Research Article

Rongchun Zhang, Meiru Jing, Guanming Lu*, Xuefeng Yi, Shang Shi, Yi Huang, and Lanfa Liu*

Building element recognition with MTL-AINet considering view perspectives

https://doi.org/10.1515/geo-2022-0506
received February 22, 2023; accepted June 10, 2023

Abstract: The reconstruction and analysis of building models are crucial for the construction of smart cities. A refined building model can provide a reliable data support for data analysis and intelligent management of smart cities. The colors, textures, and geometric forms of building elements, such as building outlines, doors, windows, roof skylights, roof ridges, and advertisements, are diverse; therefore, it is challenging to accurately identify the various details of buildings. This article proposes the Multi-Task Learning AINet method that considers features such as color, texture, direction, and roll angle for building element recognition. The AINet is used as the basis function; the semantic projection map of color and texture, and direction and roll angle is used for multi-task learning, and the complex building facade is divided into similar semantic patches. Thereafter, the multi-semantic features are combined using hierarchical clustering with a region adjacency graph and the nearest neighbor graph to achieve an accurate recognition of building elements. The experimental results show that the proposed method has a higher accuracy for building detailed edges and can accurately extract detailed elements.

Keywords: multi-view stereo images, building elements, multi-task learning, similar semantic patches, segmentation, clustering

*** Corresponding author: Guanming Lu**, College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, No. 9, Wenyuan Road, Yadong New District, 210023, Nanjing, China, e-mail: lugm@njupt.edu.cn
*** Corresponding author: Lanfa Liu**, Hubei Provincial Key Laboratory for Geographical Process Analysis and Simulation, College of Urban and Environmental Sciences, Central China Normal University, No. 152, Luoyu Road, Hongshan District, Wuhan, 430079, China, e-mail: lanfa@ccnu.edu.cn
Rongchun Zhang: School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, No. 9, Wenyuan Road, Yadong New District, 210023, Nanjing, China; College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, No. 9, Wenyuan Road, Yadong New District, 210023, Nanjing, China
Meiru Jing, Shang Shi, Yi Huang: School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, No. 9, Wenyuan Road, Yadong New District, 210023, Nanjing, China
Xuefeng Yi: Bei Fang Investigation, Design & Research Co. Ltd, No. 60 Dongting Road, Tianjin, 300222, China

1 Introduction

The reconstruction and analysis of building models are important aspects of smart city construction. Refined building models can provide a reliable data support for data analysis and intelligent management of smart cities [1]. However, owing to the diverse colors, textures, and forms of details, such as building contours, doors and windows, roofs, and advertisements, it is difficult to accurately identify the various details of buildings. An effective improvement in the accuracy and integrity of the extraction of building facade elements is imperative for the construction and development of smart cities. In recent years, semantic segmentation technology has developed rapidly in the field of building facade information extraction [2].

Semantic segmentation is crucial for understanding images in image processing and computer vision tasks [3]. The basic idea of semantic segmentation is to classify every pixel in the image and determine the category of each point (such as belonging to the background, edge, or subject). Point cloud segmentation is divided according to feature points such as space, geometry, and texture, and point clouds in the same group have similar features [4]. In recent years, methods based on image semantic segmentation and point cloud segmentation have been increasingly used in building detail recognition. (a) Image-based building information extraction: traditional building information extraction methods include methods based on K-means clustering [5], pixel-level and region-level [6], and morphological operations [7]. Most of the low-level features of the image are used, such as shadows, structures, edges, light, and dark. Huang et al. [8] proposed using a group of morphological operations to represent the inherent structural characteristics of buildings (such as brightness, contrast, and size) and

Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

automatically detect buildings from images. Zheng et al. [9] proposed the image edge regularity and shadow line indices as new features for boundary recognition of specific buildings. However, owing to the increasing amount of data and complex reality of buildings, these methods based on the artificial acquisition of low-level features in images cannot realize automatic and accurate building recognition. With the development of deep learning in recent years, convolutional neural networks (CNNs) [10] have replaced the manual and tedious feature design process by learning the semantic hierarchy of image data, thereby realizing the automation of image feature extraction. Long et al. [11] converted CNNs into fully convolutional networks (FCNs); however, the segmentation was not good in terms of details. Encoder–decoder model structures, such as SegNet [12] and U-Net [13], can effectively solve this problem. To further improve accuracy, Deeplab series improved segmentation accuracy by extending the receptive domain, learning multi-scale context information, and adding post-processing structures [14]. Although these attempts and improvements of different networks have improved the segmentation accuracy [15], they still suffer from low edge accuracy when extracting details of complex building structures [16]. (b) Point cloud-based building information extraction: the semantic features of building facades are extracted from noisy building point clouds. Traditional methods use the generated geographic reference image, coordinate information, shape features, and prior knowledge to extract windows, doors, billboards, and other elements. Some scholars proposed some other traditional methods, e.g., point cloud clustering and segmentation-based methods [17], similarity matching-based methods [18], and plane fitting-based methods [19]. In contrast, deep learning-based methods obtain higher segmentation accuracy and do not require a manual design of feature extraction operators. Therefore, using deep learning to process point cloud data of building facades has important research significance. For example, VoxNet [20], Kd-Net [21], OCT-NET [22], and others first convert irregular point clouds into regular voxel grids, but significantly increase the amount of data and their computational efficiency is low. Multi-view CNNs (MV-CNNs) [23] perform semantic segmentation by projecting a point cloud onto a 2D plane, but lose spatial information during dimension reduction. PointNet greatly improved the accuracy of point cloud segmentation. Thereafter, some scholars improved it by considering the local features of point clouds and proposed PointNet++ [24]. PointNet is an end-to-end network that directly processes point clouds. PointNet++ divides the entire point cloud into a series of local regions and then runs PointNet in each local region to extract the local features of each region, which can better handle the local and global structural information of point cloud data. Although this method improved the segmentation accuracy, it still ignored many details, resulting in under- or over-segmentation [25]. Deep learning-based methods are driven by a massive amount of data [26]. However, the current lack of a point cloud annotation reference dataset for large-scale urban building information extraction causes considerable inconvenience in the extraction of building details based on point cloud data [27,28].

