remote sensing

MDPI

Article

Dual-Channel Semi-Supervised Adversarial Network for Building Segmentation from UAV-Captured Images

Wenzheng Zhang [1], Changyue Wu [1,*], Weidong Man [1,2,3,4] and Mingyue Liu [1,2,3,4]

[1] College of Mining Engineering, North China University of Science and Technology, Tangshan 063210, China; zhangwz@stu.ncst.edu.cn (W.Z.); manwd@ncst.edu.cn (W.M.); liumy917@ncst.edu.cn (M.L.)
[2] Hebei Industrial Technology Institute of Mine Ecological Remediation, Tangshan 063210, China
[3] Collaborative Innovation Center of Green Development and Ecological Restoration of Mineral Resources, Tangshan 063210, China
[4] Tangshan Key Laboratory of Resources and Environmental Remote Sensing, Tangshan 063210, China
* Correspondence: wcy0315@ncst.edu.cn

Abstract: Accurate building extraction holds paramount importance in various applications such as urbanization rate calculations, urban planning, and resource allocation. In response to the escalating demand for precise low-altitude unmanned aerial vehicle (UAV) building segmentation in intricate scenarios, this study introduces a semi-supervised methodology to alleviate the labor-intensive process of procuring pixel-level annotations. Within the framework of adversarial networks, we employ a dual-channel parallel generator strategy that amalgamates the morphology-driven optical flow estimation channel with an enhanced multilayer sensing Deeplabv3+ module. This approach aims to comprehensively capture both the morphological attributes and textural intricacies of buildings while mitigating the dependency on annotated data. To further enhance the network's capability to discern building features, we introduce an adaptive attention mechanism via a feature fusion module. Additionally, we implement a composite loss function to augment the model's sensitivity to building structures. Across two distinct low-altitude UAV datasets within the domain of UAV-based building segmentation, our proposed method achieves average mean pixel intersection-over-union (mIoU) ratios of 82.69% and 79.37%, respectively, with unlabeled data constituting 70% of the overall dataset. These outcomes signify noteworthy advancements compared with contemporaneous networks, underscoring the robustness of our approach in tackling intricate building segmentation challenges in the domain of UAV-based architectural analysis.

Keywords: adversarial network; building segmentation; dual channel; optical flow estimation; semi-supervision; UAV

Citation: Zhang, W.; Wu, C.; Man, W.; Liu, M. Dual-Channel Semi-Supervised Adversarial Network for Building Segmentation from UAV-Captured Images. *Remote Sens.* **2023**, *15*, 5608. https://doi.org/10.3390/rs15235608

Academic Editor: Riccardo Roncella

Received: 9 November 2023
Revised: 29 November 2023
Accepted: 1 December 2023
Published: 2 December 2023

1. Introduction

Architecture is the main carrier of human life and development. Building density contains the key information of urban development. Accurate building inspection data play a vital role in environmentally friendly urban planning, commercial planning, land use change detection, national defense construction, and disaster monitoring and early warnings [1,2]. Due to the diversity of building types and sizes and the influence of complex background environments, it is still a key research direction to accurately and efficiently extract buildings from high-resolution UAV images [3]. The emergence of deep convolutional neural networks has ushered in a revolutionary stride in semantic segmentation endeavors [4,5]. Nevertheless, these approaches frequently hinge on intricate pixel-level annotations and comprehensive building outlines, which are not readily obtainable for drone-sourced images [6,7]. Furthermore, the inherent feature similarity between buildings and their backgrounds might result in internal inconsistencies within the segmentation outcomes [8], potentially leading to the misclassification of buildings as background entities.

The attention mechanism in network models helps them to focus on areas of interest in images [9]. SENet [10] incorporates an attention mechanism into the channel dimension, enhancing the importance of each feature channel. Woo et al.'s [11] lightweight attention module combines channel and spatial attention modules to refine feature maps. Vaswani et al.'s [12] multi-head attention module uses parallel attention mechanisms to extract vital features across different feature spaces. However, attention mechanisms alone do not perform well in complex, multiangle scenes.

Dual-channel network strategies are widely used in this field because of their ability to capture features from multiple angles and achieve high precision in segmenting complex backgrounds. For example, the EtoE-Fusion dual-channel network proposed by Wei et al. [13] preserves both global and local information. Different from previous studies, Wen et al. [14] designed PDSNet, in which each image in the obtained dataset is a dual-channel image, which is connected via a 2D pavement image and a corresponding 3D pavement image. Although this design is able to extract information from higher dimensions, it may increase model complexity and computational resource consumption. Therefore, You et al. [15] proposed an end-to-end dual-channel integrated cross-layer residual algorithm (TIC-Net) based on deep learning, which can learn from the feature fusion and residual calculation of different semantic information, and then realize the joint mining of features.

However, it is important to note that as network structures become more complex, the need for large amounts of labeled data to ensure model accuracy increases. In response to these practical obstacles, a range of semi-supervised methodologies has emerged, including AffinityNet [16], AdvSemiSeg [17], SemiCycleGAN [18], and CCVC [19]. AffinityNet can obtain better segmentation results from a small amount of labeled data by adopting self-supervised pretraining and introducing an affinity clustering mechanism. While these methods have shown promise, they still face limitations in effectively dealing with complex morphological and textural features that are prevalent in drone-captured architectural images [16]. In this case, balancing building form and texture properties remains a challenge that requires further research [20].

Optical flow estimation plays an important role in image segmentation due to its sensitivity to changes in light and shade in image textures and object shapes [21]. Optical flow estimation tracks pixel movements and changes within an image, thereby proving invaluable for tackling a multitude of segmentation challenges, including foreground–background separation, object tracking, and boundary detection [22]. By analyzing pixel displacements within an image, optical flow estimation enhances the understanding of both the structural layout and motion dynamics within the image [23]. The challenges in building segmentation include sensitivity to illumination and weather conditions, and without constraints on optical flow estimation, it is difficult to capture complex textures and precise shapes [24].

This paper presents a dual-channel semi-supervised segmentation network within an adversarial network framework for UAV-based building segmentation. The network uses optical flow estimation channels and integrated features from the building-aware ASPP module within the Deeplabv3+ architecture. This method aims to comprehensively segment building structures and textures in drone-captured images. To address complexities from variations in lighting and textures, a supplementary method is proposed, which includes symmetry calculation, connection domain feature mapping, and consistency calculation of the convex hull area. Furthermore, we introduce the compound loss function and Zhu et al.'s [25] adaptive attention mechanism to solve the feature redundancy problem caused by the two-channel network strategy. We designed a composite loss function specifically tailored for building segmentation, aiming to extract crucial texture information while simultaneously directing the network's attention towards building shape and structure. Moreover, we introduced the adaptive attention mechanism within the feature fusion module to address the redundancy issue arising from the dual-channel strategy. The main contributions of this study can be summarized as follows:

(1) This paper presents a unique dual-channel semi-supervised segmentation network within an adversarial network framework, aimed at improving the accuracy of complex building image segmentation. The network efficiently combines optical flow estimation channels with building-aware ASPP (BA-ASPP) features. It incorporates advanced modules, including hierarchical channel attention modules (HCAM) and multilevel feature fusion modules (MFFMs), to achieve a comprehensive understanding of building structures and textures.
(2) To address challenges related to lighting and texture, this paper presents a method that complements optical flow results with building-related information, encompassing symmetry and connected domain features. This innovative approach substantially diminishes the dependency on labeled data, rendering it well suited for semi-supervised tasks with just a 30% labeled sample set.
(3) Our network contains complementary components such as an adaptive attention mechanism feature fusion module and a composite loss function.
(4) Our network was evaluated on the drone building dataset and the publicly available UDD6 [26] dataset.

2. Materials and Methods

2.1. Data Acquisition and Dataset Construction

In this study, Tangshan City, located in the northeast coastal area, was selected as the data collection area. Tangshan City is located at 39°37′46″ 118°10′26″ north latitude and has a population of 771.8 million. It is a warm temperate semi-humid continental monsoon climate with 2600–2900 h of sunshine throughout the year and a high degree of urbanization. Because the construction of a drone building dataset requires samples with diverse architectural forms and large illumination changes, the Caofeidian area was selected as the main data collection area. The drone building data were acquired using a DJI Mavic 3 Pro drone equipped with a Seer mapping tilt camera, PSDK 102S V3. The camera used a 23.1 × 15.4 mm sensor, 3.76 μm pixel size, and 35 mm tilt lens to capture images with a resolution of 6144 × 4096 pixels. The UAV followed a precise tic-tac-toe flight path to ensure comprehensive coverage of the area. The images were mainly in RGB format and preprocessed to correct internal distortion. Data augmentation techniques were applied to increase the dataset from 700 to 1400 images. Due to computational constraints, the images were resized to 1536 × 1024 pixels for model training and evaluation.

This paper also uses the public UAV dataset UDD6 to verify the effectiveness of the network. The UDD6 dataset consists of images of cities captured by drones at low altitudes. These images are characterized by varying dimensions, including 4096 × 2160 pixels, 3840 × 2160 pixels, and 4000 × 3000 pixels, and the distribution of image counts per size is 32, 16, and 93, respectively. The images in the UDD6 dataset are in the RGB format, and each image is subdivided into five distinct categories. The dataset, consisting of 1050 images, was processed and segmented into 1024 × 720 pixels for training and validation. Buildings were assigned a 1 label, while other categories received a 0.

2.2. Methodology

2.2.1. Architecture Overview

The network simulates building motion dynamics to generate synthetic labels for segmentation. It uses a dual-channel approach for semi-supervised image semantic segmentation, including a generator, discriminator, feature fusion module, and composite loss function, as depicted in Figure 1.

In each training batch, dual key channels are generated: a morphology-driven optical flow prediction channel for estimating motion information in captured images and generating pseudo-labeled images for semi-supervised learning, and an enhanced multilayer sensing channel for extracting semantic information and building features.

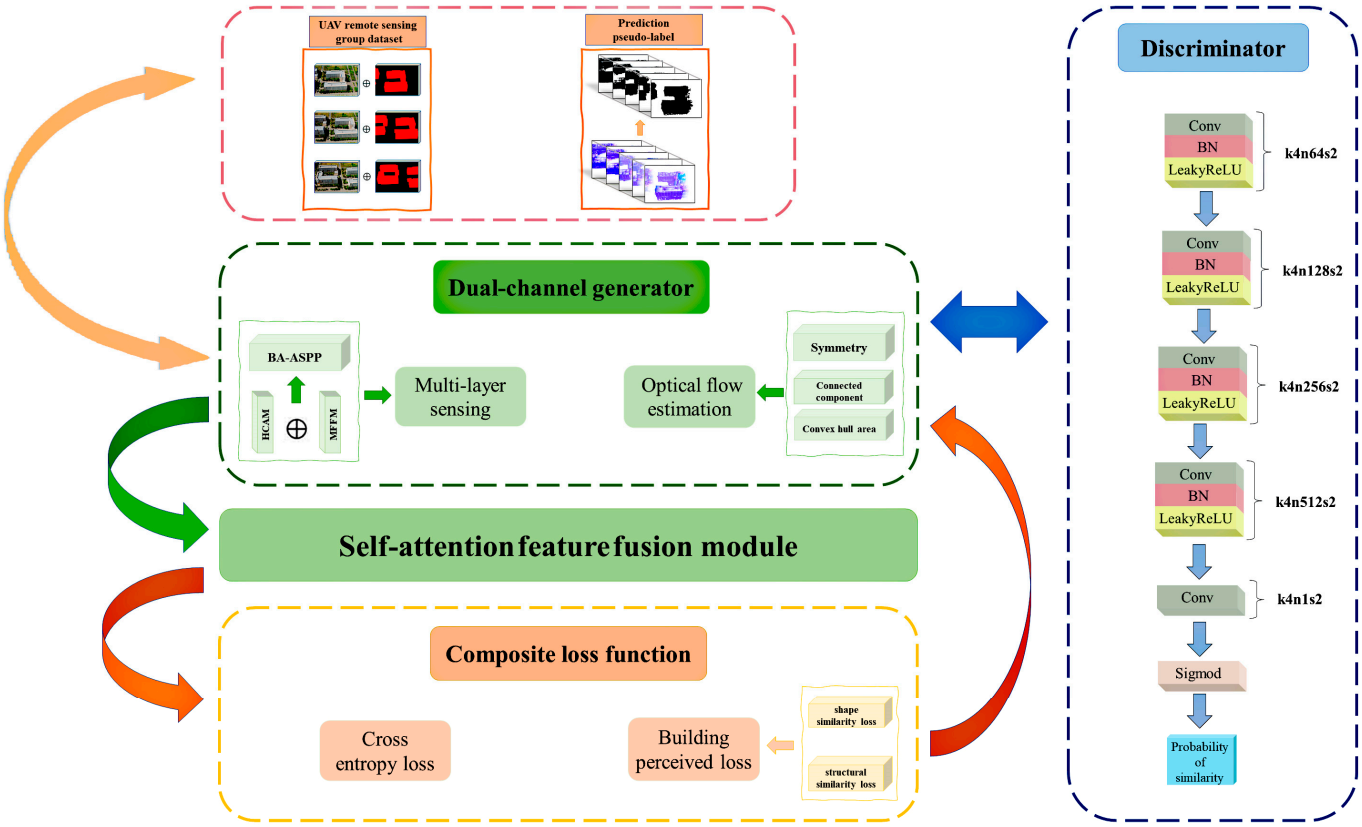

Figure 1. Total flow chart of the algorithm. The dotted lines represent the different modules in the network, and the arrows represent the order in which the network operates. The images and corresponding true labels are processed by the optical flow estimation channel and the improved Deeplabv3+ module. The features from both channels are fused, evaluated by the discriminator, and fine-tuned using the composite loss function for network convergence.

