Model + R-DSC
Potsdam-RGB	OA	97.76	97.98	98.14	98.17	**98.36**
	F-score	96.37	96.94	97.13	97.17	**97.50**
	IoU	90.82	91.2	92.53	92.61	**92.80**
Vaihingen-IRRG	OA	96.08	96.56	96.86	96.88	**96.91**
	F-score	94.80	95.07	95.59	95.68	**95.96**
	IoU	87.07	87.87	88.03	88.19	**88.84**
USGS-RGB	OA	95.48	95.86	96.02	96.17	**96.32**
	F-score	91.37	91.88	92.98	93.07	**93.24**
	IoU	78.51	78.68	79.73	79.84	**79.96**

model ('Model'), respectively. The output result of columns (b) shows that prediction maps of the building contain more semantic information than in Columns (a), which do not incorporate the semantic information of the deeper layers. In other words, it can be concluded that adding more semantic information to low-level features could increase the performance of the proposed model and avoiding the loss of information.

3.4.4. Effects of R-DSC module

In this section, to evaluate the proposed R-DSC module, according to Fig. 4, the CMGFNet model is tested with different decoders. In terms of time complexity per epoch and IoU results, Fig. 12 compares the decoder parameters of standard convolution with depth-wise separable convolution for the Vaihingen-IRRG dataset. Experiments show that the CMGFNet model, which uses the R-DSC in decoder parts of RGB and DSM stream, is superior to the same models, which only use standard convolution. In addition, the comparison between different residual-like decoders shows that the modification in the location of ReLU and BN improves the training process and achieves better results in the building extraction task. More extensive semantic information about building objects from the deep convolutional blocks is addressed to the primary layer during the up-sampling action due to the modified pre-activation RCU. As shown in Table 2, the comparison between different decoder shows that the proposed R-DSC improve the building extraction result of F-score and IoU metrics in all three datasets.

3.4.5. Different loss function

The most basic loss for the CMGFNet is BCE loss, which is supervised on P_{rgb} and P_{fusion}. Furthermore, to deal with the class-imbalance issue in segmentation of the building object, dice loss is utilized, which is supervised only on P_{fusion}. The CMGFNet model is trained in two modes, with and without dice loss, to see how adding Dice loss affected network training. In Table 1, the comparison between the eighth and ninth columns shows the effect of the combination of dice loss with BCE loss. The result obtained by combining BCE and dice losses as the loss function is superior in evaluation metrics.

The BCE loss is a distribution-based loss. This category of loss function works best when the data distribution between classes is equal. As mentioned above, the distribution of pixels involving building features is not the same as background pixels. This causes the network training process to get stuck into the local optimum if the learning rate parameter is fine-tuned. For this reason, using dice loss can help with this problem. Dice loss is based on Region-based loss functions. The goal of the dice loss function is to maximize the overlap regions or minimize the mismatch between the prediction map and the ground-truth. Thus, the combination of dice loss with BCE loss could better clarify the issue of class imbalance while improving the result of the building extraction.

4. Discussion

4.1. Comparisons with state-of-the-art single modal networks

A variety of single modal DCNN models based exclusively on three-channel input images have been introduced in recent years. These models improved segmentation accuracy compared to traditional methods. However, one of the significant limitations of them is the impossibility of using additional modality during model training. In this section, four state-of-the-art FCN models, such as SegNet (Badrinarayanan et al., 2017), PSPNet (Noh et al., 2015), Res-Unet (Xu et al., 2018), and Deeplapv3+ (Chen et al., 2018), were used to evaluate further the effectiveness of the RGB stream of the CMGFNet methods. These methods were chosen because they have all been shown to be effective in building extraction, and they are all open source and simple to use. RGB stream of the CMGFNet, similar to Section 3.4.1, is obtained by removing the DSM stream and is named 'Model–DSM'. On the other hand, the Resnet-34 has been selected as the encoder in all networks to ensure fairness in the results. The experimental results are shown in Table 3. For all datasets, it can be seen that our method outperforms other methods in IoU. The number of parameters for each model is shown in the fourth column of Table 3. The lower number of training parameters of the proposed model is specifically related to the R-DSC block presented in this research. When compared to traditional methods, using depth-wise separable convolution reduces not only the number of training parameters but also the multiply-and-accumulates (MACs) of the model. It is important to indicate that the PSPNet has only one up-sampling operator with superficial convolution layer in the decoder part compared to other methods, hence its MACs are less than other models. In addition, the multi-level feature fusion and residual-like unit in the R-DSC block retain more features and have made the proposed method superior to other models. The P-R curves shown in Fig. 13, denotes the Deeplapv3+ and 'Model–DSM' perform better in the building extraction has a high Precision and Recall rate.

Fig. 14 shows the building extraction result of some test images. Visually, Deeplabv3+ and 'Model–DSM' methods have similar outputs, but the model suggested in this study differs significantly from Deeplabv3+ in terms of both the number of training parameters and the cost of computing time. As a result, it is more efficient to use the proposed method in multimodal structures.

4.2. Comparison between different encoder networks

As described in Section 3.1, the encoder part of the CMGFNet model can be easily replaced with other classification networks. Due to the limited GPU memory, several classification networks similar to ResNet-34, such as ShuffleNet_v2_x1_0 (Zhang et al., 2017b), VGG-16bn (Simonyan and Zisserman, 2014), and DenseNet-121 (Huang et al., 2017), are compared as the primary encoder networks of the CMGFNet to show that the ability of proposed ResNet-based encoder. All of these networks have been initialized with ImageNet pre-trained weights. Table 4 shows the quantitative test results of the CMGFNet for three datasets using different encoders. In terms of IoU score, ResNet outperforms other encoders. In this comparison, the CMGFNet with ShuffleNet-based encoder has the lowest IoU score. This lightweight network uses a depth-wise separable convolution module and requires 9.46 G multiply-and-accumulates (MACs) per image. The CMGFNet model with a VGG-based encoder has high MACs and computational times per image than the other encoders due to its complex CNN structure. The DenseNet network, similar to the ResNet network, uses the residual block in its architecture. The DenseNet network aims to decrease the parameters while increasing the number of layers. However, due to the complex structure of dense connections between different blocks in this architecture, the computations are higher than the use of the ResNet network in the training and validation process.

