MDBES-Net: Building Extraction From Remote Sensing Images Based on Multiscale Decoupled Body and Edge Supervision Network

Shengjun Xu, Miao Du, Yuebo Meng, Guanghui Liu, Jiuqiang Han, and Bohan Zhan

***Abstract*—The extraction of buildings in aerial remote sensing applications is an important and challenging task. Most existing methods extract buildings based on local area attention, ignoring the loss of accuracy due to the global structure of the building. However, global structural features of buildings with strong coupling relationships in complex scenes are difficult to extract, such as the edges and bodies of buildings, leading to discontinuous results. Therefore, multiscale decoupled body and edge supervision network (MDBES-Net), which can consider both edge optimization and inner consistency, is proposed to solve these problems. MDBES-net consists of the body-mask-edge consistency constraint base network (BMECC), decoupling the body and edge aware module (DBEA), and the channel decoupled attention module (CDA). First, body-mask-edge consistency constraint supervision is established by body and edge labels to jointly improve the segmentation effect in the BMECC base network. Second, in the mutiscale DBEA module, building features are warped by a learnable flow field to make body parts more consistent and edges more detailed. Finally, the CDA module performs adaptive calibration of the recoupled feature map channel response to minimize external background noise interference. Experiments on the open Massachusetts building dataset, WHU Building Dataset show that the proposed MDBES-Net can accurately extract buildings in complex scenarios, enabling complete building segmentation with refined boundaries and improved internal consistency.**

***Index Terms*—Body and edge decoupled awareness, boundary optimization, building extraction, flow field, remote sensing image.**

I. Introduction

AS THE world's land resources become increasingly scarce and the pressure on the environment intensifies, it is important to balance land development with the carrying capacity of resources and the environment. As one of the main constituents of the resource environment, buildings have inevitably become the focus of researchers' attention [1]. Remote sensing imagery is an important means to obtain information on land features, and the extraction of buildings from high resolution remote sensing imagery is crucial for applications, such as the unauthorized building monitoring, automatic extraction of urban areas, map updating, urban change monitoring, urban planning, three-dimensional (3-D) modeling and digital city establishment [2].

In the past, traditional image processing methods enabled automatic extraction of buildings from remote sensing images. Building feature extraction predominantly relied on traditional feature extraction algorithms, including corner detection operator [3], edge detection operator [4], image transform [5], and histogram [6]. Some researchers have applied active contour region segmentation methods [7], [8] to construct building structural information and segment images into regions with similar and homogeneous properties. Nonetheless, these conventional methods are incapable of fully extracting the structural characteristics of buildings in complex environments, resulting in significant loss of edge texture information during down-sampling. Additional information, such as digital elevation model data [9], [10] and GIS data [11], contains richer spatial geometric information from remotely sensed images, which can enhance the accuracy and robustness of the model. However, these methods are expensive, inefficient, and exhibit weak generalization performance. Human-computer interaction [12] relies on a manual interpretation method guided by personal and expert experience. It offers visualization, analysis, and convenience, with a "what you see is what you get" feature. However, it is prone to human biases and subjectivity, making it unsuitable for remote sensing image research.

In recent years, remote sensing techniques based on deep learning image processing have developed rapidly. The extensive use of semantic segmentation methods has been shown to be effective for a variety of pixel-level classification tasks [13]. Among them, the typical fully convolutional network (FCN) [14] breaks the traditional piecewise barrier by replacing the final fully connected layer with a convolutional layer, successfully realizes pixel-level classification, and performs well in many important semantic segmentation tasks. Some studies have added innovations such as attention modeling and expanding sensory

Manuscript received 8 September 2023; revised 20 October 2023 and 6 November 2023; accepted 6 November 2023. Date of publication 9 November 2023; date of current version 23 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 52278125; in part by the Natural Science Foundation of Shaanxi Province under Grant 2023-JC-YB-532; and in part by the Key Research and Development Projects of Shaanxi Province under Grant 2021 SF-429. *(Shengjun Xu and Miao Du are co-first authors.) (Corresponding author: Miao Du.)*

Shengjun Xu, Miao Du, Yuebo Meng, and Guanghui Liu are with the School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China (e-mail: duplin@sina.com; 1067103658@qq.com; mengyuebo@163.com; guanghuil@163.com).

Jiuqiang Han and Bohan Zhan are with the School of Electronic Science and Engineering, Xi'an Jiaotong University, Xi'an 710055, China (e-mail: jqhan@mail.xjtu.edu.cn; zbh@xauat.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3331444

field based on the FCN framework, and achieved good building extraction results [15], [16], [17]. However, accurate building extraction from remote sensing images is still affected by various factors. On the one hand, the buildings themselves have spatial diversity, complex morphology, different sizes, and roof material transformations, leading to incomplete building segmentation results and lack of internal consistency. On the other hand, the lack of morphological structural features in building imaging due to geographic interference from roof shadows and tree occlusion leads to blurring and discontinuity of extracted edges.

Accurate building segmentation performance requires accurate modeling of building edges [18]. Recently, several studies have used post-processing techniques [19], or enhanced boundary information description [20], [21], to refine edge segmentation by extracting edge features and assisting in generating building segmentation results with enhanced boundary information. Some other researchers have used a priori fully connected conditional random fields [22] to address the problem of building detail texture loss and edge smoothing during down-sampling. However, these methods cannot address the hazard of coupling edge features to the complete features of the object during feature extraction, resulting in the loss of structural features, which to some extent hinders the complete extraction of edges of interest. Several studies have attempted to use multitask learning networks for semantic segmentation and edge detection [23], where building extraction and edge detection are categorized into dual streams of learning and information fusion interactions, which are typically combined with HED edge detection [24]. This dual stream of network learning prevents the model from accurately obtaining contextual information about the edges. There is also some work on frame field learning [25] and boundary constraints [26] to learn the process of aligning features at different scales during down-sampling to capture more information about the spatial localization of edge features. Some researchers have also modeled boundary direction fields along building boundaries in the decoder [27], [28] to guide the inward aggregation of edge features. However, these methods do not take into account the fact that the edge features are strongly coupled in the overall features of the building, and the background complex noise hinders the presentation of the edge characteristics, resulting in blurring and discontinuity of the edges of the building.

Channel attention mechanism plays an important role in improving the performance of internal integrity of objects [29]. Inspired by the classical SENet [30], researchers have proposed a series of networks that combine spatial and channel attention, such as DenseNet [31], SCAttNet [32], WFCA [33], FCANet [34], ECA-Net [35], and SKNeT [37]. SCAttNet [32] semantic segmentation network integrates lightweight spatial and channel attention modules that adaptively refine features to address the problem of small inter-class variance in semantic segmentation. WFCA [33] and FCANet [34] added attention weights in both frequency and spatial domain dimensions, respectively, to learn meaningful features in the channel space. ECA-Net [35] proposes a local cross-channel interaction strategy that does not require dimensionality reduction while overcoming performance and complexity tradeoffs. ReXNet [36] devised an efficient search method for channel configuration by means of block-indexed piecewise linear functions. SKNeT [37] proposed a dynamic selection mechanism for convolutional kernels that can capture target objects at different scales and maximize the generalization performance of the model. However, these studies have not explored the importance of inter-channel correlation structural features that capture the intrinsic distribution of the feature space and characterize the full diversity of features.

To solve the problems of blurred edges and internal inconsistencies and discontinuities, MDBES-Net is proposed for building segmentation in complex remote sensing scenes. First, we construct a body-mask-edge consistent constraint (BMECC) based network to explore the implicit spatial relationship of the interaction between edge and body, and mine the global structural feature information of the building. We then further introduce a decoupling the body and edge aware (DBEA) module, which use a learnable semantic flow field to warp each pixel toward the inner part of the object to maintain the consistency of the body part of the building and generate fine edges. Finally, the channel decoupled attention (CDA) module is introduced to capture the global structural information of the remote sensing image building by mining the structural correlation between the body and edge reconstruction coupled feature map channels to suppress background noise. Experiments on the open Massachusetts building dataset and the WHU aerial building dataset show that MDBES-Net achieves excellent performance in generalization and robustness.

The contributions of this article are threefold.

1) BMECC supervises the segmentation of buildings at three structural levels: edge; body; and object. The loss function for Body-Mask-Edge consistency constraints is designed, with different loss functions and weights allocated to each region to achieve multilevel fine segmentation.
2) DBEA is proposed to decouple the overall structural features of a building at three different scales by warping each pixel towards the interior of the building through a learnable semantic flow field, which improves the consistency of the building body and refines the edges of the building.
3) CDA is developed to extract the spatial distribution relationship of features by coupling edge and body features. This improves the compatibility of body and edge features in the channel dimension, suppressing background noise, and guiding edge fine segmentation.

The rest of this article is organized as follows. Section II introduces the details of the proposed MDBES-Net. Section III presents the implementation details of the experiments and discussion. Finally, Section IV concludes this article.