The discriminator evaluates the quality of a fused feature representation to determine its similarity to real data [27,28]. It is built on a fully convolutional neural network with a 4×4 convolutional kernel, data batch normalization method, and nonlinear activation function. The sigmoid function outputs the probability value. The network's accuracy is improved by optimizing the composite loss function. The error signal is passed back through multilayer sensing channels, and adjustments are made to produce a more accurate building estimate [29]. The process is repeated in multiple training batches until a fit state is reached.

2.2.2. Semi-Supervised Optical Flow Estimation Channel in Dual-Channel Generator

The generator [5,30,31] receives a multifaceted input comprising various components: an initial frame image, an intermediary frame image, a concluding frame image, and their corresponding labeled image. In this study, due to the relative displacement of the UAV when photographing the buildings, the external shape of the buildings does not change with its movement; therefore, we classify this 'building movement' as rigid motion, leading us to adopt an optical flow model with a uniform smoothing strategy.

The process involves defining a building's spatial extent using a binary mask, determining its shape attributes using metrics like symmetry scores, fractional characteristic maps, and consistency calculations, and computing morphological attributes, forming a morphological vector [32–34].

This integration enables a more comprehensive understanding of the motion characteristics of buildings, as visually illustrated in Figure 2.

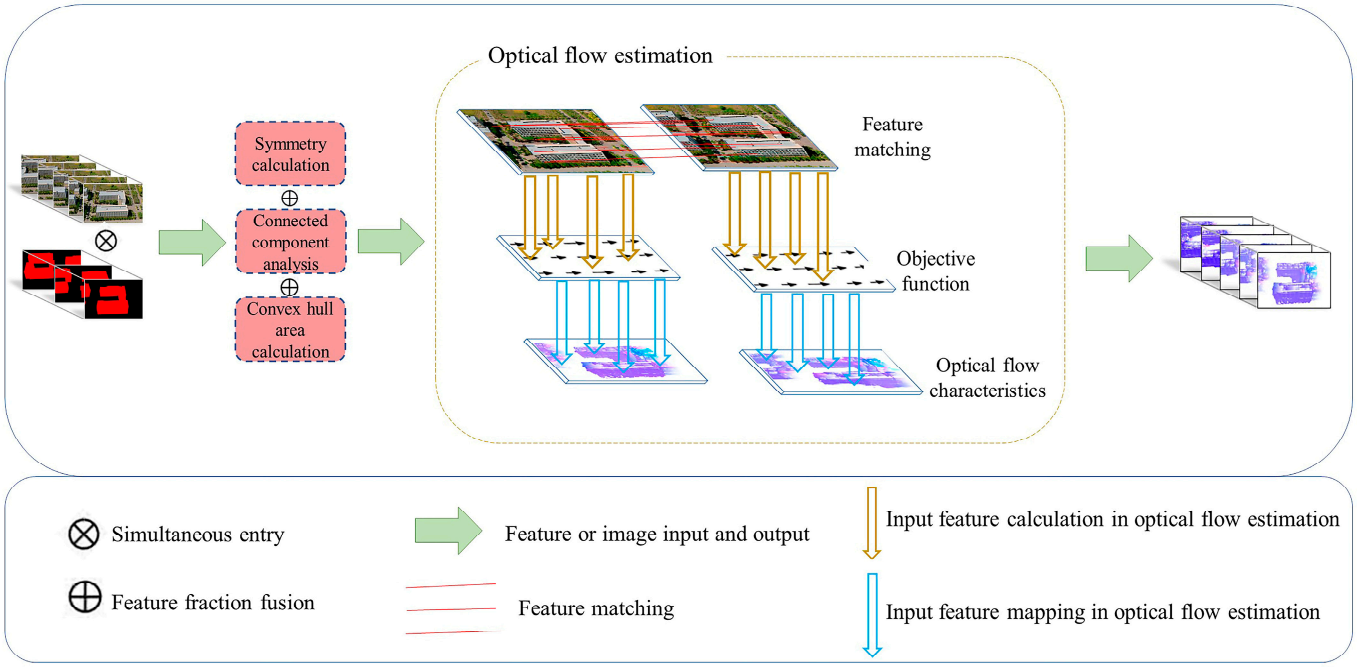

Figure 2. Schematic diagram of shape-driven optical flow estimation channel. The RGB images and keyframe label maps go through building form constraint algorithms before entering the optical flow estimation channel. This channel handles feature extraction and pseudo-label generation, including constrained feature matching, displacement calculation to establish the objective function, and the generation of optical flow characteristic values.

The overarching objective of optical flow estimation is to compute pixel displacements while striving to minimize the discrepancies inherent in the optical flow field. In the context of this study, we adopt the Horn–Schunck optical flow method, which is firmly rooted in the analysis of brightness gradients [35]. The core mathematical expression characterizing this method is as follows:

$$E(u,v) = \iint \left(\nabla I \times (u,v) + \frac{1}{2}\alpha \left(\left|\nabla u\right|^2 + \left|\nabla v\right|^2 \right) \right) dxdy \tag{1}$$

where (u, v) is the displacement field, I is the brightness of the image, α is the smoothness weight, and ∇ is the gradient operator. The first term of the integrand measures the dot product of the luminance gradient within the displacement field, which expresses the rate of change in luminance in the displacement direction. The second term is the smoothness term, which facilitates the smoothness of the displacement field to reduce noise.

The main goal is to optimize the energy function $E(u, v)$ for the optimal displacement field, minimizing errors and maintaining consistent image brightness. The energy function formulation incorporates morphological building information to constrain and mitigate extraneous noise beyond the building structure.

The calculation formula is as follows:

$$E(V) = E_{data}(V) + \lambda E_{smooth}(V) \tag{2}$$

where V is the motion energy field,$E_{data}\ (V)$ measures the difference in optical flow before the observation data, and $E_{smooth}(V)$ measures the smoothness of the motion field. To combine morphological information, the capability function is modified to

$$E(V) = E_{data}\ (V) + \lambda E_{smooth}\ (V) + \mu E_{morph}\ (V, F_{mid}) \tag{3}$$

In this formula, $E_{morph}\ (V, F_{mid}\ )$ is a morphological information term, which is used to represent the influence of morphological information on the motion field. The specific definition of this item can be designed according to the different morphological information of different buildings:

$$E_{\text{morph}}\ (V, F_{\text{mid}}) = s \times E_{\text{symmetry}}\ (V) + c \times E_{\text{connectivity}}\ (V) + a \times E_{\text{convexity}}\ (V) \quad (4)$$

Within this context, three distinct energy components come into play:$E_{\text{symmetry}}\ (V)$ gauges the impact of symmetry scores,$E_{\text{connectivity}}\ (V)$ delves into the distribution of connectivity across buildings, and $E_{\text{convexity}}\ (V)$ quantifies the congruence of the motion field in relation to the convex hull area.

The symmetry scores measure symmetry in a building's boundaries, which can become asymmetrical due to lighting or weather changes. They help segmentation algorithms assess if building sections exhibit symmetry or near-symmetry, improving their ability to discern the building's shape and structure.

The calculation process is outlined as follows:

$$d = \sqrt{(x - x_c)^2 + (y - y_c)^2} \quad (5)$$

For each pixel position (x, y) on the feature plot F, we calculate the Euclidean distance d of the geometric center position of the building (x_c, y_c), which is obtained by cutting the pixel from the labeled image and the RGB image. And we calculate the symmetry fraction S of the pixel, where the feature map size is M × N, as shown in Equation (6):

$$S = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} \frac{1}{1 + d(x, y)} \quad (6)$$

The connectivity feature maps are vital for segmentation algorithms, enhancing accuracy in complex structures and interactions. They capture relationships between building components and surroundings, preventing optical flow discontinuities and ensuring robust performance even in adverse conditions. The generalized calculation formula for connectivity feature maps is expressed as follows:

$$F(C) = g(f1(C), f2(C), f3(C)) \quad (7)$$

where $F(C)$ represents the connected domain feature map, C represents the set of connected regions, and C of each region contains a set of pixels. $f1(C)$, $f2(C)$, and $f3(C)$ are the connected regions' area characteristics, the connected regions' circumference characteristics, and the connected regions' eccentricity characteristics, respectively.

The consistency calculation for convex hull area measures the overlap between two buildings' convex hulls, assessing their consistency. Changes in light and shadows can blur or obscure building edges. This measure constrains building characteristics, improving boundary accuracy and aligning segmentation results with actual building shapes. The calculation formula is as follows:

$$\text{Consistency} = \frac{|Area(A_c h) - Area(B_c h)|}{(Area(A_c h) + Area(B_c h))} \quad (8)$$

In this formula, A_c represents the convex hull of building A, and B_c represents the convex hull of building B. $Area(A_c h)$ and $Area(B_c h)$ are the areas of the convex hulls of building A and building B, respectively.

Through the meticulous optimization of the composite energy function $E(V)$, the resultant motion vector field V after the incorporation of morphological insights can be attained. This integration culminates in the generation of a pseudo-labeled image for the intermediate frame.

Precisely, for every pixel encompassed within the intermediate frame image, its coordinates (x, y) and the associated motion vector (u, v) from the amalgamated motion vector field V are used to compute the position (x + u, y + v) within the source image. To determine the pixel value at this new position (x′, y′) in the initial frame, bilinear interpolation is applied. The formula governing this interpolation is articulated as follows:

$$\begin{array}{c} I_{\text{pseudo}}\left(x,y\right) = (1-\delta)(1-\gamma)I_{\text{start}}\left(x_{\text{start}}, y_{\text{start}}\right) + \\ \delta(1-\gamma)I_{\text{start}}\left(x_{\text{start}}+1, y_{\text{start}}\right) + \\ (1-\delta)\gamma I_{\text{start}}\left(x_{\text{start}}, y_{\text{start}}+1\right) + \\ \delta\gamma I_{\text{start}}\left(x_{\text{start}}+1, y_{\text{start}}+1\right), \end{array} \tag{9}$$

Here, $\delta\gamma$ is the offset of $(x_{\text{start}}, y_{\text{start}})$ with respect to the integer coordinates.