With regard to building detail recognition, the recognition results of image-based segmentation are still not ideal owing to a lack of 3D geometric information. Point cloud-based segmentation methods are ineffective for color semantic extraction when there are no obvious geometric features. For example, architectural images are affected by external elements such as shadows, occlusions, and shading, which make 2D color and texture semantics susceptible to interference [29]. In the case of building facades with different directions and roll angles, it is difficult to extract all the details based on color and texture features alone and to accurately extract information with obvious texture features, such as advertisements on flat walls, using only a point cloud [30]. Therefore, it is difficult to realize the accurate recognition of building elements using only 2D or 3D information. For most deep network architectures, the standard convolution operations are defined under regular grids, which largely affects the processing efficiency of irregular grids. As an end-to-end trainable network, superpixel segmentation with FCNs introduced nondifferentiable modules. However, both the skip-connect operation and the low–level pixel–pixel relationship have bad effects on the segmentation results. The AINet directly predicts the pixel–superpixel relationship by integrating the association implantation (AI) module into the FCN, which effectively improves the segmentation efficiency. In addition, the loss function considering the boundary-perceiving loss helps to improve the edge consistency of superpixels. Based on the aforementioned analysis, this study proposes a building element recognition method, called MTL-AINet (Multi-Task Learning AINet), based on 2D color and texture, 3D direction, and roll angle semantics. The proposed building facade segmentation method based on semantic projection correlation graphs of color, texture, direction, and roll angle addresses the shortcomings of traditional semantic segmentation methods based on image texture and improves the accuracy of building element recognition.

2 Methodology

In this study, the association learning of 2D color and texture features and 3D direction and roll angle features was

realized to enable detail recognition and extraction of building facades in different scenes. A flowchart of the proposed method is shown in Figure 1.

First, a dense point cloud model is generated through a multi-view stereo (MVS) reconstruction technique using building sequence images, and then, the direction and roll angle semantics are calculated. Considering the building facade orientation, it is divided into different viewpoints for planar image projection. Finally, the point cloud model is projected onto a 2D image to obtain pixel-level correlated direction, roll angle, and color maps. The fusion probability Q is obtained through multi-task integrated learning, and then, multi-semantic homogeneous patches are obtained. According to the different requirements of building element extraction tasks, a multi-semantic hierarchical clustering strategy was adopted to obtain the clustering results of building facades and achieve fine recognition and extraction of building elements under different scenarios.

2.1 Projections from multiple-view perspectives

Generally, both the color and texture are used for super-pixel segmentation. A building surface has rich color and texture semantics. Building texture refers to the similar structures presented by the color and material of the building surface, and the color and texture can characterize the 2D features of different objects on the building