II. Methodology

A. Overview of Proposed Method

The overall structure of the network is shown in Fig. 1. First, in the encoding stage, the baseline network uses multi-layer convolutional operation stacking to extract features from the input remote sensing images, and obtains low-level semantic features with spatially textured complete and strongly correlated high-level semantic information, respectively. Second, in the decoding stage, the proposed DBEA module decouples the

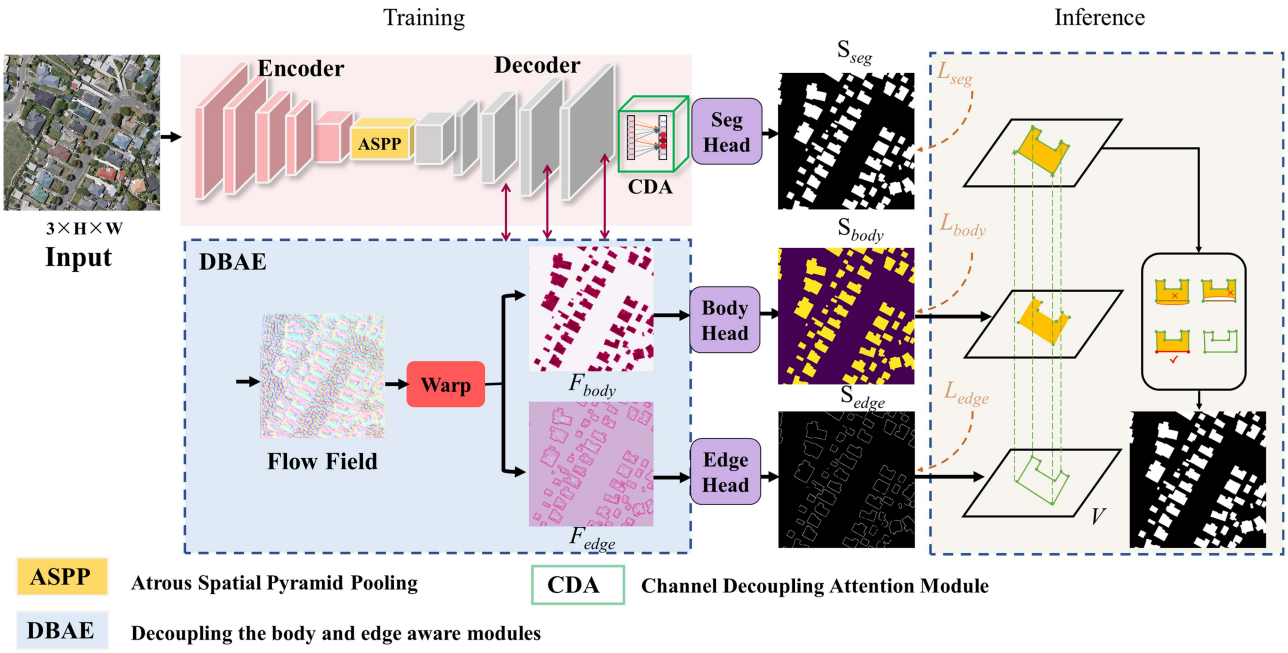

Fig. 1. MDBES-net network architecture.

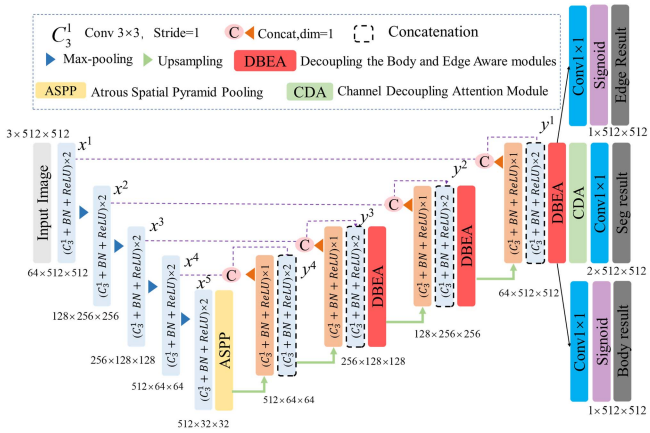

Fig. 2. Detailed process of the body-mas-edge consistency constraint framework.

decoded features containing multi-scale feature information into body features and edge features, and conducts deep supervision on both features to explore the coupling relationship between the intra-class features. Then, Atrous spatial pyramidal pooling (ASPP) is embedded between encoding stage and decoding stage to achieve inference of global contextual information and better utilize multi-scale feature information of buildings. Finally, the CDA is used to eliminate inter-class noise interference factors in the recoupled feature map and suppress the background noise of building images.

B. Body-Mask-Edge Consistency Constraint Base Network

The BMECC baseline network in this article uses a typical coding-decoding structure as shown in Fig. 2. The input tensor is expressed as $I \in \mathbb{R}^{C\times H\times W}$ where the number 3 refers to the number of channels of the input image, and H, W are the height and width of the feature map, respectively. There are five layers in the coding stage. We use the convolution block to extract image features and increase the number of channel dimensions. The convolution block is two convolution layers with a kernel size of 3×3 followed by batch normalization and ReLU activation, and $\times 2$ indicates that the convolution block is stacked two times. To obtain more high-level semantic features, the maximum pooling operation is applied to sample deeper network depth, and the convolutional block of the previous layer and the two-fold down-sampling operation are repeated in turn to obtain feature maps of $\{x^1, x^2, x^3, x^4, x^5\}$ feature map, respectively.

The decoding process can be divided into four layers of feature map of different sizes $\{y^l|\forall l \in [1, 2, 3, 4]\}$. The proposed method uses a bilinear interpolation algorithm with a scale factor of 2 for up-sampling. In order to fuse the high-level semantic features with the low-level texture detail features, the feature maps x^l and y^l with the same size in the coding and decoding stage are connected together in the channel dimension, and then restored to the pre-stitching channel dimension after the convolution operation. Finally, convolution with a kernel size of 1 is used to supervise the body feature, edge feature and final segmentation result in the prediction stages to learn more about the structural features of the building. In particular, in order to enhance the multiscale feature expression ability of the network on the codec structure, in the last layer of the encoding stage, ASPP is embedded, and the multi-scale context information is encoded by parallel convolution to enhance the representation of convolutional features and learn more about the global contextual information of the building. The detailed process of network decoding is as follows.

First, input I can obtain multiscale low-level feature maps x^l by convolution and down-sampling operations in turn. The contextual feature representation is enhanced at the deepest layer

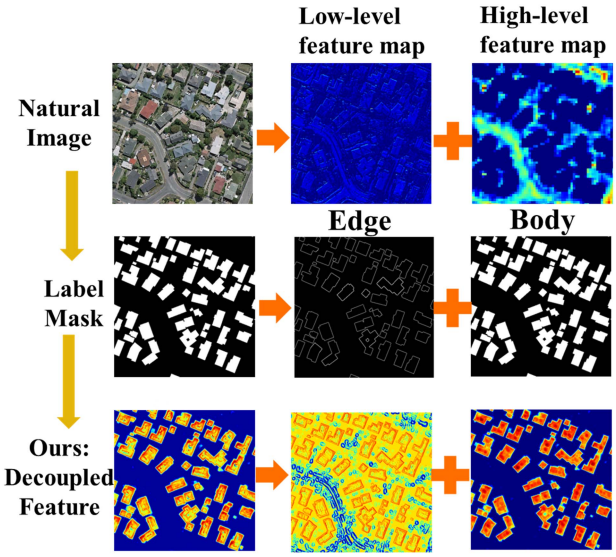

Fig. 3. Thoughts on our decoupled features map.

of the network using ASPP. The formula is expressed as follows:

$$x^l = \begin{cases} \text{Conv}(I), & l = 1 \\ \text{Conv}[\Phi(x^{l-1})], & l \in \{2, 3, 4\} \\ \text{ASPP}\{\text{Conv}[\Phi(x^{l-1})]\}, & l = 5 \end{cases} \quad (1)$$

where $\text{Conv}(\cdot)$ represents the convolution operation with kernel size 3. $\Phi(\cdot)$ denotes the $2\times$ Max pooling operation. $\text{ASPP}(\cdot)$ stands for ASPP operation.

Second, the decoding layer feature map y^{l-1} and the corresponding encoding layer feature map x^{l-1} are stitched together in the channel dimension. The spliced features are then subjected to a dimensionality reduction operation by convolution, and the overall process can be expressed as follows:

$$y^l = \psi(x^{l-1}, y^{l-1}) \quad (2)$$

where $\psi(\cdot)$ denotes concatenate splicing operation in the channel dimension.

Finally, the DBEA module is proposed to decouple feature map y^l into edge feature map y_e and body feature map y_b, and then the corresponding elements of them are added and recoupled into the full structure of the building segmentation feature map y_s. The following formulas can express it as follows:

$$\{y_e^l, y_b^l\} \leftarrow \text{Decpl}\left(y^l | \forall l = 1, 2, 3\right) \quad (3)$$

where $\text{Decpl}(\cdot)$ denotes the feature decoupling process.

Note that the DBEA module is embedded before feature map y^1, y^2, and y^3, respectively, and the CDA module is embedded before the final prediction. For more information, please refer to following section. The final layer of the network uses three classification heads to make predictions for the body, edge, and final building segmentation, respectively, followed by a Sigmoid activation function to obtain the extraction results for the building objects and edges.

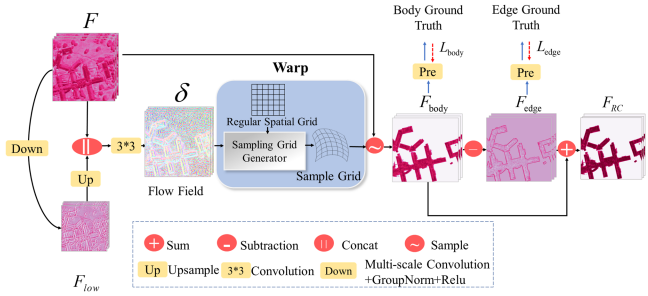

Fig. 4. Illustration of the decoupling body and edge aware module.

C. Decoupling the Body and Edge Aware Module

As shown in the first row of Fig. 3, the natural remote sensing building images are input to CNN network, which can be decoupled into high-level feature map and low-level feature map according to the depth of the network. The low-level feature map in shallower networks contains more spatial detail information from the images, and the high-level feature map have richer semantic information in deeper networks [39]. In remote sensing image labels, the building labels can be divided into edge part and body part from the structural level. Through this idea of decoupling semantically and structurally, building extraction in remote sensing images usually perceives the entire building through the body and edges of the building's roof, however there is a strong coupling relationship between the internal smooth body and the external abrupt edge. The body has an outward expanding and inward compacting quality, and edge has a quality of limitation and restriction. Therefore, if the network cannot distinguish between the extended body and the constraint edge features at the true edge position, it will often cause the building edge to expand and deform outward, resulting in blurred building segmentation edges. Due to the strong coupling between body and edge in the high-dimensional image space of remote sensing buildings, it is challenging to effectively decouple the structural feature information of body and edge and learn the interactions between the two. Inspired by the theory of decoupling, building features can also be decoupled into two structural feature components at the feature map: the building body feature and the building edge feature.

1) Flow Field Generation: The flow field can be seen as an implicit spatial representation of the squeezing process along the direction normal to the edge of the object [38], [39]. To generate a flow that points primarily to the center of an object, highlighting the features inside the center of an object as an explicit guide is a reasonable way to generate a more consistent representation of features for pixels within the same building. As shown in Fig. 4, a learnable flow field is designed, which is used to warp the original feature map in order to obtain an explicit representation of the body features of the building.