Following a series of iterative cycles, each pixel within the intermediate frame image undergoes interpolation, culminating in the acquisition of a comprehensive pseudo-labeled image, I_{pseudo} . This synthesized image effectively encapsulates the building's movement progression, wherein every pixel encompasses motion details spanning from the initial frame to the middle frame. In the final stages, the utilization of dilated convolution facilitates the transformation of I_{pseudo} into a coherent pseudo-label, thus establishing a logical connection across the building's components.

2.2.3. Improved Deeplabv3+ Module in Dual-Channel Generator

Indeed, the shape-driven optical flow channel is designed to capture building morphology and motion details, generating essential pseudo-labels for semi-supervised studies. However, it has limitations in understanding texture intricacies, fine details, and contextual cues, and is sensitive to illumination fluctuations and occlusion [36–38].

Figure 3 shows the architectural design of the Deeplabv3+ model, a building segmentation approach, with ResNet50 as the primary backbone, based on prior research for feature extraction [39–41].

The BA-ASPP module is the core of our model, enhancing feature extraction by capturing contextual information from different receptive field sizes. It uses parallel dilated convolutions and introduces a cascading structure, inspired by cascading networks, to incorporate critical details across scales [12]. This structure combines image features and morphological characteristics of buildings, initiating average and maximum pooling operations along the channel dimension of the feature map.

1. Building-Aware Atrous Spatial Pyramid Pooling network construction

The ASPP feature enhancement network balances void rate and multiscale information extraction while maintaining a large receptive field. Pooling enhances the ASPP module's ability to sense remote contextual information. The hierarchy channel attention module (HCAM) is integrated to extract multiscale information and features from various receptive fields.

The structural design, as shown in Figure 4, aims to heighten the model's sensitivity to multiscale information while maintaining a balanced utilization of atrous convolutions. The input feature layer in the HCAM network is first extracted through the average pooling and maximum pooling layers, which reduces the calculation amount and retains the significant features of the building to the greatest extent. The calculation formula is as follows:

$$\begin{array}{l} F_{\text{avg}}^{\text{c}} = Avg(F_{\text{c}}) \\ F_{max}^{v} = Avg(F_{\text{c}}) \end{array} \tag{10}$$

In this formula, F_{c} represents the feature map after applying the feature extraction network, $F_{\text{avg}}^{\text{c}}$ represents the feature map after average pooling, and F_{max}^{v} represents the feature map after maximum pooling.

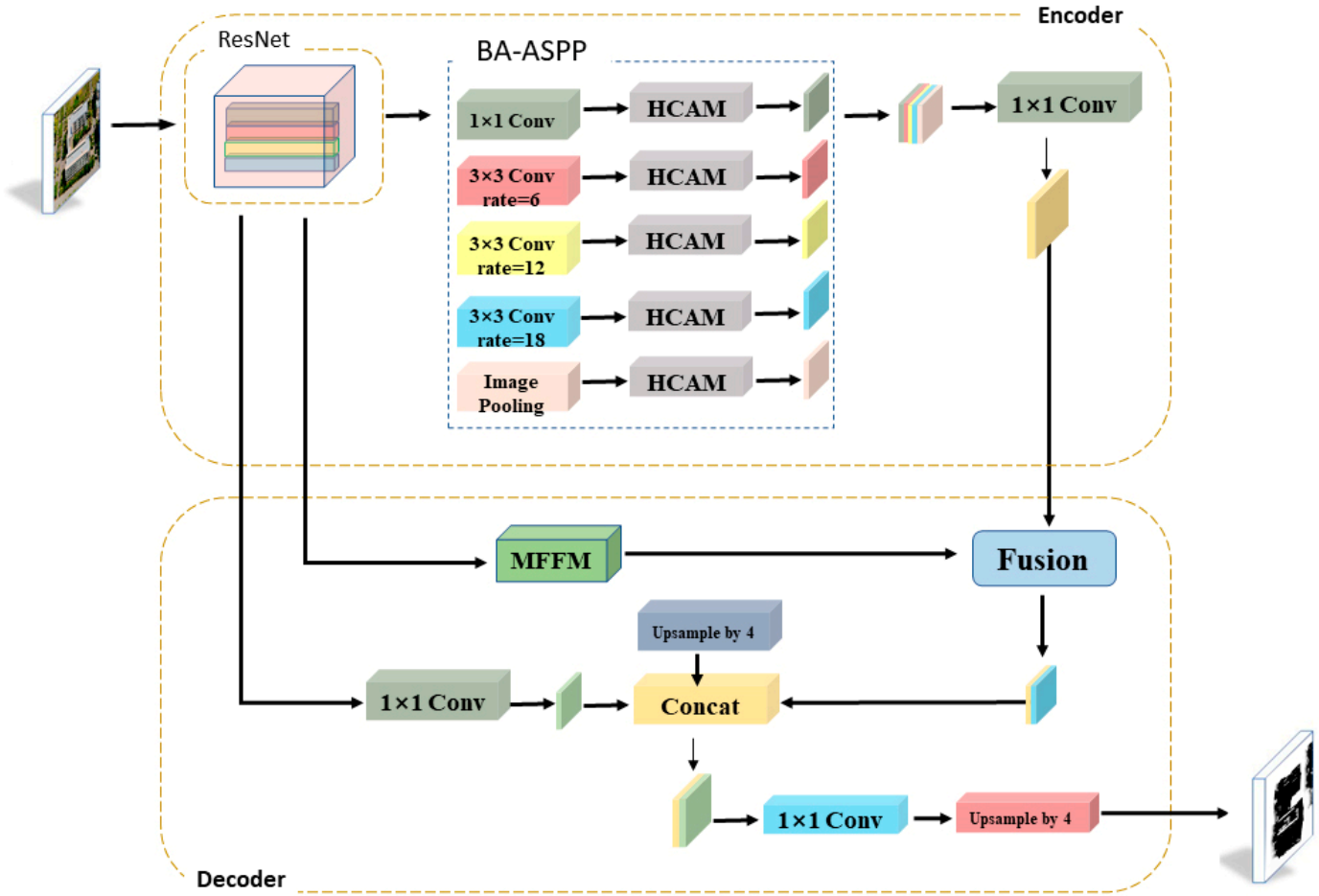

Figure 3. Improved Deeplabv3+ module structure chart. The images are initially passed through BA-ASPP, MFFM, and 1 × 1 conv using the features extracted by ResNet. The features input to BA-ASPP undergo processing by different conv layers and HCAM. They are then fused with features constrained by MFFM and subjected to 1 × 1 conv for sampling fusion. Lastly, the output feature map is upsampled.

In this enhanced structure, the process begins with obtaining first-order features. These features are subsequently fused and concatenated with the morphological characteristics of buildings, which encompass symmetry fraction [42], connectivity distribution of buildings [43], and consistency calculation of the convex hull area [44].

After the feature fusion process, the combined features undergo a series of additional operations to further refine and enhance their representation. To begin, we apply a 1 × 1 convolution operation to the original image. This operation helps generate second-order features by capturing the relationships and dependencies within the image. These second-order features are then intricately integrated with the previously fused features, resulting in the formation of third-level features. This integration enables the model to capture complex hierarchical information, enhance feature representations, and prepare them for subsequent processing. In the following stages, these third-level features are subjected to a 5×5 convolution layer for advanced convolutional operations to enhance fine-grained perception. This step allows the model to extract fine-grained details, intricate patterns, and higher-level contextual information from the features. Ultimately, this multistage refinement process contributes to the generation of the final feature map through the sigmoid activation function.

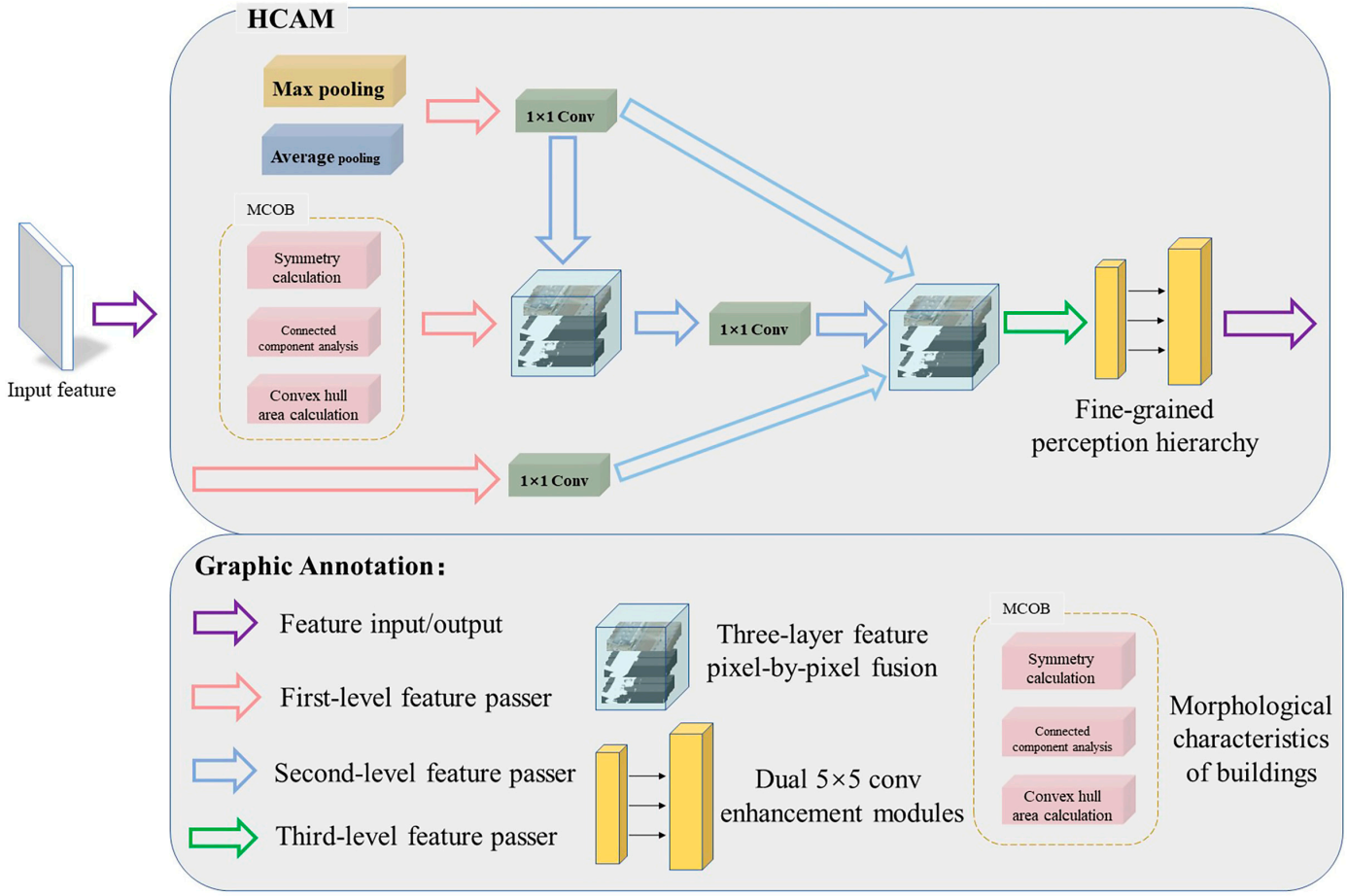

Figure 4. Hierarchy channel attention module (HCAM) structure flow chart. Input features are extracted again via max pooling, average pooling, MCOB, and 1 × 1 conv. The max pooling and average pooling features are merged with MCOB-constrained features using 1 × 1 conv. The resulting features are further combined with those obtained from initial convolution. Finally, these integrated features are passed to the fine-grained perception hierarchy for amplification.

2. Multilevel Feature Fusion Module (MFFM)

Figure 5 provides a schematic of the structure of our multilevel feature fusion module (MFFM).

The MFFM segmentation task uses three inputs: F1, F2, and F3. F1 is high-resolution and undergoes operations like convolutional operation, batch normalization, and coordinate attention. F2 captures intermediate-level features and undergoes upsampling. F3 provides a broader view and undergoes dilated convolutions to capture more contextual information. After processing, their features are fused together, combining high-level semantic information from F1 and contextual information from F2 and F3. The resulting feature representation is passed through the rectified linear unit activation function for complex feature interactions.