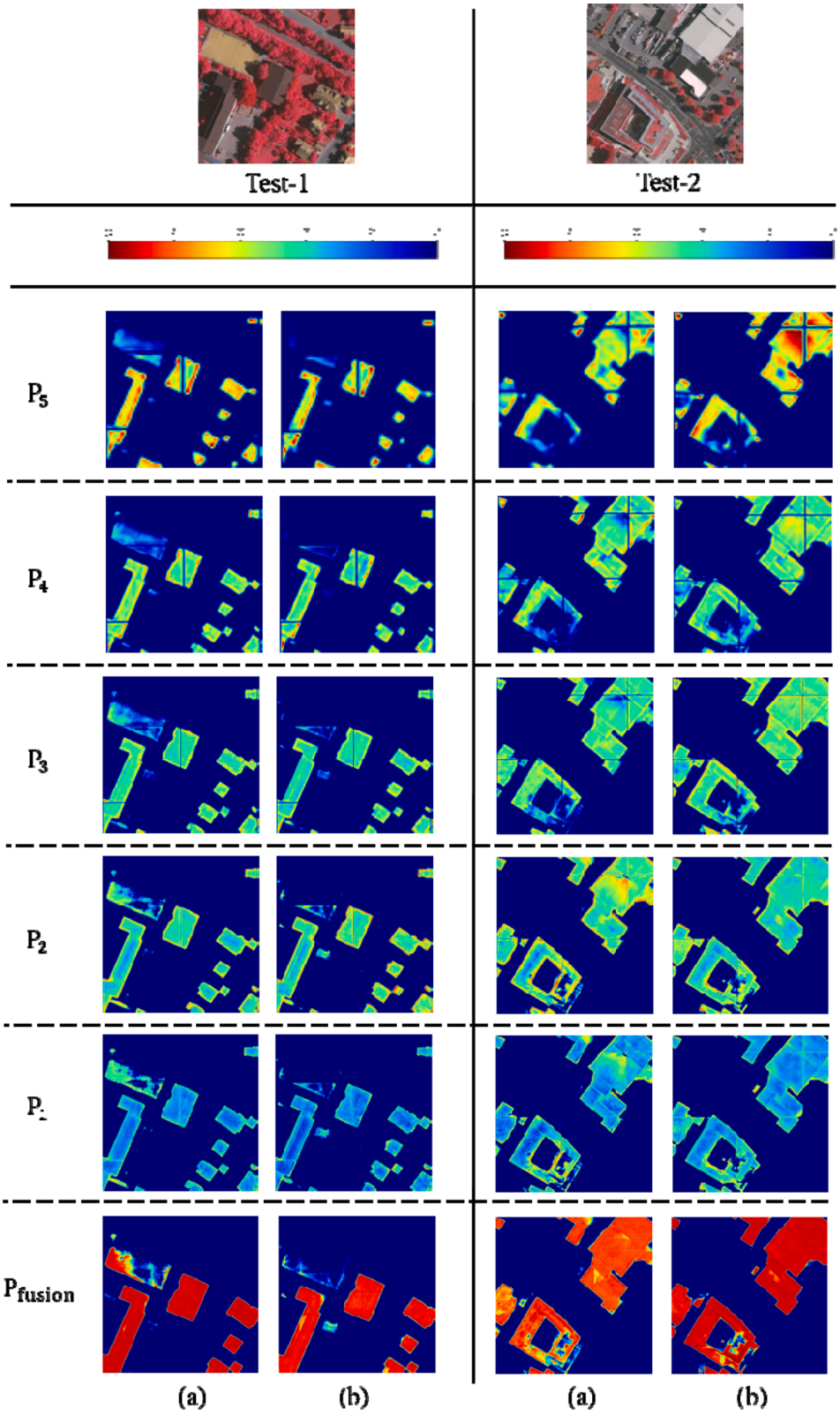

Fig. 11. The effect of multi-level feature fusion on different layers of prediction maps correspond to each DSM decoder. (a) Model–SC; (b) The CMGFNet model.

4.3. Comparisons with state-of-the-art fusion networks

for further evaluation of the proposed model, Three state-of-the-art deep-learning-based fusion models are compared to the CMGFNet, including FuseNet (Hazirbas et al., 2017), V-FuseNet (Audebert et al., 2018), and HAFNet (Zhang et al., 2020a). Note that the encoder-decoder part of all these state-of-the-art networks is built on SegNet (Badrinarayanan et al., 2017), and all of them use VGG-16bn architecture in the encoder part. Then, to guarantee that the comparison result is fair, the encoder part of the CMGFNet changed from ResNet34 to VGG-16bn