First, low-level features provide detailed location information for edge prediction, high-level features highlight the complete semantic information inside a complete building. Dimensionality reduction compression of input F by two consecutive convolution operations with no change in channel dimension. The feature map after dimensionality reduction is F_{low}. The

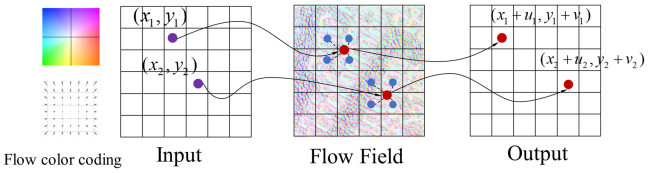

Fig. 5. Warp procedure of the DBEA module.

formula is expressed as

$$F_{\text{low}} = Br[l\text{conv}(F)] \tag{4}$$

where $l\text{conv}(\cdot)$ denotes convolution operation (kernel size is 3, stride is 2, padding is 0), $Br(\cdot)$ denotes batch normalization and ReLU.

Second, up-sample F_{low} to size $H \times W$ by differentiable bilinear sampling mechanism [40], this allows the network model to learn invariance to translations, scale transformations, rotations, and more common distortions. and F_{low} concatenate splice with F in the channel dimension. Finally, applying 3×3 depthwise convolution layer to the stitched feature map to obtain the flow field $\delta \in \mathbb{R}^{2\times H\times W}$. The following formulas can express it:

$$\delta = D\text{conv}(F \parallel F_{\text{low}}) \tag{5}$$

where $D\text{conv}(\cdot)$ represents depthwise convolution operations, $\parallel$ represents a combination of up-sampling and channel splicing operations.

2) Feature Warp: The warp process is flow-based extrusion feature generation method, the continuous convolutional operation for dimensionality reduction is utilized to simulate the dynamic extrusion process of features. $\delta \in \mathbb{R}^{2\times H\times W}$ is obtained from the flow field generation, where the value within δ signifies the pixel offset. 2 channels denote the pixel offset in the x and y axes correspondingly. A positive x value indicates a feature offset to the left, and a negative value indicates an offset to the right. Similarly, a positive y value indicates an upward feature offset, while a negative value indicates a downward feature offset. When there is no corresponding original pixel after coordinate transformation, the algorithm will use bilinear interpolation. the specific warp process is shown in Fig. 5. The input pixel point (x_1, y_1) is transformed using the obtained flow field offsets (u_1, v_1) to determine its new position $(x_1 + u_1, y_1 + v_1)$, which denotes the degree to which the low-resolution feature map is aligned with the high-resolution feature map features via the warp process. However, the coordinates of this corresponding point are not necessarily integer values, so interpolation or neighborhood values are used. In particular, Each position p_l on standard spatial grid Ω_l is mapped to a new point $\hat{p}_l$ via $p_l + \delta_l(p_l)$, then we use the differentiable bilinear sampling mechanism to approximate each point p_x in F_{body}. The formula is expressed as follows:

$$F_{\text{body}}(p_x) = \sum_{p\in N(p_l)} w_p F(p) \tag{6}$$

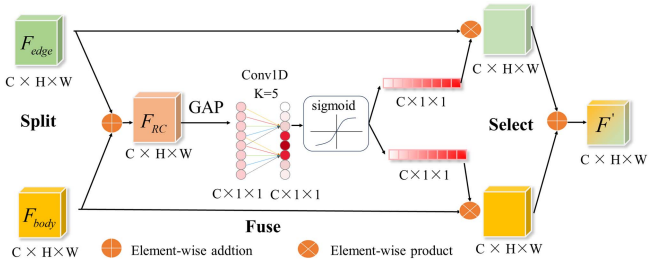

Fig. 6. Illustration of the CDA module.

where w_p indicate bilinear kernel weights on warped spatial gird, calculated from flow map δ. N denotes the total number of feature pixels input to the target to be segmented.

3) Decoupled Edge: At the level of the building object structure, the edge feature and the body feature are strongly coupled in a complementary geometric space dimension. Our approach is to learn that the flow field generated by the network itself warps F towards the interior of the object to obtain F_{body}. The edge feature retains the fine edge geometry and is obtained by explicit subtraction with constrained properties. Specifically, the original feature F is subtracted from the pixel value of the corresponding position of F_{body} to the remaining F_{edge}. The formula is expressed as follows:

$$F_{\text{edge}} = conv(F - F_{\text{body}}) \tag{7}$$

where $\text{conv}(\cdot)$ denotes convolution operations with kernel size 3. Specifically, we embed the DBEA module into y_1, y_2, y_3 layers. However, we only supervise F_{edge} and F_{body} obtained by decoupling layer y_1, The output of the DBEA module in layers y_2 and y_3 is the body and edge recoupled feature F_{RC}. The specific formula is as follows:

$$F_{\text{RC}} = F_{\text{edge}} + F_{\text{body}} \tag{8}$$

$$F_{\text{predge}} = \text{Sigmoid}[\text{Conv}(F_{\text{edge}})] \tag{9}$$

$$F_{\text{prbody}} = \text{Sigmoid}[\text{Conv}(F_{\text{body}})] \tag{10}$$

where $\text{Conv}(\cdot)$ is the prediction head (the convolution operation with kernel size 1. output channel is 1), F_{predge}, F_{prbody} denote building edge prediction results and body prediction results.

C. Channel Decoupling Attention Module

In order to fully couple edge features and body features in the spatial dimension and to improve the structural feature representation ability of the coupled feature mapping, a channel decoupling attention module is proposed. We mainly use to model the spatial distribution relationship between edge features and body features in the channel dimension, simulate the interdependence between them so that the edge features and body features align structural feature responses one by one in the channel dimension, form complementary relationships in each channel, and adaptively calibrate the channel dimension feature responses. As shown in Fig. 6, the CDA module can be divided into three parts: split; fuse; and select.

1) Split: The inputs to the module are edge and body feature maps. For the given $F_{\text{edge}}, F_{\text{body}} \in \mathbb{R}^{C\times H\times W}$, the decoupled feature maps are weighted in the spatial and channel dimensions, aiming at mining the correlation relationship between the edge structure and the body structure embedded in the channel.

2) Fuse: For the given $F_{\text{edge}}, F_{\text{body}} \in \mathbb{R}^{C\times H\times W}$, we first use element-wise addition spatial scale transformation to structurally recouple the edge features and body features of the building to obtain F_{RC}, then the global average pooling is used to deflate the F_{RC}, and the 1-D convolution is used to construct the coupling correlation relationship between edge features and body features in the channel dimension. Finally, the channel weights are obtained using the Sigmoid activation function. The process is expressed as follows:

$$\{V, S\} = \text{Sigmoid}\left\{\text{Conv}\left(\text{GAP}\left(F_{\text{RC}}\right)\right)\right\} \tag{11}$$

where GAP$(\cdot)$ denotes global average pooling. Conv$(\cdot)$ denotes convolutional operations (kernel size is 1×5). Sigmoid$(\cdot)$ denotes sigmoid activation function. V, S represent the edge and body channel weight, respectively.

3) Select: The correlation between structural features and edges of building bodies in remote sensing imagery is captured in the channel dimension, and the correlation of structural features between channels is learned by adaptively enhancing channels with significant structural features between adjacent channels and suppressing channels with minor features. The process is expressed as follows:

$$F' = (F_{\text{edge}} \otimes V) + (F_{\text{body}} \otimes S) \tag{12}$$

where $\otimes$ denotes element-wise product of channel.

D. Body-Mask-Edge Consistency Constraint Loss Function

In order to improve the smoothness of segmentation results, and prevent the overfitting of the network model, the loss function of the building edges and the body consistency constraint is proposed, which decouples the building edge features from the body features, and learns the structural features separately in a data-driven way to refine the edges, which improves the overall segmentation accuracy. Different loss functions are used according to the different segmentation regions. As shown in Fig. 7, $\hat{F}_{\text{seg}}, \hat{F}_{\text{edge}}, \hat{F}_{\text{body}}$ represent building segmentation labels, edge labels and body labels, where the edge label is obtained by traditional Canny edge detection [41] and the body is obtained by subtracting the edge label from the building segmentation label. $F_{\text{seg}}, F_{\text{edge}}, F_{\text{body}}$ represent building prediction results, edge prediction results and body prediction results, respectively. Corresponding labels and prediction maps two by two do loss to each other to get the total loss L as follows.

Sub represents subtractive operation.

$$L = \lambda_1 L_{\text{seg}} + \lambda_2 L_{\text{edge}} + \lambda_3 L_{\text{body}} \tag{13}$$

where $L_{\text{seg}}, L_{\text{edge}}, L_{\text{body}}$ denote final building segmentation loss, edge loss, and body loss, respectively. $\lambda_1, \lambda_2, \lambda_3$ are the hyperparameters that controls the weighting between the three losses.

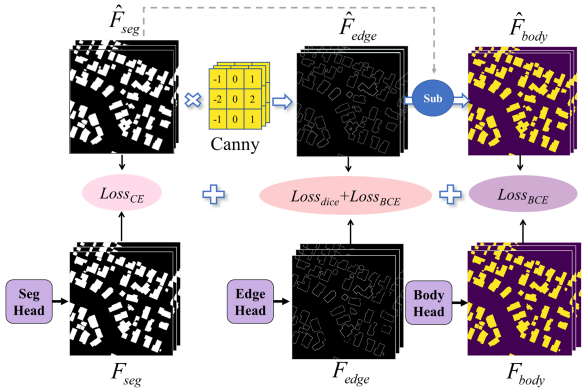

Fig. 7. Structure of body-mask-edge consistency constrains the loss of function.