2.2.4. Feature Fusion Module

This study proposes an approach using an attentional mechanism to integrate adaptive weights between morphology-driven optical flow estimation channels and an improved Deeplabv3+ module. This allows the network to autonomously discern the significance of each channel at different spatial locations, resulting in more precise weight allocation during the fusion phase. This technique improves context-aware fusion processes and reduces feature redundancy problems in dual-channel networks.

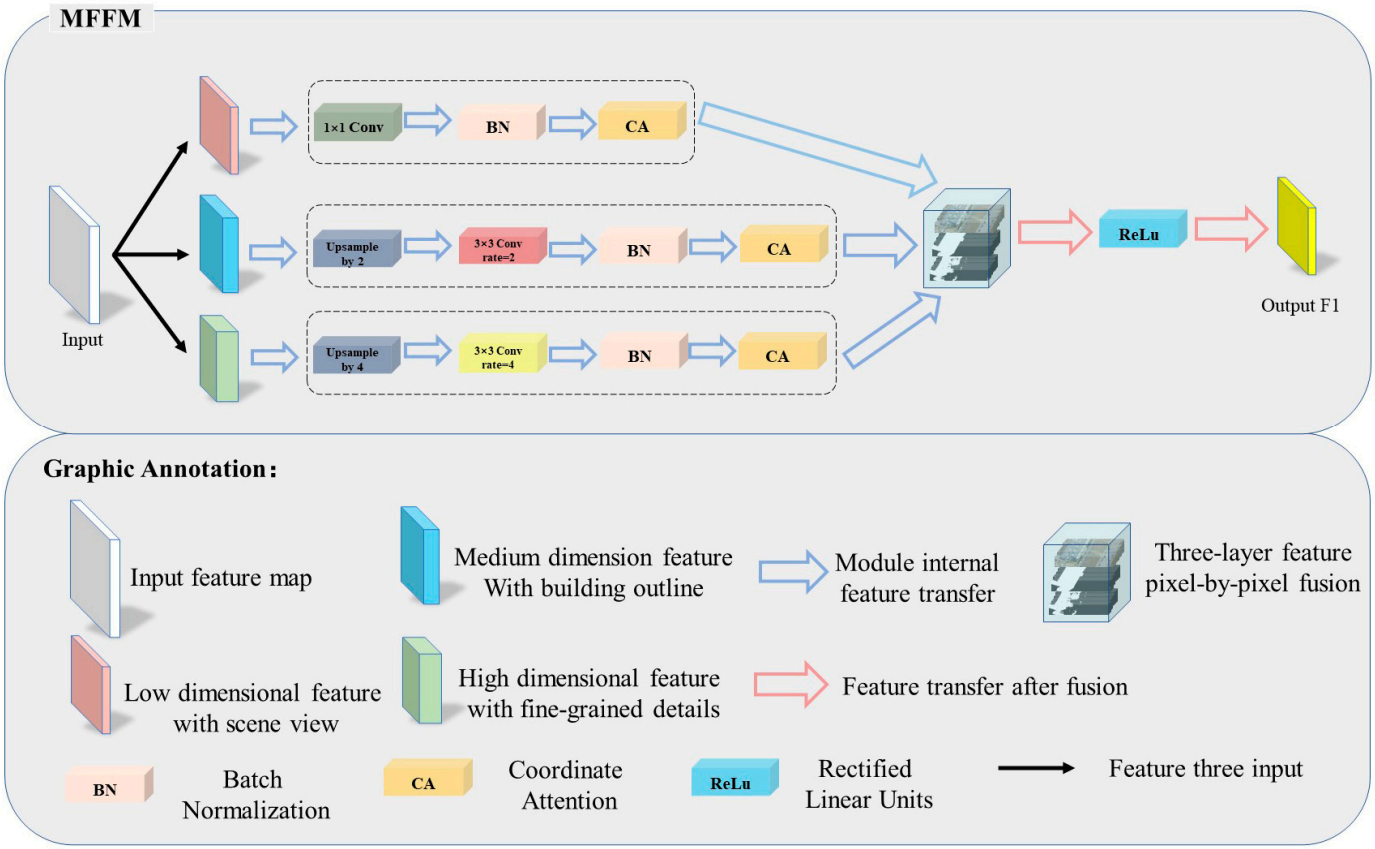

Figure 5. Multilevel feature fusion module (MFFM) structure flow chart. Input features are sampled in three dimensions: low-dimensional features through 1 × 1 conv, batch normalization, and coordinate attention; medium-dimensional features via upsampling, 3 × 3 conv (expansion rate of 2), batch normalization, and coordinate attention; high-dimensional features with 3 × 3 conv (expansion rate of 4), batch normalization, and coordinate attention. The final output is obtained by applying the ReLu activation to these features after multilayer feature fusion.

The self-attention mechanism is the linchpin for determining these adaptive weights. The process begins with a 3 × 3 convolution, employed to map the features from each channel into a shared, low-dimensional space. Subsequently, the self-attention weight is computed by assessing the similarity between feature points. This calculation unfolds as follows:

$$\begin{aligned} S_{\text{flow}} &= U_{\text{flow}} \times (U_{\text{flow}})^T \\ S_{\text{deeplab}} &= U_{\text{deeplab}} \times \left(U_{\text{deeplab}}\right)^T \end{aligned} \tag{11}$$

Within this context, S_{flow} and S_{deeplab} denote the feature similarities inherent in the optical flow estimation channel and the improved Deeplabv3+ module, respectively. In parallel, U_{flow} and U_{deeplab} signify the features subsequent to mapping for both the aforementioned channels.

To establish a standardized distribution of attention weights, the incorporation of normalized convolution becomes imperative. This step yields the attention weight matrices A_{flow} and A_{deeplab} by means of normalization procedures.

Ultimately, the calculated attention weights are harnessed to assign weights to the features originating from both channels. This process culminates in the formulation of the fused feature, denoted as F_{fused}, as succinctly demonstrated below:

$$F_{\text{fused}} = A_{\text{flow}} \times U_{\text{flow}} + A_{\text{deeplab}} \times U_{\text{deeplab}} \tag{12}$$

2.2.5. Loss Function

To elevate the efficacy of the semi-supervised building segmentation network while upholding image-specific detail features [45,46], a dual-pronged strategy incorporating both cross-entropy loss and building perception loss, rooted in the tenets of measurable generative adversarial network performance, is introduced. This innovative amalgamation of loss functions amalgamates diverse types of losses, resulting in the construction of a composite loss function.

1. Cross-Entropy Loss Function

Cross-entropy is an important concept in information theory, and it is also a commonly used loss function in neural networks [47]. Its calculation formula is

$$CE = \sum_{k} p(k) \times \log\left[\frac{1}{q(k)}\right], \tag{13}$$

where k is the sample of class k; p is the true category distribution; and q is the predicted marker distribution.

Given the prevalent imbalance between false and true labels within this study, adopting the conventional loss function, as represented in Equation (13), can inadvertently incline the model towards categorizing samples as the larger class [48].

To mitigate this issue, an innovative remedy emerges in the form of weighted improved cross-entropy. By apportioning distinct weights to each category, this mechanism effectively addresses the label imbalance predicament.

$$L_{\text{weighted\_cross\_entropy}} = -\frac{1}{N}\sum_{i=1}^{N} w_i \times y_i \times \log(p_i) \tag{14}$$

where N is the number of samples, w_i is the weight of the i class, y_i is the true label of the i class, and p_i is the prediction probability of the model. Furthermore, w_i can be set to the reciprocal of class occurrence frequency to balance different classes.

2. Building Perceived Loss Function

The cross-entropy loss function is a method used to distinguish between buildings and non-buildings by comparing predicted outcomes with actual pixel labels [49]. However, it can lead to overfitting and may not capture intricate building shape and detail information, resulting in less precise segmentations [50,51].

Building perception loss is a strategy that enhances geometric shape and structural intelligence in model training. It harmonizes motion attributes and shape components, allowing models to better understand the geometric nuances of buildings through optical flow techniques. In a more granular context, building perception loss comprises two distinct components: shape similarity loss and structural similarity loss.

The quantification of shape similarity loss L_{shape} entails a process of gauging the dissimilarity in shape characteristics between features:

$$L_{\text{shape}} = \sum_{i=1}^{N}\sum_{j=1}^{M} \| F_{\text{flow}}(i,j) - F_{\text{deeplab}}(i,j) \|_2^2, \tag{15}$$

Here, F_{flow} is the feature of optical flow channel fusion and F_{deeplab} is the feature of the improved Deeplabv3+ module.

The structural similarity loss $L_{\text{structure}}$ can be calculated using the structural similarity index (SSIM) [51,52].

$$L_{\text{structure}} = \sum_{i=1}^{N}\sum_{j=1}^{M}(1 - \text{SSIM}(F_{\text{flow}}(i,j), GT(i,j))), \tag{16}$$

Here, GT denotes the image representing the ground truth label in the context of this process.

$L_{\text{structure}}$ and L_{shape} exist as a condition of the loss function for morphological constraints. The ultimate building perception loss L_{building} is an amalgamation achieved through weighted combination of both the shape similarity loss and the structural similarity loss.

$$L_{\text{building}} = \alpha L_{\text{shape}} + \beta L_{\text{structure}} \tag{17}$$

3. Composite Perceptual Loss Function

The integration of building perception loss in semi-supervised building segmentation tasks optimizes the model by utilizing diverse information. The cross-entropy loss function and building perception loss contribute distinct strengths, compensating for each other's limitations, resulting in superior segmentation outcomes.

In summation, the composite loss function is defined as follows:

$$L_{\text{total}} = \alpha L_{\text{shape}} + \beta L_{\text{structure}} + \lambda L_{\text{weighted\_cross\_entropy}} \tag{18}$$

Here, α, β, and λ represent the weights of each loss, which can be set up with actual drone building samples.

2.2.6. Benchmark Methods

To evaluate the validity of our proposed approach, we conducted a comprehensive comparison with four well-known benchmark methods for semi-supervised language segmentation. These methods include the following:

1. AffinityNet [16] leverages class activation mapping (CAM) to accentuate localized discriminative areas of the target, thereby enhancing segmentation.
2. AdvSemiSeg [17] is grounded in adversarial training principles and leverages a generator–discriminator tandem to fuse semi-supervised signals, ultimately enhancing segmentation performance.
3. SemiCycleGan [18] is based on cyclic generative adversarial networks and employs cyclic consistency and adversarial loss in its generator to achieve improved segmentation results.
4. CCVC [19] uses a two-branch co-training framework to encourage learning distinct features from irrelevant viewpoints. The CVC strategy promotes consistent prediction scores for input.

2.3. Implementation Setting and Evaluation Indicators

2.3.1. Evaluation Metrics

In order to quantitatively analyze the comparison between our method and other methods, we used precision, mean pixel intersection over union ($mIoU$), and the $F1$ score (F1) as evaluation indicators. The formulae for calculating these metrics are shown in Equations (19)–(21).

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN} \tag{19}$$

The F1 score is defined as

$$F1 = \left(\frac{2+\frac{FP}{TP}+\frac{FN}{TP}}{2}\right)^{-1} \tag{20}$$

Precision is defined as

$$\text{Precision} = \frac{TP}{TP+FP} \tag{21}$$

where TP, FP, FN, and k represent true positive, false positive, true negative, false negative, and the number of categories, respectively.

2.3.2. Preparation for the Experiments

The experiments involved using the PyTorch deep learning framework on a Windows operating system, with an 11th Gen Intel Core i7-11800H CPU and a GTX3060 GPU. The training parameters were standardized and aligned. The optimization algorithm was Adam, and the batch size was 4. The weights of the composite loss function were chosen based on the image data's specific requirements, with α and β set to 0.25 and λ to 0.5, respectively, to accommodate standard tasks due to the diversity of building shapes.