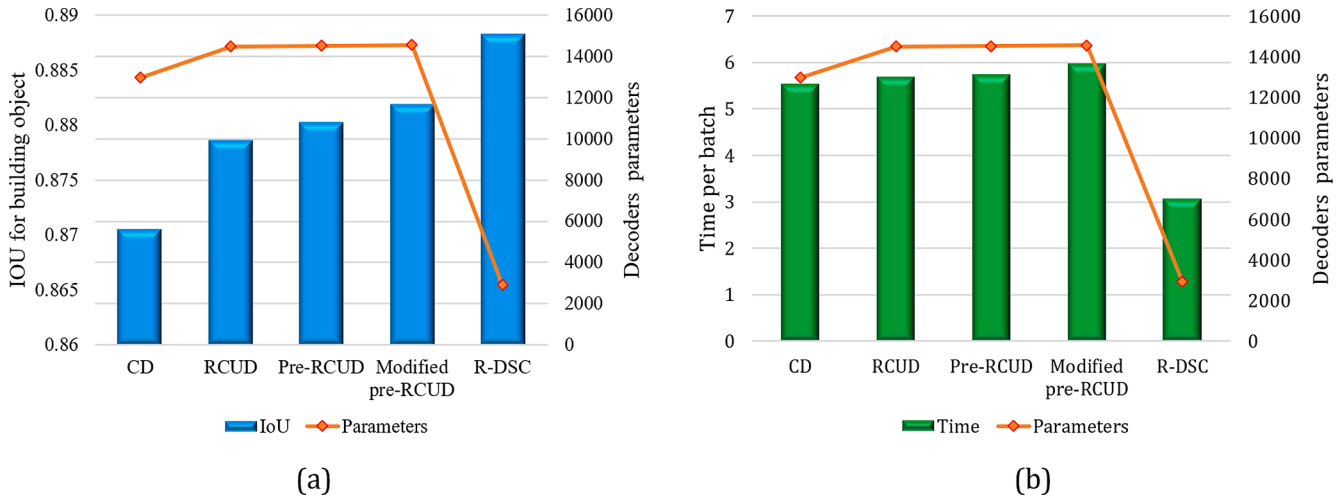

Fig. 12. (a) Comparison of the IoU Metric with the number of parameters at different decoders. The bar diagram illustrates the IoU on the Vaihingen-IRRG dataset. (b) Comparison of training time per batch and the number of parameters at different decoders. The line chart for both diagrams represent the number of parameters of different decoders.

Table 3
Comparison of different state-of-the-art single modal networks with proposed method.

Model	IoU(%)			Params(M)	MACs(G)	Image/S	
	Vaihingen-IRRG	Potsdam-RGB	USGS-RGB			Train-mode	Test-mode
SegNet	84.64	88.60	68.85	29.44	151.11	1.43	0.246
PSPNet	83.99	88.27	67.40	21.87	17.47	4.2	0.145
Res-UNet	84.80	89.30	70.81	29.94	92.68	1.08	0.198
DeepLabv3+	85.06	89.61	71.54	26.01	341.73	0.83	0.301
Model-DSM	85.28	89.76	72.95	21.84	51.84	2.38	0.178

network. Also, the same training, validation, and testing sample is used to train and test for all models, and pre-trained coefficients are adopted to start training.

Table 5 shows the quantitative experimental result of different models. In all three datasets, the CMGFNet model outperforms the other models in both IoU and F-score. As shown in Fig. 15, the P-R curve of the proposed method has a high Precision and Recall rate compared to other approaches.

Fig. 16 shows the building extraction results of some selected buildings. The CMGFNet outperformed the other approaches. In the first to third rows of Fig. 16, the selected buildings have a regular geometric shape. The proposed method and other methods have a high ability to distinguish such buildings. However, the proposed method produced a few FN (pink) and FP (cyan) in building extraction results. On the other hand, visual results show, as the complexity of buildings is increased in VHR aerial images, the FuseNet, V-FuseNet, and HAFNet networks are unable to detect these types of buildings. These complexities include differences in the sizes and shape of buildings, hidden areas, and changes in lighting conditions and textures. For instance, in the fourth and fifth row of Fig. 16, the selected sample of the building has a large size with a complex shape. Because of the large size of the building, the type of material on the roof varies greatly, causing changes in texture on the building's roof, particularly in VHR aerial images. However, the CMGFNet model has a higher ability than other methods to detect this type of building completely. In rows sixth and seventh of Fig. 16, the selected buildings have different shapes and scales. In these cases, the details in the building footprints are great, and the ground-truth of the data is not accurate. In addition, DSM data has a lot of ambiguity in the edge areas of buildings and does not help much in improving the extraction of building footprints. However, the results show that the CMGFNet accurately identified irregularly shaped buildings with different scales.

4.4. Generalization ability and applicability of the proposed model

The generalization ability of DCNNs-based methods is vital for automation, but its performance in remote sensing applications has been disappointing (Ji et al., 2019). This is because that a source dataset is significantly different from a target dataset. For instance, these differences could include: differences in the building features of different areas, other combinations of color spectrums, different distribution of buildings in urban and non-urban areas, and differences in the spatial resolution of images. In this section, experiments on both transfer learning and fine tuning are evaluated further to examine the generalization ability of the proposed CMGFNet. The CMGFNet was trained using the two modes of the Potsdam dataset (RGB and IRRG images), the Vaihingen-IRRG dataset, and the USGS-RGB dataset with a spatial resolution of 0.05, 0.09, and 0.3 m/pixel respectively. The IoU score after 60 epochs was obtained 92.80% and 92.73% for RGB and IRRG test images in the Potsdam dataset, 88.84% in Vaihingen-IRRG, and 79.96% in the USGS-RGB test image dataset, respectively. In transfer learning, the trained network was tested on the target datasets without considering any augmentation on the spectral band. Table 6 illustrates the test results for several datasets. As shown in the third column of Table 6, when different dataset with similar color spectrums combination is tested, the CMGFNet obtained the high IoU score compared with the other variety of color spectrums of VHR images. Another point to consider is the spatial resolution of the data used in the transfer learning process. In the CMGFNet, the transfer learning performance on the Potsdam and Vaihingen datasets is better than the USGS dataset, likely owing to the ISPRS datasets' more accurate image registration processing and higher image resolution.

In the next stage of the experiments, a fine-tuned transfer learning technique is used. In this scenario, after the network weights are initialized by the parameters of the trained network, just half of the