The specific losses for each component are defined as follows:

$$L_{\text{seg}} = -\frac{1}{K}\sum_{i=1}^{N} \hat{F}_{\text{seg},i} \cdot \rho\left(\hat{F}_{\text{edge},i}\right) \cdot \log\left(F_{\text{seg},i}\right) \tag{14}$$

$$\begin{aligned} L_{\text{edge}} &= \lambda_4 loss_{\text{BCE}}\left(F_{\text{edge}}, Canny(\hat{F}_{\text{seg}})\right) \\ &\quad + \lambda_5 loss_{\text{dice}}\left(F_{\text{edge}}, Canny(\hat{F}_{\text{seg}})\right) \\ &= \lambda_4 loss_{\text{BCE}}\left(F_{\text{edge}}, \hat{F}_{\text{edge}}\right) + \lambda_5 loss_{\text{dice}}\left(F_{\text{edge}}, \hat{F}_{\text{edge}}\right) \end{aligned} \tag{15}$$

$$\begin{aligned} L_{\text{body}} &= loss_{\text{BCE}}\left(F_{\text{body}}, \hat{F}_{\text{seg}} - \hat{F}_{\text{edge}}\right) \\ &= loss_{\text{BCE}}\left(F_{\text{body}}, \hat{F}_{\text{body}}\right) \end{aligned} \tag{16}$$

$$\begin{aligned} \text{loss}_{\text{BCE}}\left(y_i, \tilde{y}_i\right) &= -\frac{1}{N}\sum_{i} y_i \times \log\left(\tilde{y}_i\right) \\ &\quad + (1 - y_i) \times \log\left(1 - \tilde{y}_i\right) \end{aligned} \tag{17}$$

$$\text{loss}_{\text{dice}}(y_i, \tilde{y}_i) = 1 - \frac{2\sum_i^N y_i \times \tilde{y}_i + 1}{\sum_i^N y_i^2 + \sum_i^N \tilde{y}_i^2 + 1} \tag{18}$$

where $F_{\text{seg},i}$ is the predicted probability of pixel i, $\hat{F}_{\text{seg},i}$ is the label value of pixel i, N denotes total number of pixels. Canny$(\cdot)$ indicates Canny edge detector.

For the overall loss L, we empirically set the parameters $\lambda_1 = 1$, $\lambda_2 = 25$, $\lambda_3 = 1$ and $\lambda_4 = 1$ according to their loss magnitudes to balance the small percentage of pixels on the edge areas. For L_{seg}, we use edge prior knowledge combined with weighted cross entropy loss [42]. First, let the total number of edge pixels be n, when pixel i is on the building boundary, $\rho[\hat{F}_{\text{edge},i}] = N - n/N$, and when pixel i is on a nonbuilding edge or background area, $\rho[\hat{F}_{\text{edge},i}] = n/N$. The goal is to balance the situation where the edge pixels account for a small fraction of the building segmentation by rewarding the model for paying more attention to the segmentation at the edges. Finally, we set $K = 0.1 \times n$ to balance positive and negative sample imbalances. For L_{edge}, the best performance of the model is achieved when we set $\lambda_5 = 0.4$ or 0.004. See ablation experiments in Section III-F-4) for a detailed description.

Body-mask-edge consistency constraint loss function can can divide different regions according to the structural features of the building and assign different weights to explore the structural features in an end-to-end manner and refine the edge features to achieve optimal network performance.

III. Experiments and Discussion

A. Implementation Details

The experimental platform is equipped with an Intel Xeon E5 2650 processor (a 376 GB RAM, Intel Corporation, Santa Clara, California, USA), four NVIDIA 2080Ti 12G graphics cards (Nvidia Corporation, Santa Clara, California, USA). The deep learning framework uses pytorch-1.8, and NVIDIA's CUDA11.2 GPU running platform and cuDNN8.0 deep learning GPU acceleration library. During network training, the proposed network is set training parameters batch size of 12, epoch of 131, initial learning rate of 0.001, and optimized using Adam optimizer. During the training process, the learning rate is adjusted by decaying every 10 epoch indices with a decay coefficient of 0.1.

B. Dataset

To evaluate the effectiveness of the proposed MDBES-Net in this article, we perform validation experiments using the Massachusetts building dataset [43] and the WHU aerial building dataset [44]. The WHU aerial building dataset consists of aerial and satellite imagery. In our experiments, we select the widely used aerial subset to evaluate the MDBES-Net algorithm has effectiveness and robustness. It consists of more than 220 000 individual buildings covering an area of more than 450 square kilometers with a ground resolution of 0.3 m. 4912 of these images were used as the training set, 1637 images for the validation set, and 1638 images for the test set, and each image was cropped to 512 × 512 pixels. The Massachusetts building dataset consists of 151 aerial images of the Boston area, each of which is 1500 × 1500 pixels in size, with a spatial resolution of 1 m and an area of 2.25 square kilometers. Meanwhile, the dataset covers an area of about 340 square kilometers, of which 91 images are used as a training set, 30 for testing, and 30 for validation. To facilitate training and evaluation, each image was cropped to 256 × 256 pixels.

C. Evaluation Metrics

To quantitatively facilitate the analysis of the segmentation performance of the proposed MDBES-Net, the semantic segmentation evaluation metrics used in this article are: precision; recall; intersection over union (IoU); and F1-score. The specific formula is expressed as follows:

$$\mathrm{IoU} = \mathrm{TP}/\left(\mathrm{FN} + \mathrm{FP} + \mathrm{TP}\right) \tag{19}$$

$$\mathrm{F1score} = 2 \times \mathrm{TP}/\left(2 \times \mathrm{TP} + \mathrm{FN} + \mathrm{FP}\right) \tag{20}$$

$$\mathrm{Recall} = \mathrm{TP}/\left(\mathrm{FN} + \mathrm{TP}\right) \tag{21}$$

$$\mathrm{Precision} = \mathrm{TP}/\left(\mathrm{FP} + \mathrm{TP}\right) \tag{22}$$

where, TP, TN, FP, FN represent the true positives, true negatives, false positives, and false negatives, respectively. N represents the total number of samples.

To quantitatively assess the structural characteristics of learning edges. We apply two boundary metrics in our analysis: the Hausdorff distance (HD) [45] and the structural similarity (SSIM) [46], for effective quantitative evaluation. This is described as follows:

$$\begin{aligned} dt_{\mathrm{HD}}(T,P) &= \max\{dt_{\mathrm{TP}}, dt_{\mathrm{TP}}\} \\ &= \max\{\max_{t\in T}\min_{p\in P}(t,p), \max_{p\in P}\min_{t\in T}(t,p)\} \end{aligned} \tag{23}$$

where T and P represent the ground true and prediction map. $dt_{HD}(t,p)$ is the shortest maximum distance from one point in a point-set to another point set.

To eliminate the effect of very small outlier subsets, multiplies HD by 95% to obtain the final evaluation metric (95% HD). HD distance is negatively correlated with the similarity of the labeled image shapes. SSIM is used to evaluate the similarity between two images. The range of SSIM is $(-1, 1)$. Its value is positively correlated with the similarity. This is as follows:

$$\begin{aligned} \mathrm{SSIM}(T,P) &= F(l(T,P)), c(T,P), s(T,P) \\ &= \frac{(2\mu_t\mu_p + C_1)(2\sigma_{tp} + C_2)}{(\mu_t^2 + \mu_p^2 + C_1)(\sigma_t^2 + \sigma_p^2 + C_2)} \end{aligned} \tag{24}$$

where μ, σ and σ_{tp} denote mean, variance and covariance, respectively. The constants C_1, C_2 are 6.50 and 58.52. $l(t,p)$ is luminance function, $c(t,p)$ is contrast function and $s(t,p)$ is structure function.

D. Results and Discussion

1) Analysis of Experimental Results: On the WHU dataset, comparison experiments are conducted with Unet [47], DeepLabv3+ [48], PSPNet [49], Map-Net [50], BOMSC-net [28], DR-Net [51], MBR-HRNet [52], and CFENet [53] networks, respectively. The findings from the qualitative experiments are presented in Fig. 8. The obstacles impacting segmentation accuracy in the WHU data are broadly classified into five main types of problems: (a) transformations of varying roof materials; (b) sizable buildings; (c) shadow masking; (d) and complicated building structures.

In the given roof material transformation case in (a), Unet and PSPnet exhibit considerable instances of false red classification, while MDBES-Net avoids semantically segregating metal shelves into buildings. For row (b) medium to large buildings, BOMSC-net reveals more instances of missed blue areas for extended large buildings, while MDBES-Net provides a more comprehensive extraction for larger buildings. For the shaded tree occlusion in row (c), the BOMSC-net and PSPnet networks exhibit greater instances of missed predictions for blue occluded parts. In contrast, the MDBES-Net demonstrates accurate predictions with more regular edge areas. Regarding the complex structural buildings in row (d), Unet and DeepLabv3+ demonstrate more blue omissions and red misclassifications in the structurally complex edge regions. Conversely, MDBES-Net's

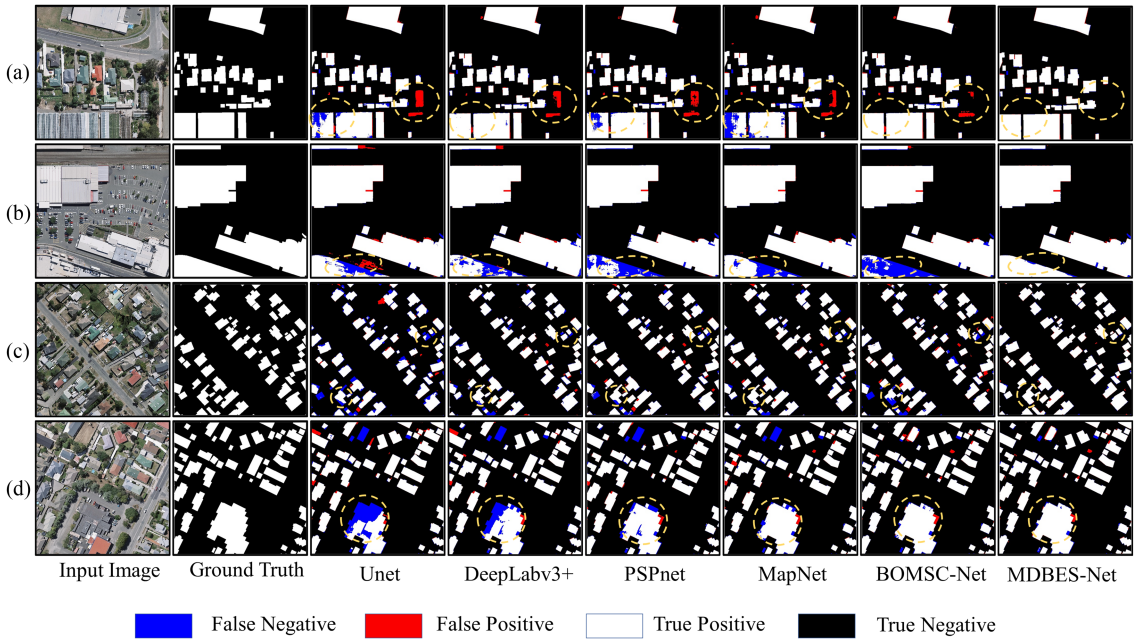

Fig. 8. (a–d) Comparison of the results obtained from the WHU test dataset using different segmentation methods (comparative experiment).