The weight values for building perception loss (λ) are influenced by the size of building samples and their complexity. For extensive datasets with intricate shapes, a higher weight may be assigned to λ to prioritize fine-grained details and structural accuracy. These weight adjustments aim to optimize the model's performance for the specific segmentation task. In this study, unlabeled images accounted for approximately 70% of the overall dataset.

3. Results

3.1. Qualitative Analysis of Comparative Experimental Results

The segmentation capabilities of diverse semi-supervised segmentation methods for segmenting buildings were examined in a comprehensive manner. The segmentation performance was analyzed individually for the two datasets, and a visual juxtaposition of the segmentation images obtained using the network described in this paper is presented, aiming to underscore its effectiveness.

Figure 6 shows the visualization results of network segmentation of the drone building dataset. The AffinityNet, AdvSemiSeg, SemiCycleGan, and CCVC networks extract most buildings, but there are still problems of extraction errors and omissions. An analysis of Figure 6c, focusing on the second and third rows, reveals that AffinityNet primarily relies on localized context data for segmentation. However, this emphasis on local context may fall short of capturing the broader global shape and motion attributes of buildings. Consequently, inaccuracies in boundary delineation and shape representation might manifest in complex scenes. Despite AffinityNet's utilization of feature affinity graphs for fusion, its fusion mechanism may not optimally exploit multichannel information. Moreover, this method exhibits sensitivity to input image noise due to its reliance on local feature interdependence. This sensitivity can lead to unstable segmentation results, particularly in lower-quality images. The inherent variability in UAV-captured buildings, which differs substantially from that in structures captured in remote sensing images, poses an added challenge. The absence of explicit integration of building morphology in AffinityNet makes it ill equipped to handle building boundaries and shapes effectively.

Figures 6d and 6e respectively show the experimental results of AdvSemiSeg and SemiCycleGan on the drone building dataset. The extraction of buildings is a formidable task due to the abundant presence of shadows and irregular lighting conditions, particularly within the buildings outlined in the first and fourth rows, marked by the green boxes. While both AdvSemiSeg and SemiCycleGan manage to enhance the shape information related to the buildings in the region, the extraction results still exhibit some noise, leading to a degree of blurriness in the boundary details. AdvSemiSeg and SemiCycleGan show comparable performance in the segmentation of the second- and third-row images because the lighting and shadow conditions vary little. This issue can be attributed to the primary focus of the generator components in both networks on image generation. When a drone captures a relatively constrained area featuring large buildings, it can be challenging to capture the full extent of these structures within a single image, thus making it difficult to maintain precise building shape information. The absence of mechanisms guiding these networks to prioritize building characteristics contributes to the reduction in the accuracy of segmenting built areas.

Figure 6f presents the visualization results of the advanced semi-supervised CCVC network. Notably, certain building details exhibit similarities to those shown in Figure 6g. However, its performance appears less robust in areas where noticeable shading changes occur due to building shadows and lighting variations; this can be clearly seen in the green

box in the third row. This discrepancy can be attributed to the two-branch cooperative training framework in CCVC, which encourages both subnetworks to acquire multilayer information features. While the conflict-based pseudo-labeling (CPL) method effectively reduces network crashes caused by a limited number of labels, it demonstrates reduced sensitivity to the light and dark variations often encountered in building segmentation tasks, without corresponding supplementary measures.

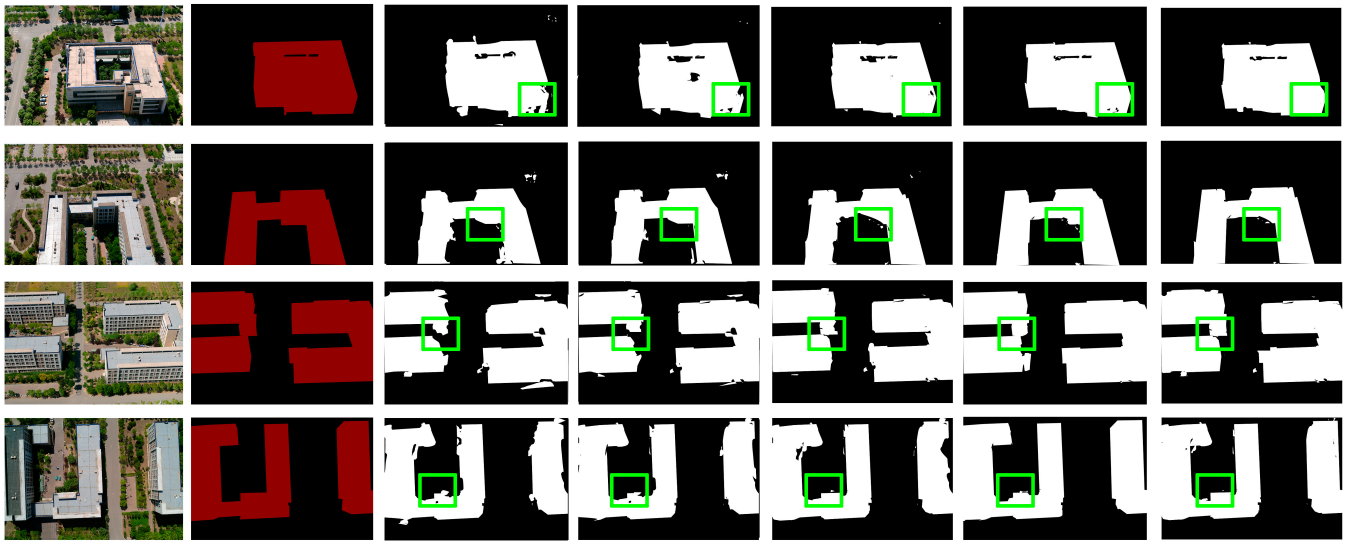

Figure 6. Comparison of the results of each network in the drone building dataset.

From the green boxes shown on lines 2 and 4, it can be seen that for large buildings, the extraction results of the comparison methods have relatively serious holes and building edge errors. The HCAM module proposed in this paper extracts multiscale context information about buildings via multiple cascaded and parallel technologies, while the MCOB can effectively suppress features other than buildings, which helps the network to better pay attention to the boundaries, shapes, and details of buildings.

These observations highlight the advantages of the networks outlined in this paper, particularly in the ability to capture global shape and motion properties, which yield more precise and stable segmentation results in complex scenes.

In Figure 7, the segmentation outcomes of each network using the UDD6 dataset are showcased. This dataset exhibits a wide array of building shapes, ranging from irregular forms to structures with sharp boundaries.

The AdvSemiSeg segmentation approach relies on the fusion of real and pseudo-labels through adversarial training, as mediated by adversarial networks. This phenomenon can be observed in the green box of the first row. However, in scenarios with intricate backgrounds, adversarial training may face interruptions, leading to the generation of inaccurate images. This is evident when comparing the results in the first row, where the complexity of building shapes and backgrounds in the UDD6 dataset posed a challenge for AdvSemiSeg. AffinityNet primarily centers on leveraging connectivity information within an image. Regrettably, this focus prevents it from adeptly capturing disjointed regions and subtle features that manifest in complex building structures. This limitation may hinder the accurate division of this complex building, which can be clearly seen in the green boxes in the first and second rows. SemiCycleGan operates on the principle of image translation via cyclic consistency loss. However, this mechanism often leads to notable discrepancies between generated images and real building representations. From the green boxes in the third and fourth rows, it is evident that complex backgrounds, along with imbalances and variations in shadows, can hinder the convergence of network consistency loss, thereby affecting the segmentation quality.

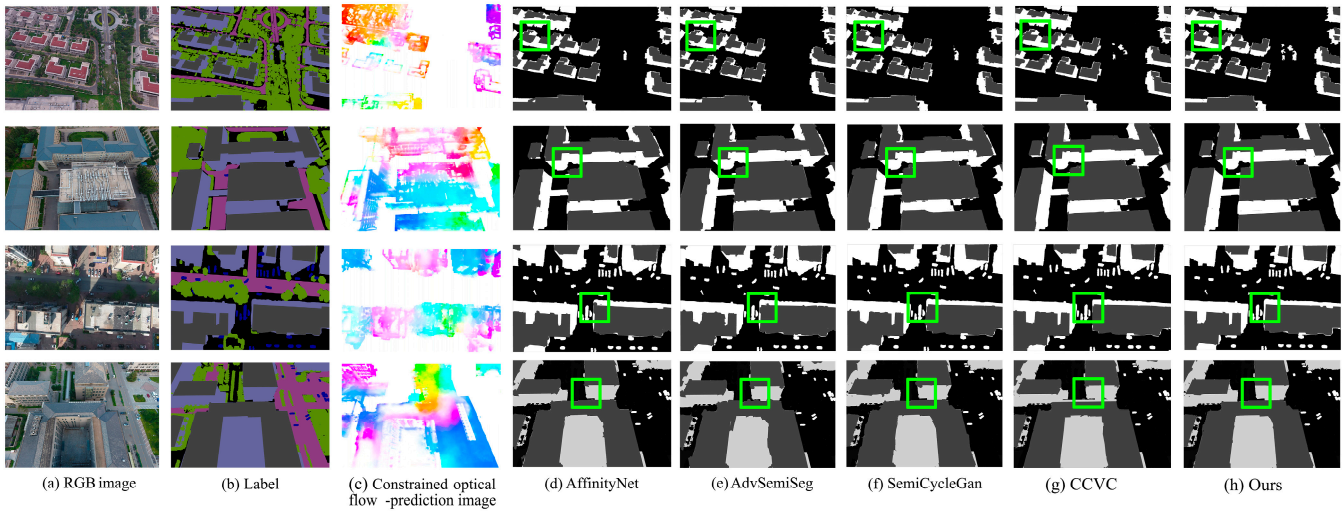

Figure 7. Comparison of the results of each network in the UDD6 dataset.

CCVC demonstrates both strengths and weaknesses in its performance. Its two-branch cooperative training framework facilitates the extraction of multilayer information features, which can be beneficial in scenarios requiring comprehensive feature learning. However, as can be seen from the images in the second and fourth rows, CCVC has limitations in dealing with complex lighting changes, especially in areas where building shadows and lighting change.

As depicted in Figure 7, the proposed network excels in extracting building information, owing to the MFFM and building perceived loss. An examination of the green boxes in the first and fourth rows reveals that the network introduced in this paper adeptly captures complete building outlines, regardless of whether the buildings are large or small. From the green box in the third row, it is evident that despite significant variations in light and shadow, the network can effectively compensate by utilizing other building features, resulting in an improved segmentation outcome. Its comprehensive approach to capturing global shape, motion properties, and intricate morphologies enhances its ability to yield accurate segmentation results.

3.2. Quantitative Analysis of Comparative Experimental Results

Table 1 presents the results of a comprehensive assessment of the performance of each network across the two datasets.

Table 1. Comparison of segmentation results of each network.

Method	Drone Building Dataset			UDD6 Dataset		
	F1 Score (%)	mIoU (%)	Precision (%)	F1 Score (%)	mIoU (%)	Precision (%)
AffinityNet	70.31	74.56	73.34	69.71	70.18	71.64
AdvSemiSeg	77.82	76.93	78.47	75.63	77.65	76.82
SemiCycleGan	69.75	73.59	72.63	68.56	71.89	70.02
CCVC	**80.18**	80.26	79.82	77.25	78.59	77.49
Ours	79.36	**82.69**	**80.56**	**77.68**	**79.37**	**79.43**

Note: Black bold represents the highest level of the same evaluation criteria.

In the drone building dataset, the network proposed in this paper achieved an F1 score of 79.36%, an mIoU of 82.69%, and a precision of 80.56%, reaching an excellent level in the comprehensive index. Lines 1–3 show the evaluation standards of AffinityNet, AdvSemiSeg, and SemiCycleGan. Compared with other semi-supervised networks, the F1 score of the network proposed in this paper increased by 9.05%, 1.54%, and 9.61%, respectively. The mIoU improved by 8.13%, 5.76%, and 9.10%, while precision improved by

7.22%, 2.09%, and 7.74%, respectively. The quantitative analysis of CCVC demonstrated an F1 score of 80.18%, an mIoU of 80.26%, and a precision of 79.82%. These results reveal that despite CCVC enhancing the feature perception of individual subnetworks through feature transfer during the inference stage, its ability to process multilevel image details is not as precise as that of our proposed cascade structure, the HCAM. The morphology-driven dual-channel network proposed in this paper outperformed all other networks across every evaluation metric in the drone building dataset. Its mIoU and precision surpass those of CCVC by 2.43% and 0.74%, respectively.