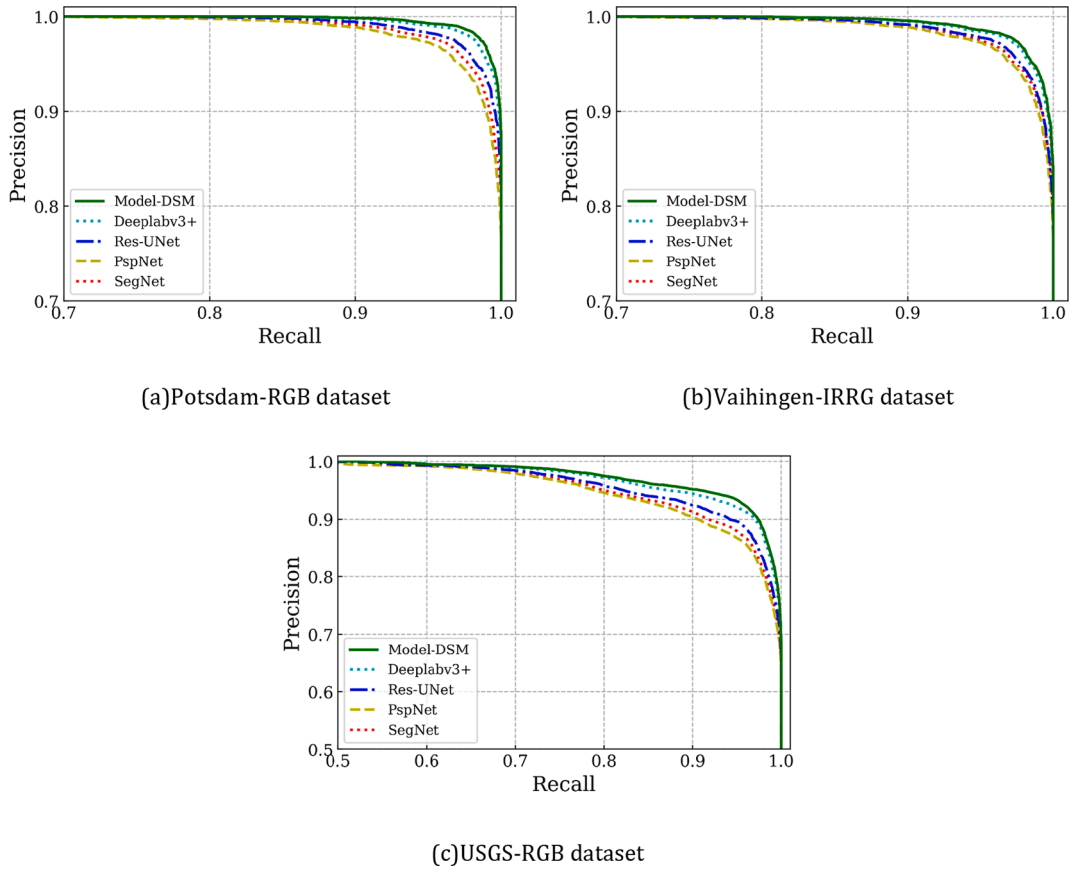

Fig. 13. P-R Curves comparison of SegNet, PSPNet, Res-UNet, DeepLabv3+, and Model-DSM.

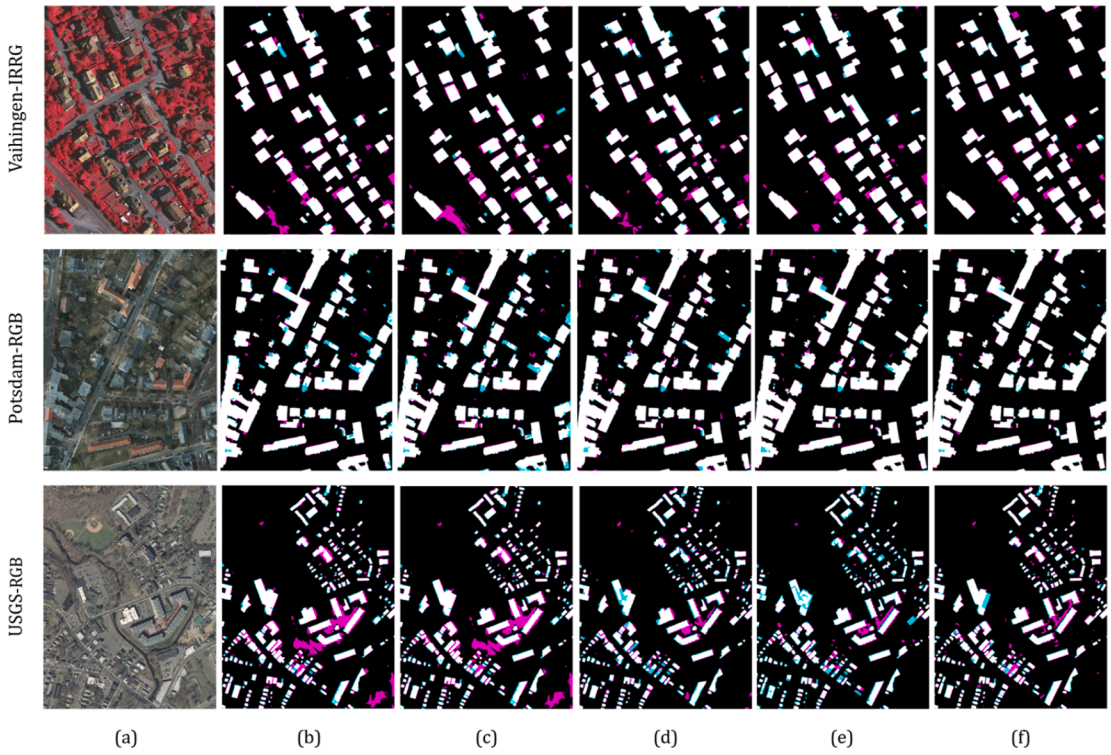

Fig. 14. Comparison of the proposed Model–DSM method and four state-of-the-art single-modality models. (a) Original input image. (b) The result of SegNet. (c) The result of PSPNet. (d) The result of Res-UNet. (e) The result of DeepLabv3+, and (f) the result of the proposed Model–DSM. The TP, FP and, FN are marked in white, cyan, and pink, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of different encoder networks that adopted to the CMGFNet.

Encoder type	IoU(%)			Params(M)	MACs(G)	Image /S	
	Vaihingen-IRRG	Potsdam-RGB	USGS-RGB			Train-mode	Test-mode
ShuffleNet_v2_x1_0	86.92	91.01	78.21	7.06	9.46	2.13	0.156
VGG-16bn	88.32	92.18	79.69	30.06	259.41	1.01	0.294
DensNet-121	88.39	92.43	79.74	16.13	123.77	1.28	0.201
ResNet-34	88.53	92.80	79.96	43.68	98.14	1.58	0.197

Table 5
Evaluation metrics score (%) of FuseNet, V-FuseNet, HAFNet and the CMGFNet.