TABLE I
QUANTITATIVE EVALUATION ON THE WHU BUILDING DATASET

Network	Common Metrics				Boundary Metrics		Params (M)
	IoU (%)	Precision (%)	Recall (%)	F1-score (%)	95% HD	SSIM (%)	
Unet	85.51	91.88	92.52	92.77	93.11	88.83	17.26
PSPNet	86.12	93.46	92.55	92.73	88.47	91.89	53.58
DeepLabv3+	85.78	92.51	91.06	92.20	97.58	89.03	15.31
DR-Net	88.30	94.31	94.38	93.84	91.35	90.74	9.00
CFENet	89.30	94.30	94.39	94.35	88.56	91.86	171.00
MAP-Net	89.94	95.59	93.84	94.70	84.24	92.36	24.00
BOMSC-Net	90.15	95.14	94.50	94.80	80.03	92.43	129.32
MBR-HRNet	91.31	95.48	94.88	95.18	79.37	92.89	31.02
MDBES-Net	**91.78**	**95.74**	**95.63**	**95.68**	**77.69**	**93.84**	26.42

predictions have sharper edges and are closer to the ground truth overall.

The quantitative evaluation results of MDBES-Net are given in Table I. Our MDBES-Net outperforms Unet, PSPNet, DeepLabv3+, DR-Net, CFENet, MAP-Net, BOMSC-Net, and MBR-HRNet in terms of IoU, precision, recall, and F1-score. Our network has an IoU of 91.78, a precision of 95.74, a recall of 95.63, an F1-score of 95.68 on the WHU dataset. Additionally, it performed best in 95% HD and SSIM with 77.69 and 93.84, respectively. Due to the combination of our designed DBAE and CDA modules, we can effectively distinguish the edge features and body region features, mine the spatial structure features as well as enhance edge detail features. This process preserves texture information while fully extracting edges.

For the Massachusetts dataset, there are four main building types: coastal marinas, densely distributed small buildings (with a high number of buildings per unit pixel), large buildings, and shadow occlusion. As illustrated in Fig. 9, MDBES-Net effectively mitigates misclassification of densely clustered buildings in types (a), (b), and (c) compared to traditional segmentation methods such as Unet, PSPNet, and DeepLabv3+. For sizable constructions containing intricate frameworks and obstructed edges, both MAP-net and BOMSC-net exhibit increased blue omissions and red misclassifications in their latest segmentation results, while MDBES-Net exhibits a superior understanding of the edges and internal consistency at the feature decoupling level. This results in sharper and better segmentation results at the edges, compared the other networks evaluated.

Table II gives the evaluation results. Our MDBES-Net consistently achieves the highest IoU, precision, recall, and F1-score compared to Unet, PSPNet, DeepLabv3+, DR-Net, CFENet, MAP-Net, BOMSC-Net, and MBR-HRNet networks. Specifically, it has an IoU of 75.55, a precision of 86.88, a recall of 84.21, and an F1-score of 85.52 on the Massachusetts dataset. Additionally, it performed best in 95% HD and SSIM, with scores of 203.75 and 82.17, respectively. The data indicates that

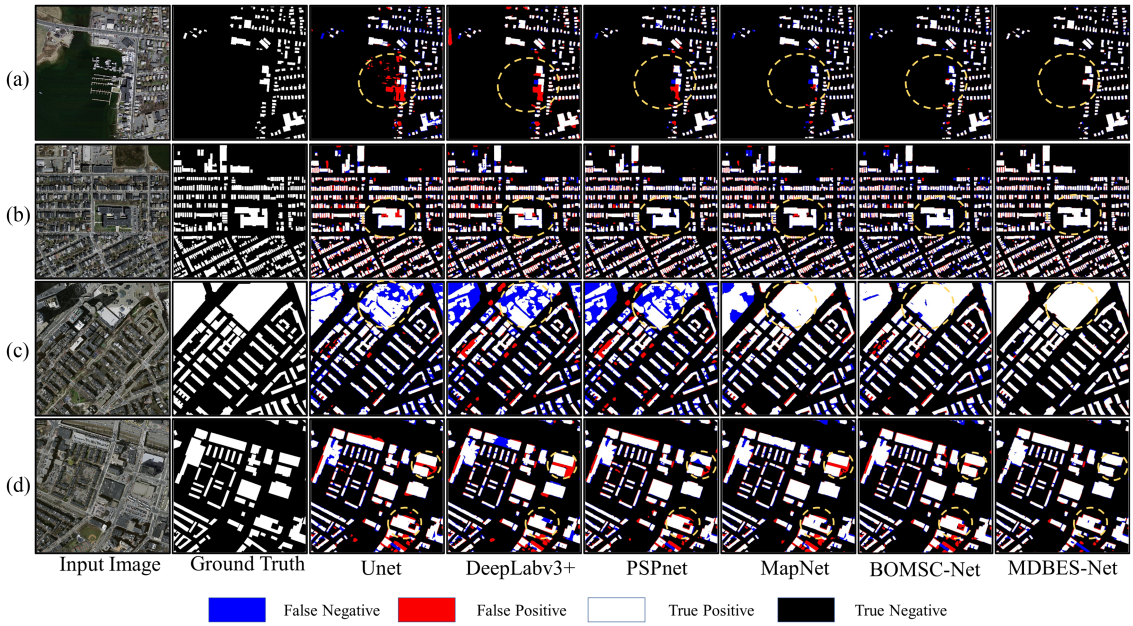

Fig. 9. Comparison of the results obtained from the massachusetts building test dataset using different segmentation methods (comparative experiment).

TABLE II
QUANTITATIVE EVALUATION ON THE MASSACHUSETTS BUILDING DATASET

Network	Common Metrics				Boundary Metrics		Params (M)
	IoU (%)	Precision (%)	Recall (%)	F1-score (%)	95% HD	SSIM (%)	
Unet	69.98	80.36	84.40	82.34	321.00	74.12	17.26
PSPNet	68.04	79.76	82.24	81.00	331.05	75.92	53.58
DeepLabv3+	67.38	78.44	82.75	73.47	315.31	70.94	15.31
DR-Net	66.00	80.77	83.12	79.50	343.44	74.82	9.00
CFENet	68.02	79.35	82.68	80.97	332.18	75.73	171.00
MAPNet	71.51	86.84	80.20	83.39	289.68	77.13	24.00
BOMSC-Net	74.71	86.64	83.68	85.13	217.10	80.54	129.32
MBR-HRNet	70.97	86.40	80.85	83.53	296.94	77.89	31.02
MDBES-Net	**75.55**	**86.88**	**84.21**	**85.52**	**203.75**	**82.17**	26.42

MDBES-Net demonstrates robustness in identifying low spatial resolution buildings and refining their edges. This improvement is particularly noticeable in the Massachusetts dataset when compared to other networks.

F. Ablation Analyses

1) Influence of Different Modules: To evaluate the impact of each component of the MDBES-NET model on the experimental findings, we conducted an ablation study on the WHU dataset and the Massachusetts dataset. The proposed MDBES-NET consists of the BMECC base network BMECC, the DBEA, and the CDA module CDA.

In Fig. 10, we have selected four representative building maps: (a) complex structures; (b) large buildings; (c) buildings with overshadowing; and (d) material transformations and roof shading. BMECC already has a basic building extraction capability, but lacks internal consistency with the presence of holes within the large buildings in row (a) and row (b). The results in column (D) show that the combination of BMECC + CDA modules significantly improves the building extraction capability, improving the hole phenomenon for large buildings in row (b).The results in column (E) show that the predicted segmentation maps for the combination of BMECC + DBEA modules are more advantageous, eliminating the hole phenomenon for large buildings in row (a) and row (b), and the extraction results are inherently uniform, with the building edge predictions in row (d) more closely matching the Ground Truth's edges. The results in column (F) show that the MDBES-net building prediction results are clearer, with more continuous and regular predictions at the edges, and accurate segmentation can be achieved at the corners of the edges of the complex buildings in row (a). The results in column (G) show that BMECC + DBEA + CDA learns the structural features of the boundaries, successfully predicts the building edges, and that the edge predictions are closer to the ground truth.

As given in Table III, on the WHU dataset, BMECC has an IoU of 84.79, a precision of 91.98, a recall of 92.52, an

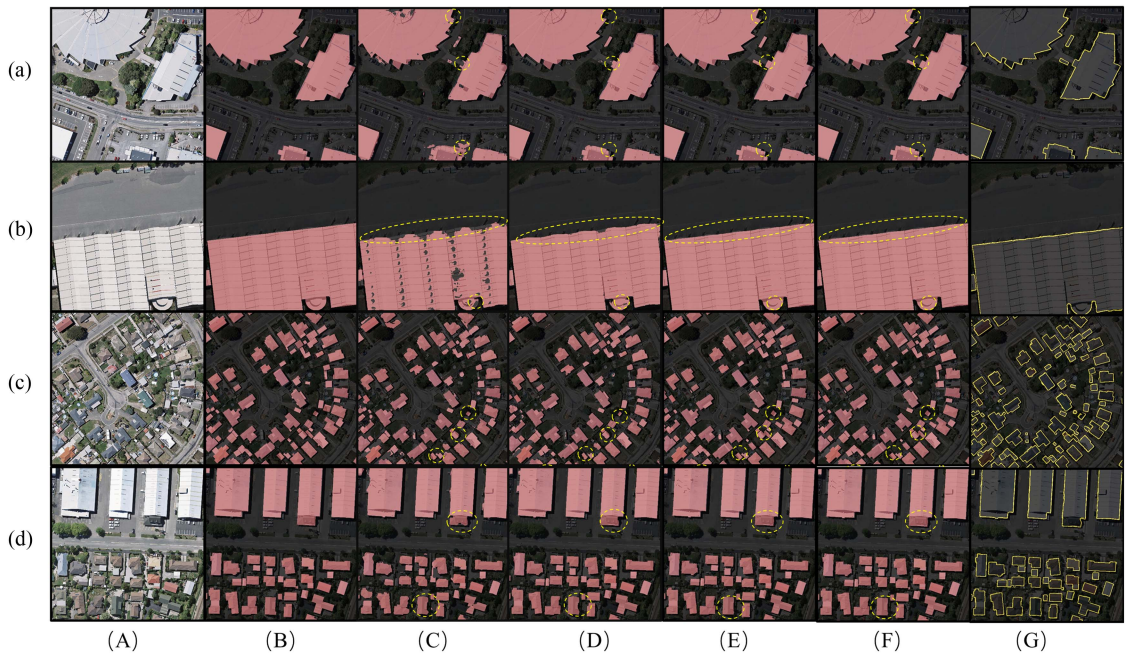

Fig. 10. (a–d) Visualization of prediction results of different modules (comparative experiment). (a) Input images. (b) Ground truth. (c) "BMECC" segmentation results. (d) "BMECC + CDA" segmentation results. (e) "BMECC + DBEA" segmentation results. (f) "MDBES-NeT" segmentation results. (g) "MDBES-NET" edge segmentation results.