When examining the UDD6 dataset, which presents greater complexities in terms of building types, shapes, and backgrounds, all networks experienced a general performance decline. Nevertheless, AdvSemiSeg maintained a relatively stable level of accuracy, delivering an F1 score of 75.63%, an mIoU of 77.65%, and an accuracy of 76.82%, all of which still reach commendable levels. Conversely, both SemiCycleGan and AffinityNet encountered limitations when dealing with the dataset's intricacies, resulting in compromised model performance and an inability to effectively capture intricate building shapes. Specifically, the results of the network proposed in this paper were compared with the results of the networks from the first row to the third row. The F1 score showed improvements of 7.97%, 2.05%, and 9.12%, while the mIoU increased by 9.19%, 1.72%, and 7.48%, and accuracy improved by 7.79%, 2.61%, and 9.41%, respectively.

Despite the complexity of the UDD6 dataset, the performance indicators of the proposed network are still the best. Compared with the advanced CCVC, the F1 score, mIoU, and accuracy of the network were improved to varying degrees. These findings highlight the superiority of the proposed network in accurately capturing complex architectural details. This enhancement improves the visual salience of buildings in the captured images. Importantly, our approach also preserves the necessary details.

3.3. Ablation Experiment Using the Drone Building Dataset and UDD6 Dataset

To assess the impact of different modules within the network, a series of ablation experiments were conducted. This involved dissecting the network and evaluating the roles played by the morphology-driven channel, the dual-channel generator, and the composite loss function. The outcomes of these experiments are summarized in Table 2, which provides the F1 score, mIoU, and precision values obtained from the ablation experiment conducted using the drone building dataset.

Table 2. Ablation experiment on drone building dataset.

Improved Deeplabv3+	Morphology-Driven Channel	Composite Loss Function	F1 Score (%)	mIoU (%)	Precision (%)
√	×	×	75.94	74.51	73.68
√	√	×	78.69	80.36	78.47
√	×	√	73.19	75.47	74.05
√	√	√	79.36	82.69	80.56

Note: √ means that the module is used for training, and × means that the module is not involved in computation.

The findings indicate that the employment of a dual-channel combined module results in substantial improvements. When we combine the improved Deeplabv3+ with a morphology-driven channel, HCAM plays a crucial role in capturing basic architectural features. These features include symmetry calculations, building connection analysis, and consistency measurement of convex shell areas. Therefore, the F1 score, mIoU, and precision of the combination reached the suboptimal standards of the ablation experiment, which were 78.69%, 80.36%, and 78.47%, respectively. However, when solely utilizing the improved single-channel Deeplabv3+ architecture and the composite loss function, the improvement is not as substantial, with an F1 score of 73.19%, mIoU of 75.47%, and precision of 74.05%. The presence of this phenomenon can be attributed to the continued effectiveness of the HCAM in the improved Deeplabv3+ architecture.

Importantly, with the introduction of the dual-channel strategy, the network achieves its optimal performance across all metrics. This approach yields remarkable enhancements of 3.42% in the F1 score, 8.18% in the mIoU, and 6.88% in precision when compared with using only the improved single-channel Deeplabv3+. These findings robustly affirm the indispensability of the proposed network structure for drone building segmentation tasks. The successful integration of these modules validates their synergistic effect, which collectively contributes to the network's enhanced performance.

To establish the broad applicability and persuasive potential of the new network modules, further investigation of the network was undertaken to scrutinize the individual roles of each module using the UDD6 public dataset. The findings from this ablation experiment are illustrated in Table 3.

Table 3. Ablation experiment on UDD6 dataset.

Improved Deeplabv3+	Morphology-Driven Channel	Composite Loss Function	F1 Score (%)	mIoU (%)	Precision (%)
√	×	×	69.34	71.73	70.48
√	√	×	75.59	75.44	73.87
√	×	√	74.39	76.17	75.39
√	√	√	77.68	79.37	79.43

Note: √ means that the module is used for training, and × means that the module is not involved in computation.

It is important to note that using only the improved Deeplabv3+ module yielded the lowest results in the experiment, with the F1 score, mIoU, and precision reaching 69.34%, 71.73%, and 70.48%, respectively. However, the network's accuracy significantly improved when combined with the morphology-driven channel, reaching suboptimal standards of 75.59%, 75.44%, and 73.87%, respectively. This improvement can be attributed to the fact that building information constraints in the form drive can be better adapted to the variety of building shapes present in the UDD6 dataset, including irregular and well-defined boundaries. The highest evaluation level was achieved by the network that combined the dual channel and the composite loss function, with the F1 score, mIoU, and precision reaching 77.68%, 79.37%, and 79.43%, respectively. These values were 2.09%, 3.93%, and 5.56% higher than those of the second-best combination.

Figure 8 showcases the visual outcomes of the ablation experiment conducted using the UDD6 dataset, with conclusions that are consistent with those drawn based on the results presented in Table 3. As depicted in Figure 8d in the second and third rows, the segmentation task accompanied by the integration of the morphology-driven channel excels in capturing edges and finer intricacies. In contrast, Figure 8e in the first row reveals that the inclusion of the composite loss function yields a solid portrayal of simpler buildings; however, when confronted with multiscale structures, its effectiveness falters. In the case of building shadows in row 4, it can be seen that both the improved Deeplabv3+ and the introduction of the form-driven optical flow estimation channel have a good inhibition effect on building shadows.

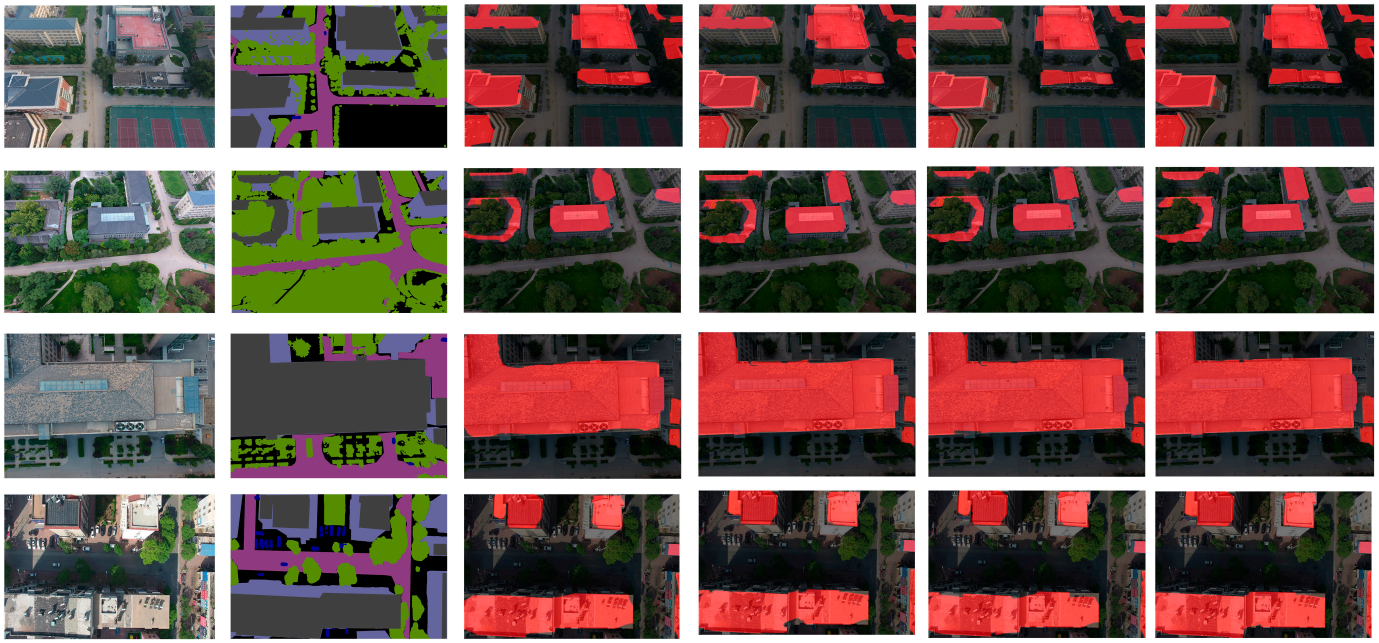

Figure 8. Ablation experiments on the UDD6 dataset.

Interestingly, the segmentation representation depicted in Figure 8f, stemming from the synergistic utilization of the morphology-driven channel and composite loss function, ensures the holistic depiction of buildings while simultaneously addressing the limitations inherent to the improved single-channel Deeplabv3+.

4. Discussion

Accurate extraction of buildings from drone images is critical for urban planning, disaster response, infrastructure monitoring, and a variety of other applications, greatly improving our ability to understand and manage urban environments [53,54]. Nonetheless, several pivotal factors influence building extraction, demanding further attention and resolution.

4.1. Influence of UAV Imaging on Model

4.1.1. Imaging Conditions of Uneven Illumination

One of the main challenges in the utilization of UAV and remote sensing images for building extraction lies in the inconsistency of lighting conditions and the presence of shadows, which can significantly impact segmentation accuracy [8–15]. The study is shown in Figure 5. The literature [55] shows that although TPT-GAN is effective in extracting foreground from backgrounds with uneven shading changes of no more than 50%, its applicability weakens when extended to complex architectural scenes. This limitation is evident as extended branches, as employed in the literature [56,57], are utilized to handle scenes with pronounced color changes, imposing high demands on the diversity and universality of labeled data. To address this challenge, our study introduces the innovative concept of an optical flow estimation channel. Table 2 shows the effectiveness of the channel. Using the characteristics of optical flow estimation, the problem of low accuracy of building extraction caused by some light and shade changes is made up for.

4.1.2. Multiresolution Imaging Properties of UAV

In the field of UAV images, as highlighted in the literature [58], the distance between the building and the camera may lead to reduced image clarity, and the size of the building will also show significant changes. As shown in Figure 2 of reference [59], relying only on

an interpolation algorithm to compensate sharpness leads to instability in texture details and spatial information, seriously affecting the accuracy of extraction. To overcome this limitation, our approach includes an HCAM module within the BA-ASPP framework. This module employs a multiclass feature extraction mechanism that integrates max pooling, average pooling, building shape features, and original features to enhance the network's sensitivity to remote contextual information.

The advantage of drone-based imagery lies in its ability to capture images at various resolutions, each revealing unique characteristics of the buildings [60]. High-dimensional images offer clearer textures, low-dimensional images prioritize contour feature extraction, and medium-dimensional images consider both aspects. As can be seen from Table 3, the MFFM integrated into our network extracts basic information from these three dimensions to ensure that image details affected by different resolutions are effectively utilized. Furthermore, it is vital for the loss function to provide meaningful feedback during each iteration. The literature [61] lacks sensitivity to building structures in the face of training that usually deals with unbalanced samples. To address these challenges, our study introduced a compound loss function with the mIoU in Table 1 at 82.69% and 79.37%, respectively, effectively forcing the network to make better decisions based on building shape.

4.2. Influence of Label Ratio on Model Accuracy

In the realm of semantic segmentation, the distribution of labeled and unlabeled data in the training set significantly impacts the model's accuracy. Therefore, the subsequent discussion on label ratios remains crucial in understanding the model's influence—specifically, that of AffinityNet, AdvSemiSeg, SemiCycleGan, and CCVC—on extracting buildings from UAV images. References [16,17] highlight the challenge of model generalization to new, unseen scenes when the training dataset lacks diversity due to an insufficient number of labeled samples, and its evaluation indicators are much lower than those in Table 1. The CCVC network used in Table 3 of reference [19] has an mIoU value of 77.3% under 25% labeled data. Since the sample set it trains is a multitype sample set, the applicability of its model is reduced when it is applied to complex buildings. As shown in Table 1 of reference [18], the accuracy tested on both Cityscapes and VOC datasets was less than 50% when using a 30% labeling rate. In contrast, when the proportion of labeled data is between 50 and 70 percent, balancing the labeling scales has been shown to help improve the accuracy of the model. As shown in Table 4, our experiment displayed a significant correlation between the unbalanced labeling rate and reduced model generalization, with the mIoU reduced by only 11.08% and 8.78% when the labeled data were reduced by 70%.