Datasets	Metrics	FuseNet	V-FuseNet	HAFNet	CMGFNet
Potsdam-RGB	OA	97.65	97.87	97.99	**98.24**
	F-score	96.57	96.89	97.12	**97.41**
	IoU	90.47	90.98	91.31	**92.18**
Vaihingen-IRRG	OA	96.17	96.34	96.42	**96.7**
	F-score	94.9	95.06	95.34	**95.64**
	IoU	86.89	87.06	87.53	**88.32**
USGS-RGB	OA	95.94	96.01	96.18	**96.27**
	F-score	92.02	92.41	92.88	**93.18**
	IoU	78.23	78.79	79.09	**79.69**

target data was used for training in each new dataset. The IoU score for the same test image in each dataset, after 5 and 10 repetitions, is given in the sixth and seventh columns of Table 6. For instance, the IoU score after ten epochs of direct learning using the Potsdam-RGB dataset is calculated to be 84.26%. However, utilizing transfer learning with fine tuning of the model with the Potsdam-IRRG dataset, the training of the same dataset yielded a result value of 90.92% after only 10 epochs. In addition, fine tuning in a model that uses higher spatial resolution images has a higher IoU, showing more convergence in repetition, and saves more computing time. Due to the result of the model's generalization, using a pre-trained model in building extraction is an intelligent decision.

However, essential factors should be considered about the applicability of the proposed model. Similar to other research in the field of deep multimodal data fusion, our method can be suffered from some challenges. In the following, we will review and propose a possible solution to overcome these issues in future research:

Data diversity: The availability and cost of extracting the high-precision DSM from LIDAR data reduces the efficiency and usability of the deep multimodal method. As a result, the size of multimodal datasets is usually smaller than the size of VHR image datasets. In addition, the available datasets are usually recorded in limited urban areas, weather conditions, and sensor settings. Data augmentation via simulation is one way to get around these limitations. In other cases, DSM data can be obtained from cheaper methods. In recent years, DSM creation from VHR images is becoming more possible thanks to advancements in Image-Dense-Matching (IDM) and Structure-from-Motion (SfM) algorithms. In some circumstances, IDM algorithms are improving to enable finer-resolution DSM creation from unmanned aerial system (UAS) data, comparable to the level of airborne LiDAR (Salach et al., 2018). Increasing the efficiency of data labeling is another technique to overcome the limits of datasets generation. In some projects, it is relatively simple to collect DSM and VHR images when creating a multimodal

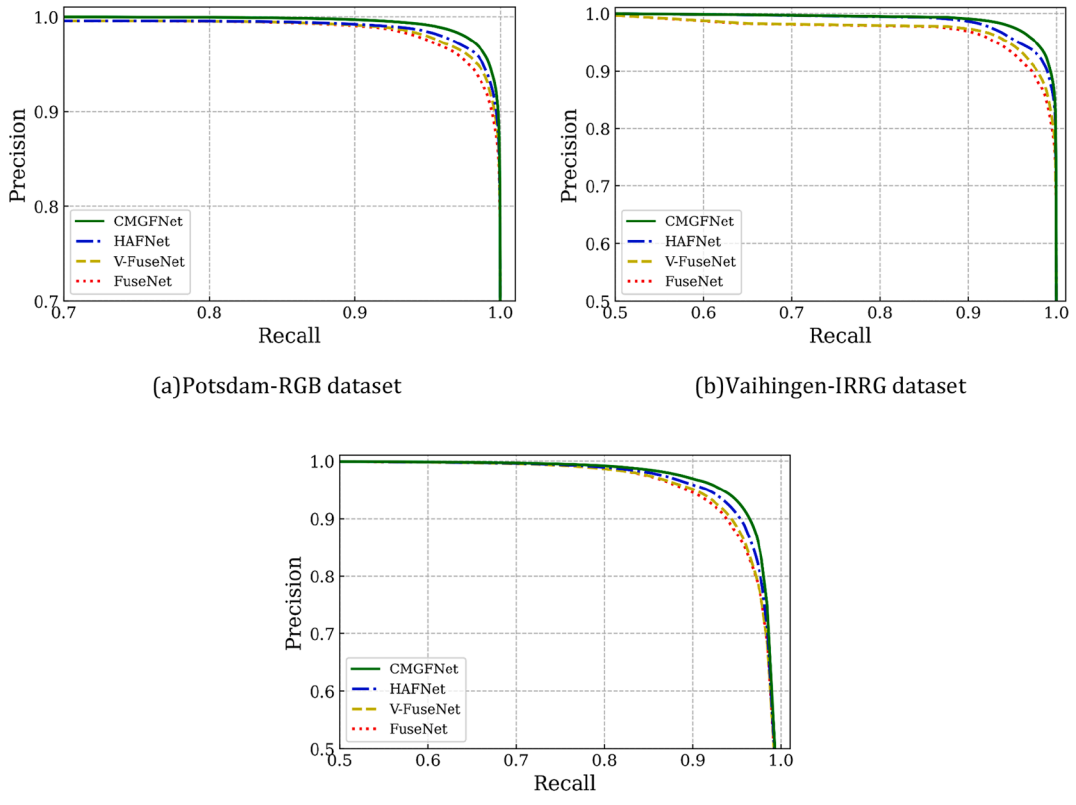

Fig. 15. P-R Curves comparison of FuseNet, V-FuseNet, HAFNet and the CMGFNet.