TABLE III
QUANTITATIVE EVALUATION OF ABLATION EXPERIMENTS

Dataset	Method	Common Metrics				Boundary Metrics	
		IoU (%)	Precision (%)	Recall (%)	F1-score (%)	95%HD	SSIM (%)
WHU	BMECC	84.79	91.98	92.52	92.25	93.11	88.83
	+DBEA	90.57	95.22	94.83	95.02	79.44	92.75
	+CDA	87.40	93.31	93.84	93.57	84.24	91.06
	MDBES-Net	**91.78**	**95.74**	**95.63**	**95.68**	**77.69**	**93.84**
Massachusetts	BMECC	69.70	83.63	82.00	82.81	303.74	77.24
	+DBEA	74.54	86.12	83.99	85.04	254.50	80.87
	+CDA	71.85	84.44	83.09	83.76	268.43	78.63
	MDBES-Net	**75.55**	**86.88**	**84.21**	**85.52**	**203.75**	**82.17**

F1-score of 92.25, a 95%HD of 93.11 and a SSIM of 88.83. BMECC+DBEA have an IoU of 90.57, a precision of 95.22, a recall of 94.83, an F1-score of 95.02, a 95%HD of 79.44 and a SSIM of 92.75. BMECC+CDA have an IoU of 87.40, a precision of 93.31, a recall of 93.84, an F1-score of 93.57, a 95%HD of 84.24 and a SSIM of 91.06. MDBES-Net has an IoU of 91.78, a precision of 95.74, a recall of 95.63, an F1-score of 95.68, a 95%HD of 77.69 and a SSIM of 93.84. On the Massachusetts, BMECC has an IoU of 69.70, a Precision of 83.63, a Recall of 82.00, an F1-score of 82.81, a 95%HD of 303.74 and a SSIM of 77.24. BMECC+DBEA have an IoU of 74.54, a precision of 86.12, a recall of 83.99, an F1-score of 85.04, a 95%HD of 254.50 and a SSIM of 80.87. BMECC+CDA have an IoU of 71.85, a Precision of 84.44, a recall of 83.09, an F1-score of 83.76, a 95%HD of 268.43 and a SSIM of 78.63. MDBES-Net has an IoU of 75.77, a precision of 86.88, a recall of 84.21, an F1-score of 85.52, a 95%HD of 203.75 and a SSIM of 82.17.the DBEA module can improve the learning ability of edge features through edge optimization and region segmentation accuracy, and better contribute to the improvement of the overall model. CDA can effectively alleviate the incomplete building extraction and improve the feature characterization. MDBES-Net fully integrates the advantages of DBEA and CDA models, which can realize the complete description of buildings and the refinement of building boundary information.

2) Influence of the DBEA: In order to verify the effectiveness of the DBEA module, this article will represent the DBEA advantage in qualitative visualization and quantitative. As shown in Fig. 11, the semantic flow of the column in (B) resembles the optical flow, which is color-coded to shrink inward from the boundary of a building to a location inside, typically near the center of the building. As can be seen in Fig. 11, the background of the body feature in columns (C) and (D) is clear, noiseless, and uniform with no gaps. The edge feature is regular, continuous,

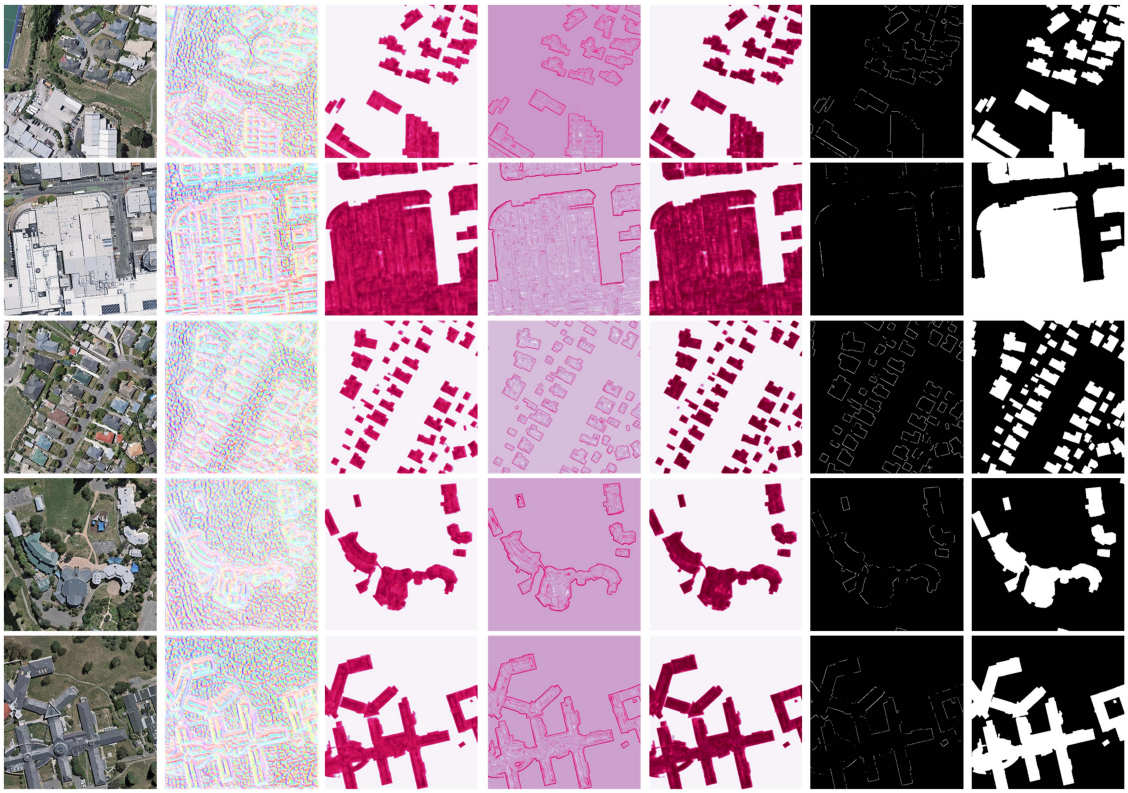

Fig. 11. Visualization results of decoupled feature map and predicted results. (a) Original image. (b) Learnable semantic flow field. (c) Body feature map. (d) Edge feature map. (e) Redecoupling feature map. (f) Building edge prediction result. (g) Building segmentation result.

and has well defined rectangular boundaries. Column (E) shows an overall improvement in the building's features and more refinement in the boundary features. Meanwhile, column (F) suggests that the proposed network remains capable of accurately predicting building edges for different building types.

In summary, since DBEA utilizes the potential flow pattern distribution laws among different frequency features to capture the body and edge features of the building structure, it has the ability to mine the target boundaries to divide the region to gain scene knowledge. Therefore, when extracting different types of buildings in different complex scenes, the proposed MDBES-Net can successfully decouple the building structure features and learn strong coupling pairs of body features and edge features, thereby optimizing the overall performance of the network.

To investigate the effectiveness of the body and edge decoupled sensing module DBEA at different decoding stages, we visualize the feature maps of both the module without and with DBEA at the decoding layer to highlight the advantages of the decoupling properties of the DBEA module. As shown in Fig. 12, y_1, y_2, and y_3 denote the first, second and third level of feature visualization in the decoding phase when using DBEA, respectively. ${y'}_1$, ${y'}_2$, and ${y'}_3$ indicate feature visualization map without DBEA. ${y''}_1$ denotes the feature visualization map using both DBEA and CDA modules at the first level of decoding. The experimental results show that the y_1, y_2, and y_3 building features without DBEA module are poorly characterized, especially in the deeper layer y_3 of the network, the building features are blurred, mixed with the background noise, and the building features cannot be effectively extracted. At shallower layer y_1 and y_2 of the network, the internal features of the building are affected by the complexity of the roof structure and material changes, and the phenomenon of "holes" in the building features and boundary areas receives little attention. However, the features of buildings ${y'}_1$, ${y'}_2$, and ${y'}_3$ using the DBEA module has been significantly enhanced, with an emphasis on the representation of the building ontology. In the deeper layer ${y'}_3$ of the network, architectural features are prominent, effectively distinguishing the building from the background noise, enabling the extracting of more complete architectural features. At shallower layers ${y'}_1$ and ${y'}_2$ of the network, the interference of background noise is significantly suppressed, and the enhanced building features contrast sharply with the background, and the building features are complete and uniform inside, with more regular edges. In particular, the marginal area of the ${y'}_1$ does not extend outward, while the attention to tiny buildings is enhanced and their boundaries are clearer.