Table 4. The effect of label ratio on segmentation results.

Label Ratio (%)	Drone Building Dataset			UDD6 Dataset		
	F1 Score (%)	mIoU (%)	Precision (%)	F1 Score (%)	mIoU (%)	Precision (%)
30	79.36	82.69	80.56	77.68	79.37	79.43
50	81.39	83.53	83.16	81.92	83.33	82.67
70	87.25	86.41	87.36	85.42	86.18	87.52
100	94.68	93.77	93.36	89.21	91.47	92.77

An intriguing aspect of our investigation was the trade-off between label ratios and the associated labor labeling costs. As mentioned in references [62,63], an increase in the number of labeled samples exhibits a positive correlation with improved model accuracy. However, this improvement comes at the cost of additional expenses and labor-intensive efforts. Without exception, as the proportion of manually labeled data increases from 30% to 70%, the networks mentioned in the literature [16–19] all show commendable segmentation accuracy. However, these networks are generalized extraction models that lack the accuracy required for specific application domains, and none of them achieve the accuracy shown in Table 1. It is not difficult to see from Tables 3 and 4 that we not only reduced the need for labeled data but also designed a semi-supervised learning mechanism that takes

into account the construction of features in a specific segmentation domain. When the proportion of labeled data is 30%, the mIoU still reaches 82.69%.

4.3. Optical Flow Estimation and Motion Image Segmentation

Optical flow estimation is commonly employed in the analysis of moving objects [21]. Figure 6 and Table 2 in reference [22] use fusion segmentation and redistribution strategies to segment images and effectively generate motion components to calculate scene flow. Although achieving PRSM: 2.04% in the five-pixel category, visual inspection reveals that the model does not incorporate specific constraints to mitigate errors induced by variations in light and shade. In another approach, Sotirios et al. focused on enhancing motion trajectory capture, ensuring the goal error within the camera field of view remained below 60 pixels. Despite the effectiveness of the polarization-based UAV attitude estimation and segmentation method proposed in the literature [24] across the entire scene, it falls short of achieving an mIoU value of 82.69% in Table 1 when confronted with low-label data and complex building structures.

The significance of improved optical flow estimation in building segmentation is further elucidated in Figure 9. In particular, Figure 9c represents the initial optical flow estimation image, while Figure 9d showcases the optical flow estimation image after the application of morphological information constraints. Notably, the constrained optical flow estimation image visibly alleviates segmentation challenges arising from uneven shadows and illumination.

4.4. Limitations and Perspective

In conclusion, while our method exhibits robust segmentation accuracy across various datasets and experimental conditions, it is essential to recognize that there is no universal solution applicable to all scenarios. The effectiveness of our network may vary depending on factors such as data acquisition methods, drone shooting angles, and adverse weather conditions. For instance, the network excels at handling shadow and lighting changes within a dataset. However, its performance in occlusion situations is likely to be similar to that of mainstream networks. While our network does not feature a dedicated module to address occlusion, it extends the extraction of contextual information through the HCAM and MFFM in the BA-ASPP to enhance building feature sensitivity [64]. Therefore, a more comprehensive assessment is needed to thoroughly understand its capabilities in different scenarios. Additionally, during network training, it is crucial to input data in accordance with the sequence of images captured by the UAV to meet the format requirements of the optical flow estimation channel. To address this limitation, future research efforts may prioritize the development of intelligent image processing [65] and sorting algorithms [66]. These advancements would enable automatic processing of out-of-order image inputs and harness the multi-view capabilities of drones to enhance occlusion modeling. This, in turn, would improve the model's adaptability and practicality. After training, the model can be integrated into programmable DJI or Pegasus experimental machines for real-time monitoring and accurate building segmentation at low and medium altitudes, with important potential applications in urban planning, building surveillance, military and intelligence operations, and navigation and mapping.

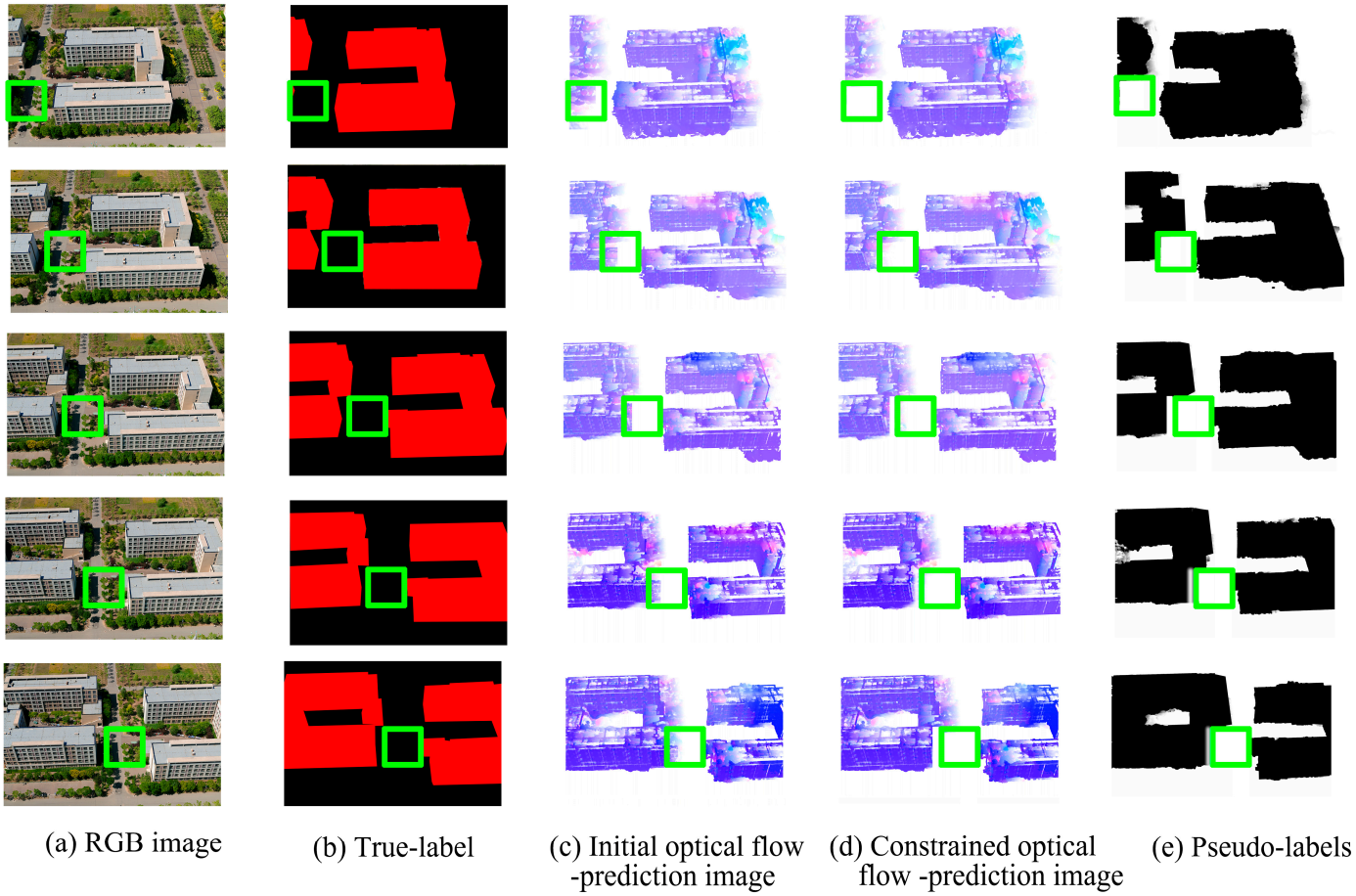

Figure 9. Morphology-driven optical flow estimation example.

5. Conclusions

A comprehensive method is adopted to address the challenges posed by diverse architectural forms and complex extraction requirements in the Caofeidian District of Tangshan. An optical flow estimation channel was introduced to improve performance under varying lighting conditions. Multilevel feature fusion modules and hierarchical channel attention modules were integrated to address texture information in different building resolutions. Weighted cross-entropy loss was applied to ensure model stability. Building perception loss provided feedback on structure information. Based on experiments conducted on two datasets, the following conclusions can be drawn:

(1) The optical flow estimation channel proves effective in compensating for complex background defects when the ratio of light and shade change in the building image is no more than 50% of the total image.
(2) In the case of UAV images exhibiting multiscale and multiresolution characteristics, the hierarchical channel attention module (HCAM) with a cascade structure captures potential building information across high, middle, and low dimensions and different spatial contexts.
(3) Even with only 30% of the labeled datasets, the mIoU of the two-channel parallel structure still reached 82.69% and 79.37% on the two UAV datasets, respectively. And when the labeled data increased from 30% to 70%, the accuracy improved the fastest.
(4) The experiment demonstrated that when irregular buildings dominated the study area, the building perception loss forced the network to prioritize the building's structural information, and the actual result was a significant improvement in key metrics, including F1 scores, mIoU, and accuracy.

Author Contributions: Conceptualization, W.Z., C.W., W.M. and M.L.; methodology, W.Z. and C.W.; software, W.Z., C.W. and W.M.; validation, W.Z.; formal analysis, W.Z. and W.M.; investigation, W.Z. and C.W.; resources, W.Z. and C.W.; data curation, W.Z. and C.W.; writing—original draft preparation, W.Z. and W.M.; writing—review and editing, W.Z. and M.L.; visualization, W.Z. and C.W.; supervision, W.Z., C.W. and W.M.; project administration, W.Z., W.M. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Central Guidance and Local Science and Technology Development Funds (grant no. 236Z3305G), the Natural Science Foundation of Hebei Province, China (grant nos. D2022209005 and D2019209322), the Science and Technology Project of Hebei Education Department (grant no. BJ2020058), the Key Research and Development Program of Science and Technology Plan of Tangshan, China (grant no. 22150221J), and the North China University of Science and Technology Foundation (grant nos. BS201824 and BS201825).