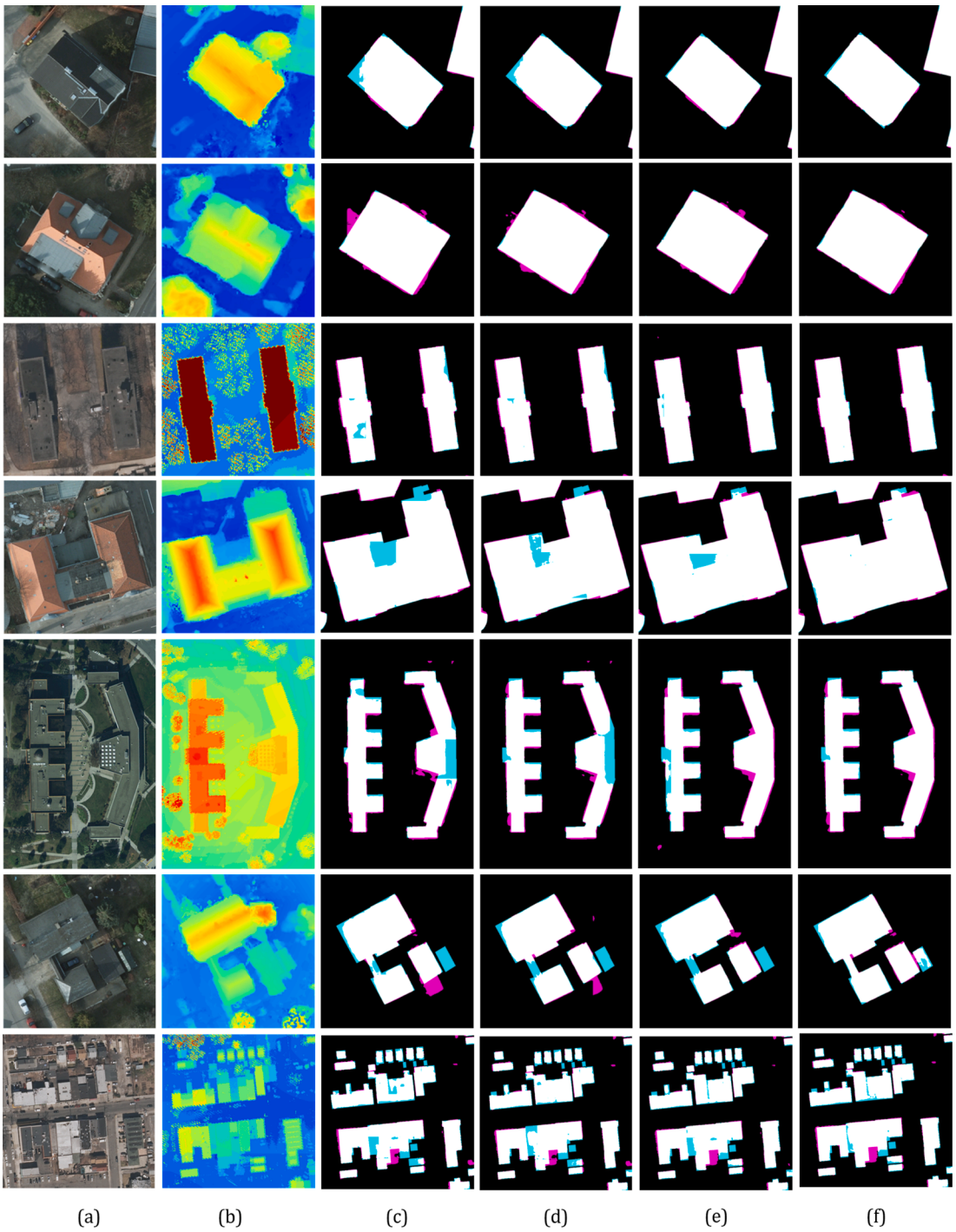

Fig. 16. Comparison of the proposed method and three state-of-the-art models. (a) Original input image. (b) DSM. (c) The result of FuseNet. (d) The result of V-FuseNet. (e) The result of HAFNet, and (f) the result of the CMGFNet model. The TP, FP and, FN are marked in white, cyan, and pink, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

training dataset for building extraction. However, it is very time-consuming and difficult to label them, especially when dealing with LiDAR-derived DSM data. Therefore, the use of optimal labeling methods such as transfer learning can overcome this problem.

Data quality: Besides the size and diversity of the training dataset, data quality significantly affects the performance of a deep multimodal fusion methods. The quality of multimodal data is affected from two perspectives in building extraction tasks, the ground-truth errors and misalignment of DSM with VHR images. Deep multimodal networks are specifically robust for random ground-truth errors. However, the presence of a bias error in ground-truth can cause a significant error in the training and validation of the proposed model. When accruing and preparing training data, temporal and spatial misalignments between DSM and optical data may arise. This could result in significant mistakes in training datasets and reduce network performance. In this paper, we utilized DSM data which were rasterized from the LiDAR point cloud and align with VHR images by the data provider. In addition, to avoid uncertainty in the building extraction results, no change was made in the quality of the original dataset.

Table 6
Transfer learning and fine tuning IoU results of different datasets on the proposed CMGFNet.

Source Datasets	Direct learning		Target Dataset	Transfer learning	Fine tuning	
	(10 epochs) IoU (%)	(60 epochs) IoU (%)		IoU (%)	(5 epochs) IoU (%)	(10 epochs) IoU (%)
Potsdam-RGB	84.26		Potsdam-IRRG	89.61	90.93	91.21
		92.80	Vaihingen-IRRG	66.03	74.23	86.72
			USGS-RGB	51.25	60.23	66.07
Potsdam-IRRG	83.90		Potsdam-RGB	89.94	90.43	90.92
		92.63	Vaihingen-IRRG	83.23	86.26	86.90
			USGS-RGB	52.04	55.52	60.67
Vaihingen-IRRG	79.98	88.84	Potsdam-RGB	78.54	84.86	90.29
			Potsdam-IRRG	80.27	82.50	85.74
			USGS-RGB	41.70	50.25	54.23
USGS-RGB	60.40		Potsdam-RGB	36.74	59.81	83.57
		79.96	Potsdam-IRRG	21.36	46.57	81.09
			Vaihingen-IRRG	24.21	48.36	77.74

5. Conclusions

The primary goal of this work is to extract the building object from VHR photogrammetry and remote sensing imagery. We propose a new end-to-end trainable cross-modal gated fusion deep network (CMGFNet) that fuse both VHR remote sensing images and DSM data. The encoder sections of the CMGFNet are built on a residual network for both RGB and DSM streams. The feature of the RGB encoder guides the learning of the DSM feature to obtain cross-modal feature fusion by the GFM module. On the other hand, the multi-level feature fusion can fuse the feature of high-level layers with low-level layers through a top-down strategy. The decoder sections use R-DSC to transmit and up-sample semantic information from the deep layer to the shallow layer. Three publicly available datasets consisting of VHR remote sensing images and corresponding DSM are used to evaluate the proposed method. These datasets include urban and non-urban areas with a great variety of buildings in size and shape. Experimental results from challenging datasets show that the CMGFNet surpasses other fusion-based methods, and the efficacy of all main elements is confirmed by the extensive ablation study. In the future, the quality improvements of DSM data will be considered in building extraction results. In addition, the single input channel DSM of the network will be developed to support other modalities.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Potsdam and Vaihingen datasets are provided by the German Society for Photogrammetry and Remote Sensing. The dataset is published by United States Geological Survey.