Table IV compare the effectiveness of DBEA modules at different decoding stages, this article conducted experiments based on the BMECC base network. +DBEA (with y_l) indicates that the DBEA module is used at layer l of the decoding. on the WHU dataset, BMECC+DBEA (with y_3) has an IoU of 85.75, a Precision of 92.23, a Recall of 93.50, an F1-score of 92.86, a 95%HD of 89.87 and a SSIM of 89.66. +DBEA (with y_2,

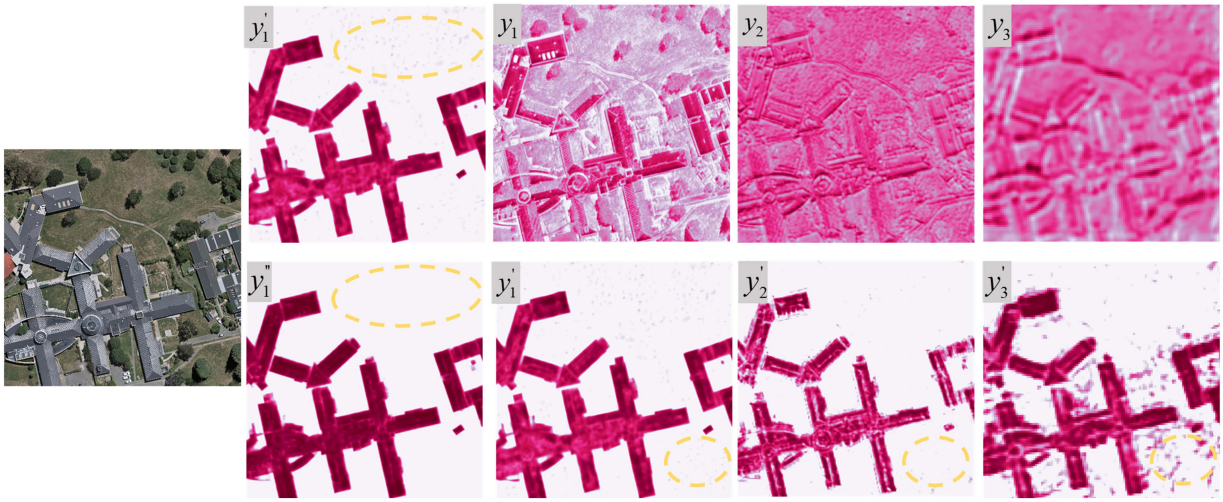

Fig. 12. Visualized feature maps of DBEA modules at different scales.

TABLE IV
INFLUENCE OF DBEA MODULE

Dataset	Method	Common Metrics				Boundary Metrics	
		IoU (%)	Precision (%)	Recall (%)	F1-score (%)	95%HD	SSIM (%)
WHU	BMECC	84.79	91.98	92.52	92.25	93.11	88.83
	+DBEA (with y_3)	85.75	92.23	93.50	92.86	89.87	89.66
	+DBEA (with y_2, y_3)	87.85	93.54	94.06	93.80	83.33	90.87
	+DBEA (with y_1, y_2, y_3)	**90.57**	**95.22**	**94.83**	**95.02**	**79.89**	**92.46**
Massachusetts	BMECC	69.70	83.63	82.00	82.81	303.74	77.24
	+DBEA (with y_3)	70.00	84.11	81.88	82.98	292.31	78.53
	+DBEA (with y_2, y_3)	72.52	85.38	82.93	84.13	277.32	79.08
	+DBEA (with y_1, y_2, y_3)	**74.54**	**86.12**	**83.99**	**85.04**	**254.50**	**80.87**

y_3) have an IoU of 87.85, a Precision of 93.54, a Recall of 94.06, an F1-score of 93.80, a 95%HD of 83.33 and a SSIM of 90.87. +DBEA (with y_1, y_2, y_3) have an IoU of 90.57, a Precision of 95.22, a Recall of 94.83, an F1-score of 95.02, a 95%HD of 79.89 and an SSIM of 92.46. On the Massachusetts, BMECC+DBEA (with y_3) has an IoU of 70.00, a precision of 84.11, a recall of 81.88, an F1-score of 82.98, a 95%HD of 292.31 and a SSIM of 78.53. +DBEA (with y_2, y_3) have an IoU of 72.52, a precision of 84.54, a recall of 82.14, an F1-score of 84.13 a 95%HD of 277.32 and an SSIM of 79.08. +DBEA (with y_1, y_2, y_3) have an IoU of 74.54, a precision of 86.12, a recall of 83.99, an F1-score of 85.04, a 95%HD of 254.50 and a SSIM of 80.87.

In summary, the DBEA module can successfully decouple the structural features of the building at multiple scales and explore the strongly coupled structural features of the interaction between the body and the edge, focus on the features of the building body, and suppress background noise interference. By learning the warping of the flow field, more precise boundaries can be generated, and these precise boundaries contribute to better segmentation results.

3) Influence of the CDA: To analysis the impact of CDA, we compared the results of all operators and without CDA. As given in Table V. On the WHU dataset, all operators have an IoU of 91.78, a precision of 95.74, a recall of 95.63, an F1-score of 95.68, a 95%HD of 77.69 and a SSIM of 93.84 and w/o CDA have an IoU of 90.57, a precision of 95.22, a recall of 94.83, an F1-score of 95.02, a 95%HD of 79.44 and a SSIM of 92.75. On the Massachusetts dataset, all operators have an IoU of 75.55, a precision of 86.88, a recall of 84.21, an F1-score of 85.52, a 95%HD of 203.75 and a SSIM of 82.17. and w/o CDA have an IoU of 74.54, a precision of 86.12, a recall of 83.99, an F1-score of 85.04, a 95%HD of 254.50 and a SSIM of 80.87. As shown in Fig. 12, y'_1 and y''_1 represent the characteristic mapping without CDA and with CDA, respectively. The experimental results show that y''_1 has a stronger ability to perceive the global building than y'_1. It has a stronger ability to characterize complex building structures, especially at the edges, and the wave point noise in y'_1 is effectively suppressed.

In summary, the qualitative and quantitative analysis shows that the channel decoupling attention module CDA can effectively suppress extraneous background noise, improve the

TABLE V
INFLUENCE OF CDA MODULE

Dataset	Settins	Common Metrics				Boundary Metrics	
		IoU (%)	Precision (%)	Recall (%)	F1-score (%)	95%HD	SSIM (%)
WHU	All operators	**91.78**	**95.74**	**95.63**	**95.68**	**77.69**	**93.84**
	w/o CDA	90.57	95.22	94.83	95.02	79.44	92.75
Massa-chusetts	All operators	**75.55**	**86.88**	**84.21**	**85.52**	**203.75**	**82.17**
	w/o CDA	74.54	86.12	83.99	85.04	254.50	80.87

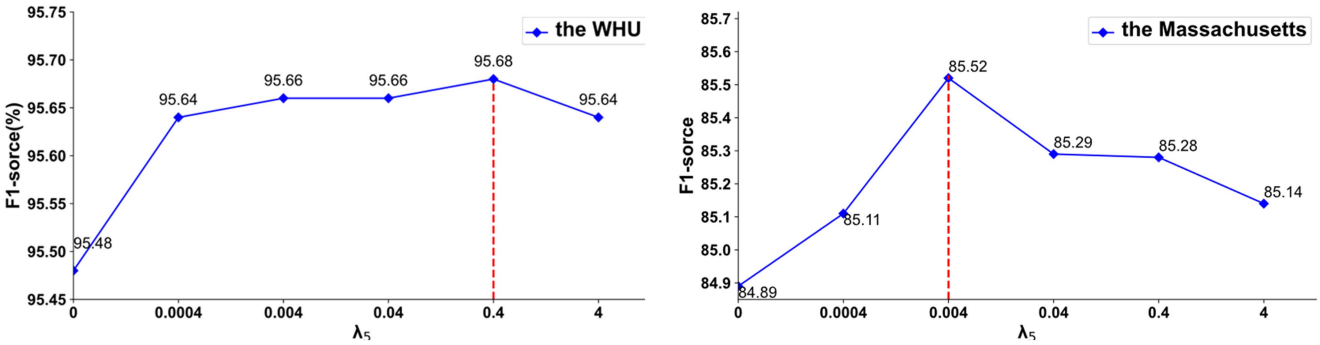

Fig. 13. Ablation study for F1-score at different hyperparameters λ_5.

characterization of building features and regularize boundary details, and CDA supports DBAE in achieving the best network segmentation.

4) Influence of the Parameters λ_5: In this article, the ablation experiment was performed out in both the WHU and the Massachusetts datasets and we fine-tune the size of λ_5 to observe the change in F1-sorce. λ_5 represent the weight of $\text{loss}_{\text{dice}}$in the edge head, The horizontal coordinates represent the λ_5 values, the vertical coordinates represent the F1-sorce values as shown in Fig. 13, In the WHU dataset, different λ_5-values of 0, 0.0004, 0.004, 0.04, 0.4, and 4 were set up, and then the corresponding F1 values of 95.48, 95.64, 95.66, 95.66, 95.68, and 95.64 were obtained. when $\lambda_5 = 0.4$, F1-sorce achieves a maximum value of 95.68. In the Massachusetts dataset, the corresponding F1 values of 95.48, 95.64, 95.66, 95.66, 95.68, and 95.64 were obtained. when $\lambda_5 = 0.004$, F1-sorce achieves a maximum value of 85.52. When the loss weighting factor λ_5 in $\text{loss}_{\text{dice}}$ is relatively larger, it is more advantageous in the WHU dataset with high resolution, complete structural information. When λ_5 is relatively small, the segmentation effect is better, although the Massachusetts dataset has a lower resolution, the buildings are densely distributed, and the positive and negative sample are relatively balanced.

In summary, the addition of $\text{loss}_{\text{dice}}$ in the edge supervision can efficiently assist the network to mine the structural feature information of the positive building samples, significantly improve the building segmentation accuracy, and optimize the overall performance of the proposed network.

5) Influence of Different Architectures: In this article, the effects of different encode-decode frameworks and backbones on the network. Semantic segmentation frameworks are FCN [54], PSPNet, DeepLabv3+, and Unet. As given in Table VI, On the WHU dataset, FCN+our modules attained an F1-score of 94.17 and an IoU of 86.12. PSP+our modules attained an F1-score of 95.04 and an IoU of 90.88. Combination of DeepLabv3+ and our modules attained an F1-score of 94.82 and an IoU of 88.29. Unet+our modules achieved the highest F1-score of 95.68 and an IoU of 91.78. On the Massachusetts dataset, FCN+our modules attained an F1-score of 81.91 and an IoU of 69.65. PSP+our modules attained an F1-score of 84.28 and an IoU of 71.67. Combination of DeepLabv3+ and our modules attained an F1-score of 75.77 and an IoU of 71.56. Unet+our modules achieved the highest F1-score of 85.52 and an IoU of 75.55.