Data Availability Statement: The authors would like to thank the team of the UDD6 dataset for the data and experiments. Due to privacy restrictions and experimental requirements, the drone building dataset is not publicly available for the time being.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Güneralp, B.; Zhou, Y.; Ürge-Vorsatz, D.; Gupta, M.; Yu, S.; Patel, P.L.; Fragkias, M.; Li, X.; Seto, K.C. Global Scenarios of Urban Density and Its Impacts on Building Energy Use through 2050. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 8945–8950. [CrossRef] [PubMed]
2. Claassens, J.; Koomen, E.; Rouwendal, J. Urban Density and Spatial Planning: The Unforeseen Impacts of Dutch Devolution. *PLoS ONE* **2020**, *15*, e0240738. [CrossRef]
3. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction from High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
4. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
5. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
6. Moghalles, K.; Li, H.C.; Alazeb, A. Weakly Supervised Building Semantic Segmentation Based on Spot-Seeds and Refinement Process. *Entropy* **2022**, *24*, 16. [CrossRef] [PubMed]
7. Yang, W.S.; Liu, X.Y.; Zhang, Y.J.; Wan, Y.; Ji, Z. Object-based building instance segmentation from airborne LiDAR point clouds. *Int. J. Remote Sens.* **2022**, *43*, 6783–6808. [CrossRef]
8. Ye, H.; Liu, S.; Jin, K.; Cheng, H. CT-UNet: An Improved Neural Network Based on U-Net for Building Segmentation in Remote Sensing Images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021.
9. Larochelle, H.; Hinton, G.E. Learning to combine foveal glimpses with a third-order Boltzmann machine. In Proceedings of the 24th Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–9 December 2010; pp. 1243–1251.
10. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
11. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision, Munich, German, 8–14 September 2018; pp. 3–19.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
13. Wei, X.L.; Li, W.; Zhang, M.M.; Li, Q.L. Medical Hyperspectral Image Classification Based on End-to-End Fusion Deep Neural Network. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 4481–4492. [CrossRef]
14. Wen, T.; Ding, S.; Lang, H.; Lu, J.J.; Yuan, Y.; Peng, Y.C.; Chen, J.; Wang, A.D. Automated pavement distress segmentation on asphalt surfaces using a deep learning network. *Int. J. Pavement Eng.* **2022**, 1–14. [CrossRef]
15. You, H.F.; Yu, L.; Tian, S.W.; Ma, X.; Xing, Y. Medical image segmentation based on dual-channel integrated cross-layer residual algorithm. *Multimed. Tools Appl.* **2023**, *82*, 5587–5603. [CrossRef]
16. Ma, T.; Zhang, A. Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods* **2018**, *145*, 16–24. [CrossRef]

17. Hung, W.C.; Tsai, Y.H.; Liou, Y.T.; Lin, Y.-Y.; Yang, M.-H. Adversarial learning for semi-supervised semantic segmentation. *arXiv* **2018**, arXiv:1802.07934.
18. Mondal, A.K.; Agarwal, A.; Dolz, J.; Desrosiers, C. Revisiting CycleGAN for semi-supervised segmentation. *arXiv* **2019**, arXiv:1908.11569.
19. Wang, Z.; Zhao, Z.; Xing, X.; Xu, D.; Kong, X.; Zhou, L. Conflict-Based Cross-View Consistency for Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 19585–19595.
20. Li, M.; Shen, Q.K.; Xiao, Y.; Liu, X.G.; Chen, Q.H. PolSAR Image Building Extraction with G^0 Statistical Texture Using Convolutional Neural Network and Superpixel. *Remote Sens.* **2023**, *15*, 23. [CrossRef]
21. Ding, J.; Zhang, Z.; Yu, X.X.; Zhao, X.W.; Yan, Z.G. A Novel Moving Object Detection Algorithm Based on Robust Image Feature Threshold Segmentation with Improved Optical Flow Estimation. *Appl. Sci.* **2023**, *13*, 19. [CrossRef]
22. Hu, F.Z.; Zhang, Z.L.; Hu, X.; Chen, T.T.; Guo, H.; Quan, Y.; Zhang, P.J. A scene flow estimation method based on fusion segmentation and redistribution for autonomous driving. *IET Contr. Theory Appl.* **2023**, *17*, 1779–1788. [CrossRef]
23. Aspragkathos, S.N.; Karras, G.C.; Kyriakopoulos, K.J. A Hybrid Model and Data-Driven Vision-Based Framework for the Detection, Tracking and Surveillance of Dynamic Coastlines Using a Multirotor UAV. *Drones* **2022**, *6*, 28. [CrossRef]
24. Shabayek, A.E.; Demonceaux, C.; Morel, O.; Fofi, D. Vision Based UAV Attitude Estimation: Progress and Insights. *J. Intell. Robot. Syst* **2012**, *65*, 295–308. [CrossRef]
25. Zhu, H.; Ma, W.P.; Li, L.L.; Jiao, L.C.; Yang, S.Y.; Hou, B. A Dual-Branch Attention fusion deep network for multiresolution remote-Sensing image classification. *Inf. Fusion* **2020**, *58*, 116–131. [CrossRef]
26. Chen, Y.; Wang, Y.; Lu, P.; Chen, Y.; Wang, G. Large-Scale Structure from Motion with Semantic Constraints of Aerial Images. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Guangzhou, China, 23–26 November 2018; pp. 347–359.
27. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
28. Sheng, C.Y. Research on the Application of Data Set Expansion Based on Conditional Generative Adversarial Network in Right Ventricle Segmentation. Ph.D. Thesis, Suzhou University, Suzhou, China, 2021. (In Chinese).
29. Wang, P.; Bai, X. Thermal infrared pedestrian segmentation based on conditional GAN. *IEEE Trans. Image Process.* **2019**, *28*, 6007–6021. [CrossRef]
30. Anilkumar, P.; Venugopal, P. An Enhanced Multi-Objective-Derived Adaptive DeepLabv3 Using G-RDA for Semantic Segmentation of Aerial Images. *Arab. J. Eng* **2023**, *48*, 10745–10769. [CrossRef]
31. Li, X.L.; Li, Y.Y.; Ai, J.Q.; Shu, Z.H.; Xia, J.; Xia, Y.P. Semantic segmentation of UAV remote sensing images based on edge feature fusing and multi-level upsampling integrated with Deeplabv3+. *PLoS ONE* **2023**, *18*, e0279097. [CrossRef] [PubMed]
32. Yu, L.J.; Zeng, Z.X.; Liu, A.; Xie, X.C.; Wang, H.P.; Xu, F.; Hong, W. A Lightweight Complex-Valued DeepLabv3+for Semantic Segmentation of PolSAR Image. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens* **2022**, *15*, 930–943. [CrossRef]
33. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
34. Cho, W.; Choi, Y. LMGAN: Linguistically Informed Semi-Supervised GAN with Multiple Generators. *Sensors* **2022**, *22*, 17. [CrossRef] [PubMed]
35. Bruhn, A.; Weickert, J.; Schnorr, C. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *Int. J. Comput. Vis* **2005**, *61*, 211–231. [CrossRef]
36. Xiang, X.Z.; Yu, Z.T.; Lv, N.; Kong, X.D.; El Saddik, A. Attention-Based Generative Adversarial Network for Semi-supervised Image Classification. *Neural Process. Lett* **2020**, *51*, 1527–1540. [CrossRef]
37. Kim, K.K.; Ban, S.W.; Lee, K.I. Motion estimation with optical flow-based adaptive search region. *IEICE Trans. Fundam. Electron. Commun. Comput* **2001**, *E84A*, 1529–1531.
38. Zheng, J.; Wang, H.Y.; Pei, B.N. Robust optical flow estimation based on wavelet. *Signal Image Video Process.* **2019**, *13*, 1303–1310. [CrossRef]
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
40. Zhang, L.; Wang, M.Y.; Fu, Y.J.; Ding, Y.H. A Forest Fire Recognition Method Using UAV Images Based on Transfer Learning. *Forests* **2022**, *13*, 20. [CrossRef]
41. Zhang, R.L.; Zhu, Y.J.; Ge, Z.S.J.; Mu, H.B.; Qi, D.W.; Ni, H.M. Transfer Learning for Leaf Small Dataset Using Improved ResNet50 Network with Mixed Activation Functions. *Forests* **2022**, *13*, 21. [CrossRef]
42. Rasin, A.G. Computation of generating symmetries. *Commun. Nonlinear Sci. Numer. Simul.* **2023**, *118*, 12. [CrossRef]
43. Brown, J.L. SDMtoolbox: A python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods Ecol. Evol.* **2014**, *5*, 694–700. [CrossRef]
44. Kim, C.E.; Stojmenovic, I. Sequential and parallel approximate convex hull algorithms. *Comput. Artif. Intell.* **1995**, *14*, 597–610.
45. Zhang, Z.L.; Sabuncu, M. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018; pp. 8792–8802.

46. Dong, Y.Q.; Zhang, L.; Cui, X.M.; Ai, H.B.; Xu, B.A. Extraction of Buildings from Multiple-View Aerial Images Using a Feature-Level-Fusion Strategy. *Remote Sens.* **2018**, *10*, 30. [CrossRef]
47. He, L.; Shan, J.; Aliaga, D. Generative Building Feature Estimation From Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 13. [CrossRef]
48. Brown, G.; Pocock, A.; Zhao, M.J.; Lujan, M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
49. Kang, J.; Fernandez-Beltran, R.; Sun, X.; Ni, J.G.; Plaza, A. Deep Learning-Based Building Footprint Extraction With Missing Annotations. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5. [CrossRef]
50. Lu, T.T.; Ming, D.; Lin, X.G.; Hong, Z.L.; Bai, X.D.; Fang, J. Detecting Building Edges from High Spatial Resolution Remote Sensing Imagery Using Richer Convolution Features Network. *Remote Sens.* **2018**, *10*, 19. [CrossRef]
51. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003; pp. 1398–1402.
52. Ma, K.D.; Wu, Q.B.; Wang, Z.; Duanmu, Z.; Yong, H.; Li, H.; Zhang, L. Group MAD Competition? A New Methodology to Compare Objective Image Quality Models. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 166–1673.
53. Huang, X.; Zhang, L. Morphological Building/Shadow Index for Building Extraction From High-Resolution Imagery Over Urban Areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 161–172. [CrossRef]
54. Ding, Q.; Shao, Z.F.; Huang, X.; Feng, X.X.; Altan, O.; Hu, B. Consistency-guided lightweight network for semi-supervised binary change detection of buildings in remote sensing images. *GISci. Remote Sens.* **2023**, *60*, 26. [CrossRef]
55. Sakkos, D.; Ho, E.S.L.; Shum, H.P.H. Illumination-Aware Multi-Task GANs for Foreground Segmentation. *IEEE Access* **2019**, *7*, 10976–10986. [CrossRef]
56. Vrsnak, D.; Domislovic, I.; Subasic, M.; Loncaric, S. Framework for Illumination Estimation and Segmentation in Multi-Illuminant Scenes. *IEEE Access* **2023**, *11*, 2128–2137. [CrossRef]
57. Zhang, Z.; Li, Y.; Shin, B.S. Robust color medical image segmentation on unseen domain by randomized illumination enhancement. *Comput. Biol. Med.* **2022**, *145*, 14. [CrossRef] [PubMed]
58. Chen, J.; Xu, W.F.; Yu, Y.; Peng, C.L.; Gong, W.P. Reliable Label-Supervised Pixel Attention Mechanism for Weakly Supervised Building Segmentation in UAV Imagery. *Remote Sens.* **2022**, *14*, 3196. [CrossRef]
59. Xu, G.; Ling, R.; Deng, L.S.; Wu, Q.; Ma, W.Y. Image Interpolation via Gaussian-Sinc Interpolators with Partition of Unity. *CMC-Comput. Mat. Contin.* **2020**, *62*, 309–319. [CrossRef]
60. Fatty, A.; Li, A.J.; Yao, C.Y. Instance segmentation based building extraction in a dense urban area using multispectral aerial imagery data. *Multimed. Tools Appl.* **2023**, 1. [CrossRef]
61. Niu, M.J.; Zhang, Y.J.; Yang, G.; Wang, Z.W.; Liu, J.W.; Cui, Z.W. Semantic segmentation for remote sensing images via dense feature extraction and companion loss neural network. *Int. J. Remote Sens.* **2021**, *42*, 8640–8660. [CrossRef]
62. Ahfock, D.; McLachlan, G.J. Harmless label noise and informative soft-labels in supervised classification. *Comput. Stat. Data Anal.* **2021**, *161*, 12. [CrossRef]
63. Lee, J.; Ilyas, T.; Jin, H.; Lee, J.; Won, O.; Kim, H.; Lee, S.J. A pixel-level coarse-to-fine image segmentation labelling algorithm. *Sci. Rep.* **2022**, *12*, 18. [CrossRef]
64. Zhou, G.Q.; Wang, Y.F.; Yue, T.; Ye, S.Q.; Wang, W. Building Occlusion Detection From Ghost Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 1074–1084. [CrossRef]
65. Cai, W.; Wen, X.D.; Tu, Q.; Guo, X.J. Research on image processing of intelligent building environment based on pattern recognition technology. *J. Vis. Commun. Image Represent.* **2019**, *61*, 141–148. [CrossRef]
66. Xue, L.L.; Zeng, P.; Yu, H.B. SETNDS: A SET-Based Non-Dominated Sorting Algorithm for Multi-Objective Optimization Problems. *Appl. Sci.* **2020**, *10*, 15. [CrossRef]