References

Ahmadi, S., Zoej, M.J.V., Ebadi, H., Moghaddam, H.A., Mohammadzadeh, A., 2010. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. Int. J. Appl. Earth Obs. Geoinf. 12 (3), 150–157. https://doi.org/10.1016/j.jag.2010.02.001.

Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A., 2020. Gated multimodal networks. Neural Comput. Appl. 32 (14), 10209–10228. https://doi.org/10.1007/s00521-019-04559-1.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Bradbury, K., Brigman, B., Collins, L., Johnson, T., Lin, S., Newell, R., Park, S., Suresh, S., Hole, W., 2016. Aerial imagery object identification dataset for building and road detection, and building height estimation. figshare. Collect. 1–22.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Aug, C. V, 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation.

Chollet, F., 2016. Xception: deep learning with depthwise separable convolutions 1251–1258.

Feng, W., Sui, H., Hua, L.i., Xu, C., Ma, G., Huang, W., 2020. Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder-decoder architecture and historical land use vector map. Int. J. Remote Sens. 41 (17), 6595–6617. https://doi.org/10.1080/01431161.2020.1742944.

Freire, S., Santos, T., Navarro, A., Soares, F., Silva, J.D., Afonso, N., Fonseca, A., Tenedório, J., 2014. Introducing mapping standards in the quality assessment of buildings extracted from very high resolution satellite imagery. ISPRS J. Photogramm. Remote Sens. 90, 1–9. https://doi.org/10.1016/j.isprsjprs.2013.12.009.

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z., Yang, X.u., 2013. Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. Int. J. Remote Sens. 34 (14), 5166–5186. https://doi.org/10.1080/01431161.2013.788261.

Hammoudi, K., Dornaika, F., 2010. A featureless approach to 3D polyhedral building modeling from aerial images. Sensors 11, 228–259. https://doi.org/10.3390/s110100228.

Han, D., Kim, Jiwhan, Kim, Junmo, 2016. Deep pyramidal residual networks. In: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-Janua, pp. 6307–6315. https://doi.org/10.1109/CVPR.2017.668.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2017. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. Int. Polit. Sci. Rev. 213–228. https://doi.org/10.1007/978-3-319-54181-5_14.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 630–645. https://doi.org/10.1007/978-3-319-46493-0_38.

Hermosilla, T., Ruiz, L.A., Recio, J.A., Estornell, J., 2011. Evaluation of automatic building detection approaches combining high resolution images and LiDAR data. Remote Sens. 3, 1188–1210. https://doi.org/10.3390/rs3061188.

Hoeser, T., Kuenzer, C., 2020. Object detection and image segmentation with deep learning on earth observation data: a review-Part I: evolution and recent trends. Remote Sens. 12, 1667. https://doi.org/10.3390/rs12101667.

Hosseinpoor, H., Samadzadegan, F., 2020. Convolutional neural network for building extraction from high-resolution remote sensing images. In: 2020 International Conference on Machine Vision and Image Processing (MVIP). IEEE, pp. 1–5. https://doi.org/10.1109/MVIP49855.2020.9187483.

Hu, J., You, S., Neumann, U., Park, K.K., 2004. Building modeling from LIDAR and aerial imagery. Aspers 4, 23–28.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Huang, J., Zhang, X., Xin, Q., Sun, Y., Zhang, P., 2019. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. ISPRS J. Photogramm. Remote Sens. 151, 91–105. https://doi.org/10.1016/j.isprsjprs.2019.02.019.

Huang, X., Zhang, L., 2012. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 5 (1), 161–172. https://doi.org/10.1109/JSTARS.2011.2168195.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K., Wang, G., 2018. Multilayer perceptron neural network for surface water extraction in Landsat 8 OLI satellite images. Remote Sens. 10, 755. https://doi.org/10.3390/rs10050755.