TABLE VI
APPLICATION ON OTHER ARCHITECTURES

Dataset	Network	F1-sorce (%)	IoU (%)
WHU	FCN	90.39	82.46
	+our modules	94.17	86.43
	PSPNet	92.73	86.12
	+our modules	95.04	90.88
	DeepLabv3+	92.20	85.78
	+our modules	94.82	88.39
	Unet	92.77	85.51
	+our modules	**95.68**	**91.78**
Massa-chusetts	FCN	78.76	65.24
	+our modules	81.91	69.65
	PSPNet	81.00	68.04
	+our modules	84.28	71.67
	DeepLabv3+	73.47	67.38
	+our modules	75.77	71.56
	Unet	82.34	69.98
	+our modules	**85.52**	**75.55**

As given in Table VII, we employed a backbone consisting of MobileNetv2, Xception65, ResNet50, ResNet101, and

TABLE VII
RESULTS OF USING DIFFERENT BACKBONE NETWORK

Method	Dataset	Backbone	F1-sorce (%)
MDBES-Net	WHU	MobileNetv2	86.45
		Xception65	90.80
		ResNet50	91.31
		ResNet101	95.03
		ResNeXt101	95.70
	Massachu-setts	MobileNetv2	78.91
		Xception65	80.50
		ResNet50	80.83
		ResNet101	85.17
		ResNeXt101	85.55

TABLE VIII
QUANTITATIVE EVALUATION OF PARAMETERS OF DIFFERENT MAIN MODULES

	BMECC	BMECC+ DBEA	BMECC+ CDA	MDBES -Net
Params (M)	26.17	26.19 (+0.02)	26.40 (+0.23)	26.42 (+0.25)

ResNeXt101. On the WHU dataset, MobileNetv2 attained an F1-score of 86.45. Xception65 attained an F1-score of 90.80. ResNet50 attained an F1-score of 91.31. ResNet101 attained an F1-score of 95.03. ResNeXt101 attained the highest F1-score of 95.70. On the Massachusetts dataset, MobileNetv2 attained an F1-score of 78.91. Xception65 attained an F1-score of 80.50. ResNet50 attained an F1-score of 80.83. ResNet101 attained an F1-score of 85.17. ResNeXt101 attained the highest F1-score of 85.55.

6) Parameter Analysis: In this article, the number of participants in the different modules was counted. As given in Table VIII, compared to the BMECC baseline network, the number of parameters for "BMECC + DBEA" is only increased by 0.02 M, and the number of parameters for "BMECC + CDA" is increased by 0.23 M. Therefore, MDBES-Net is a lightweight semantic segmentation model.

IV. CONCLUSION

To address the challenge of low segmentation accuracy caused by incomplete and discontinuous internal building segmentation as well as blurred edges in remote sensing images, a building extraction from remote sensing images based on a MDBES-Net is proposed. First, the BMECC network is built with body-mask-edge consistency constraints, allowing for hierarchical extraction of building structure features and ensuring the extracted features are rich. Second, by DBEA modules, the multiscale feature map can be split into two parts: edge features and body features. The body part enhances the building's internal consistency, while the edge part is more detailed and regular, resulting in significantly improved feature characterization ability upon coupling. Finally, the CDA module effectively reduces interference from external background noise and enhances the smoothness of building edges. Furthermore, quantitative, qualitative, and ablation experiments on publicly available and aerial image datasets, as well as the Massachusetts dataset, provide evidence of the algorithm's effectiveness and robustness. In future work, we will try to improve the decoupling process and use fine edge features guidance to improve the model results, fusing multimodal digital elevation model data for eventual application to other types of target extraction.

REFERENCES

[1] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.
[2] L. Yang, H. Wang, K. Yan, X. Yu, J. Li, and D. Man, "Building extraction of multi-source data based on deep learning," in *Proc. IEEE 4th Int. Conf. Image, Vis. Comput.*, 2019, pp. 296–300.
[3] Y. Lin, H. He, Z. Yin, and F. Chen, "Rotation-invariant object detection in remote sensing images based on radial-gradient angle," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 746–750, Apr. 2015.
[4] R. Shang, M. Liu, L. Jiao, J. Feng, Y. Li, and R. Stolkin, "Region-level SAR image segmentation based edge feature label assistance," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5237216.
[5] M. Wang, X. Zheng, and C. Feng, "Color constancy enhancement for multi-spectral remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 864–867.
[6] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 487–491, Mar. 2015.
[7] G. Liasis and S. Stavrou, "Building extraction in satellite images using active contours and colour features," *Int. J. Remote Sens.*, vol. 37, pp. 1127–1153, 2016.
[8] L. Wang et al., "Active contours driven by edge entropy fitting energy for image segmentation," *Signal Process.*, vol. 149, pp. 27–35, 2018.
[9] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1835–1838.
[10] Q. Chen, M. Sun, X. Hu, and Z. Zhang, "Automatic seamline network generation for urban orthophoto mosaicking with the use of a digital surface model," *Remote Sens.*, vol. 6, pp. 12334–12359, 2014.
[11] Z. Zhang, W. Guo, M. Li, and W. Yu, "GIS-supervised building extraction with label noise-adaptive fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2135–2139, Dec. 2020.
[12] H. Daschiel and M. Datcu, "Information mining in remote sensing image archives: System evaluation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 1, pp. 188–199, Jan. 2005.
[13] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.
[14] P. Schuegraf and K. Bittner, "Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN," *ISPRS Int. J. Geo-Inf.*, vol. 8, 2019, Art. no. 191.
[15] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.
[16] Y. Li, H. Lu, Q. Liu, Y. Zhang, and X. Liu, "SSDBN: A single-side dual-branch network with encoder–decoder for building extraction," *Remote Sens.*, vol. 14, 2022, Art. no. 768.
[17] Z. Tang et al., "Capsule–encoder–decoder: A method for generalizable building extraction from remote sensing images," *Remote Sens.*, vol. 14, 2022, Art. no. 1235.
[18] A. Eftekhari, F. Samadzadegan, and F. D. Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 117, 2023, Art. no. 103180.
[19] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan, "CGSANet: A contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1526–1542, 2022.
[20] B. Xu, J. Xu, N. Xue, and G.-S. Xia, "HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 284–296, 2023.

[21] X. Liu, Y. Chen, C. Wang, K. Tan, and J. Li, "A lightweight building instance extraction method based on adaptive optimization of mask contour," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103420.
[22] Q. Zhu, Z. Li, Y. Zhang, and Q. Guan, "Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields," *Remote Sens.*, vol. 12, 2020, Art. no. 3983.
[23] S. He and W. Jiang, "Boundary-assisted learning for building extraction from optical remote sensing imagery," *Remote Sens.*, vol. 13, 2021, Art. no. 760.
[24] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5215512.
[25] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building extraction by frame field learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5887–5896.
[26] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6106–6120, Jul. 2021.
[27] Z. Liu, Q. Shi, and J. Ou, "LCS: A collaborative optimization framework of vector extraction and semantic segmentation for building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5632615.
[28] Y. Zhou et al., "BOMSC-net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618617.
[29] Y. Wang, X. Zeng, X. Liao, and D. Zhuang, "B-FGC-net: A building extraction network from high resolution remote sensing imagery," *Remote Sens.*, vol. 14, 2022, Art. no. 269.
[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[31] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based DenseNet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020.
[32] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.
[33] Y.-C. Su, T.-J. Liu, and K.-H. Liuy, "Multi-scale wavelet frequency channel attention for remote sensing image segmentation," in *Proc. IEEE 14th Image, Video, Multidimensional Signal Process. Workshop*, 2022, pp. 1–5.
[34] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 763–772.
[35] 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.*, 2020, pp. 11531–11539.
[36] D. Han, S. Yun, B. Heo, and Y. Yoo, "Rethinking channel dimensions for efficient model design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 732–741.
[37] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[38] D. Sousa and C. Small, "Joint characterization of sentinel-2 reflectance: Insights from manifold learning," *Remote Sens.*, vol. 14, 2022, Art. no. 5688.
[39] X. Li et al., "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 435–452.
[40] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28.
[41] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[42] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8792–8802.
[43] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.
[44] 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.
[45] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
[46] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[47] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[50] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.
[51] M. Chen et al., "DR-net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, 2021, Art. no. 294.
[52] G. Yan, H. Jing, H. Li, H. Guo, and S. He, "Enhancing building segmentation in remote sensing images: Advanced multi-scale boundary refinement with MBR-HRNet," *Remote Sens.*, vol. 15, 2023, Art. no. 3766.
[53] J. Chen, D. Zhang, Y. Wu, Y. Chen, and X. Yan, "A context feature enhancement network for building extraction from highresolution remote sensing imagery," *Remote Sens.*, vol. 14, 2022, Art. no. 2276.
[54] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

Shengjun Xu received the Ph.D. degree in engineering from Xi'an Jiaotong University, Xi'an, China, in 2013.

He is currently an Associate Professor with the School of Information and Control Engineering, Xi'an University of Architecture and Technology. His research interests include image processing, artificial intelligence and automation.

Miao Du received the B.S. degree in automation in 2017 from Xi'an University of Architecture and Technology, Xi'an, China, where he is currently working toward the M.S. degree in control science and engineering.

His major research interests include computer vision, deep learning, and remote sensing image processing.

Yuebo Meng received Ph.D. degree in engineering from Xi'an Jiaotong University, Xi'an, China, in 2014.

She is currently an Associate Professor with the School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, China. Her research interests include machine learning and intelligent building technology.

Guanghui Liu received Ph.D. degree in engineering from Xi'an University of Architecture and Technology, China, in 2014.

He is currently an Associate Professor with the School of Information and Control Engineering, Xi'an University of Architecture and Technology. His research interests include machine learning, intelligent building technology.

Jiuqiang Han received the B.E. degree in automatic control from Xi'an Jiaotong University, Xi'an, China, in 1977.

He is currently a Professor with the School of Electronic Science and Engineering, Xi'an Jiaotong University. His research interests include intelligent perception theory for visual recognition, artificial intelligence and automated systems, and intelligent control theory for model simulation.

Bohan Zhan received the M.S. degree in automation from Xi'an University of Architecture and Technology, Xi'an, China, in 2022.

He is currently with the Xi'an Jiaotong University, with research interests in machine vision and image processing.