Kaiser, L., Gomez, A.N., Chollet, F., 2017. Depthwise separable convolutions for neural machine translation. arXiv.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization 1–15.
Liu, Z., Zhang, W., Zhao, P., 2020. A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. Neurocomputing 387, 210–220. https://doi.org/10.1016/j.neucom.2020.01.045.
Long, J., Shelhamer, E., Darrell, T., 2014. Fully convolutional networks for semantic segmentation. In: 2019 IEEE/CVF Int Conf. Comput. Vis. Work., pp. 847–856. https://doi.org/10.1109/ICCVW.2019.00113.
Ma, J., 2020. Segmentation Loss Odyssey. arXiv.
Ma, J., Wu, L., Tang, X., Liu, F., Zhang, X., Jiao, L., 2020. Building extraction of aerial images by a global and multi-scale encoder-decoder network. Remote Sens. 12, 2350. https://doi.org/10.3390/rs12152350.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55 (2), 645–657. https://doi.org/10.1109/TGRS.2016.2612821.
Maltezos, E., Doulamis, A., Doulamis, N., Ioannidis, C., 2019. Building extraction from LiDAR data applying deep convolutional neural networks. IEEE Geosci. Remote Sens. Lett. 16 (1), 155–159. https://doi.org/10.1109/LGRS.2018.2867736.
Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009.
Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571. https://doi.org/10.1109/3DV.2016.79.
Mohammadi, H., Samadzadegan, F., 2020. An object based framework for building change analysis using 2D and 3D information of high resolution satellite images. Adv. Sp. Res. 66 (6), 1386–1404. https://doi.org/10.1016/j.asr.2020.05.041.
Nahhas, F.H., Shafri, H.Z.M., Sameen, M.I., Pradhan, B., Mansor, S., 2018. Deep learning approach for building detection using LiDAR–Orthophoto fusion. J. Sensors 2018, 1–12. https://doi.org/10.1155/2018/7212307.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proc. 27th Int. Conf. Int. Conf. Mach. Learn., pp. 807–814.
Ngo, T.-T., Mazet, V., Collet, C., de Fraipont, P., 2017. Shape-based building detection in visible band images using shadow information. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (3), 920–932. https://doi.org/10.1109/JSTARS.2016.2598856.
Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: Proc. IEEE Int. Conf. Comput. Vis. 2015 Inter, pp. 1520–1528. https://doi.org/10.1109/ICCV.2015.178.
Osco, L.P., Junior, J.M., Ramos, A.P.M., Jorge, L.A. de C., Fatholahi, S.N., Silva, J. de A., Matsubara, E.T., Pistori, H., Gonçalves, W.N., Li, J., 2021.
Ozdarici-Ok, A., Ok, A., Schindler, K., 2015. Mapping of agricultural crops from single high-resolution multispectral images—data-driven smoothing vs. parcel-based smoothing. Remote Sens. 7, 5611–5638. https://doi.org/10.3390/rs70505611.
Pacifici, F., Chini, M., Emery, W.J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113 (6), 1276–1292. https://doi.org/10.1016/j.rse.2009.02.014.
Pan, X., Yang, F., Gao, L., Chen, Z., Zhang, B., Fan, H., Ren, J., 2019. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. Remote Sens. 11, 917. https://doi.org/10.3390/rs11080917.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. arXiv.
Piramanayagam, S., Saber, E., Schwartzkopf, W., Koehler, F., 2018. Supervised classification of multisensor remotely sensed images using a deep learning framework. Remote Sens. 10, 1429. https://doi.org/10.3390/rs10091429.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. Comput. Vis. Pattern Recognit. 1–8.
Salach, A., Bakuła, K., Pilarska, M., Ostrowski, W., Górski, K., Kurczyński, Z., 2018. Accuracy assessment of point clouds from LiDAR and dense image matching acquired using the UAV platform for DTM creation. ISPRS Int. J. Geo-Information 7, 342. https://doi.org/10.3390/ijgi7090342.
Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S., Sommai, C., 2020. BRRNet: a fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. Remote Sens. 12, 1050. https://doi.org/10.3390/rs12061050.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. Am. J. Heal. Pharm. 75, 398–406. https://doi.org/10.2146/ajhp170251.
Sirmacek, B., Unsalan, C., 2009. Urban-area and building detection using SIFT keypoints and graph theory. IEEE Trans. Geosci. Remote Sens. 47 (4), 1156–1167. https://doi.org/10.1109/TGRS.2008.2008440.
Sun, Y., Zhang, X., Xin, Q., Huang, J., 2018. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. ISPRS J. Photogramm. Remote Sens. 143, 3–14. https://doi.org/10.1016/j.isprsjprs.2018.06.005.
Tomljenovic, I., Tiede, D., Blaschke, T., 2016. A building extraction approach for Airborne Laser Scanner data utilizing the Object Based Image Analysis paradigm. Int. J. Appl. Earth Obs. Geoinf. 52, 137–148. https://doi.org/10.1016/j.jag.2016.06.007.
Vetrivel, A., Gerke, M., Kerle, N., Vosselman, G., 2015. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. ISPRS J. Photogramm. Remote Sens. 105, 61–78. https://doi.org/10.1016/j.isprsjprs.2015.03.016.
Weidner, U., 1997. Digital surface models for building extraction. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images (II). Birkhäuser Basel, Basel, pp. 193–202. https://doi.org/10.1007/978-3-0348-8906-3_19.
Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., Shibasaki, R., 2018. Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks. Remote Sens. 10, 1–18. https://doi.org/10.3390/rs10030407.
Xu, Y., Du, B.o., Zhang, L., Cerra, D., Pato, M., Carmona, E., Prasad, S., Yokoya, N., Hansch, R., Le Saux, B., 2019. Advanced multi-sensor optical remote sensing for urban land use and land cover classification: outcome of the 2018 ieee grss data fusion contest. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 12 (6), 1709–1724. https://doi.org/10.1109/JSTARS.2019.2911113.
Xu, Y., Wu, L., Xie, Z., Chen, Z., 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. Remote Sens. 10, 144. https://doi.org/10.3390/rs10010144.
Zhang, C.e., Wang, T., Atkinson, P.M., Pan, X., Li, H., 2015. A novel multi-parameter support vector machine for image classification. Int. J. Remote Sens. 36 (7), 1890–1906. https://doi.org/10.1080/01431161.2015.1029096.
Zhang, P., Du, P., Lin, C., Wang, X., Li, E., Xue, Z., Bai, X., 2020a. A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data. Remote Sens. 12, 3764. https://doi.org/10.3390/rs12223764.
Zhang, W., Huang, H., Schmitz, M., Sun, X., Wang, H., Mayer, H., 2017a. Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling. Remote Sens. 10, 52. https://doi.org/10.3390/rs10010052.
Zhang, X., Zheng, Y., Liu, W., Peng, Y., Wang, Z., Balas, V.E., Jain, L.C., 2020b. An improved architecture for urban building extraction based on depthwise separable convolution. J. Intell. Fuzzy Syst. 38 (5), 5821–5829. https://doi.org/10.3233/JIFS-179669.
Zhang, X., Zhou, X., Lin, M., Sun, J., 2017. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. Black Stud. Read. 1–488. https://doi.org/1707.01083.
Zhang, Y., Sidibé, D., Morel, O., Mériaudeau, F., 2021. Deep multimodal fusion for semantic image segmentation: a survey. Image Vis. Comput. 105, 104042. https://doi.org/10.1016/j.imavis.2020.104042.
Zhang, Z., Wang, Y., 2019. JointNet: a common neural network for road and building extraction. Remote Sens. 11, 696. https://doi.org/10.3390/rs11060696.
Zhang, Z., Zhang, X., Peng, C., Cheng, D., Sun, J., 2018. ExFuse: enhancing feature fusion for semantic segmentation pp. 1–17.