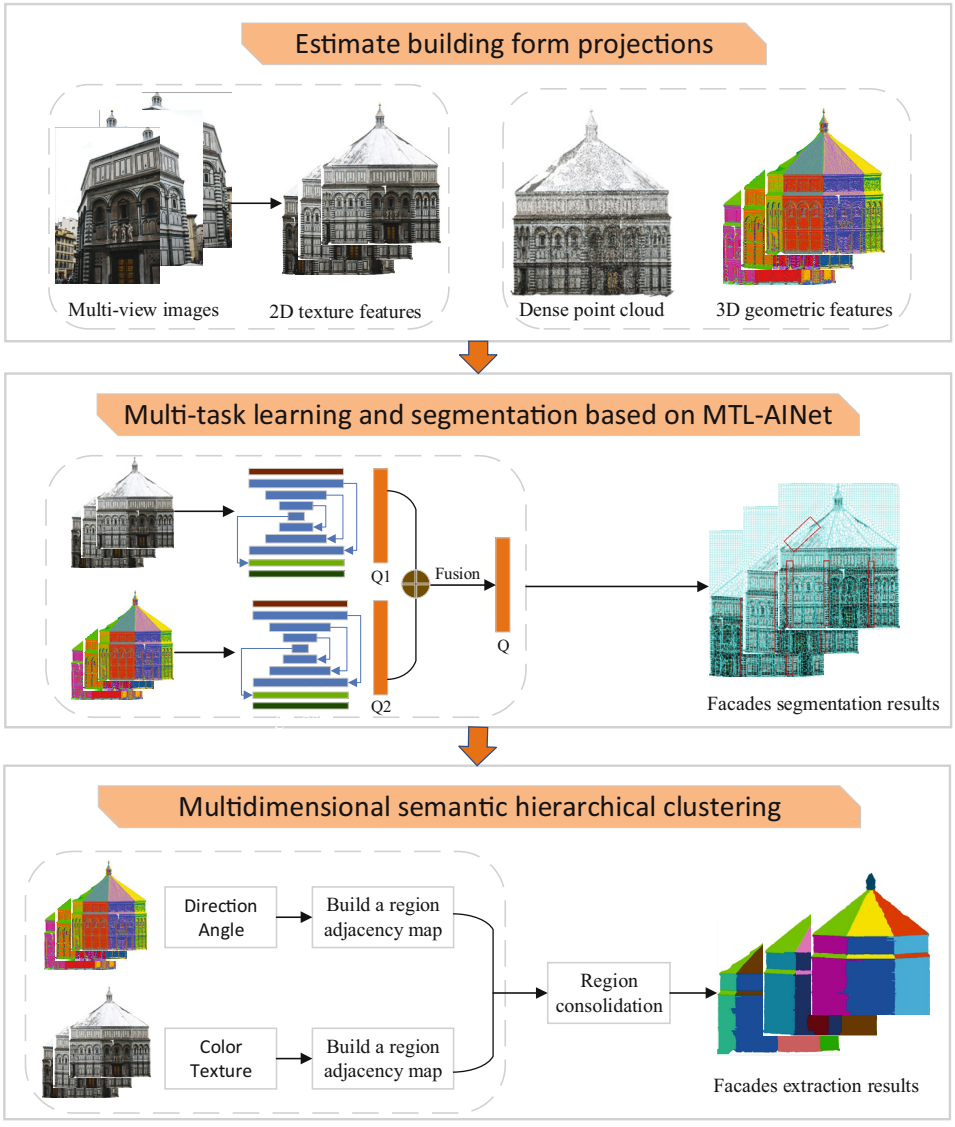


Figure 1: Flowchart of the proposed method.

surface, such as walls, windows, and doors [31,32]. For example, in Figure 2(a), the color and texture features of the building facade, windows, and roofs differ significantly, and the building detail elements such as windows, facades, and roofs can be accurately identified. However, as shown in Figure 2(b), the color and texture of the building facade are not consistent with those of the facade, and it is difficult to accurately identify the building elements using only the color and texture. The dilapidated building facade in Figure 2(c) also shows an inconsistency between the 2D color and texture and the 3D feature of the facade. In conclusion, although 2D color and texture semantics can be important for building detail recognition, it is difficult to achieve an accurate recognition of complex building details using them.

It is difficult to achieve an accurate recognition of complex building details using only 2D color and texture semantics. Adding 3D features can overcome the limitations of inaccurate 2D features and increase the recognition accuracy. Figure 3 shows the geometric schematics and projection interpretation of multiple features for a building. The color feature is displayed in Figure 3(a). As shown in Figure 3(b), $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$, respectively, denote the normal vectors for the roof and the facade of a building. The geometric relationship about the orientation and roll angle features is described in Figure 3(c). The parameter θ in Figure 3(c) represents the roll angle, i.e., the included angle between the normal vector of a plane (e.g., roofs and facades of a building in Figure 3) and the horizontal plane. $\boldsymbol{n}_1'$ represents the projection of the normal vector $\boldsymbol{n}_1$ in the horizontal plane, and the direction is represented by the intersection angle between $\boldsymbol{n}_1'$ and the true north direction. In this study, the normal vectors of the point cloud model are calculated based on the planar-

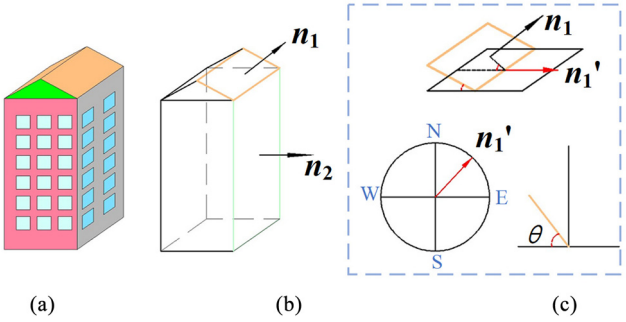


Figure 3: Geometric interpretation of multiple features for building elements: (a) a building, (b) normal vectors of the roof and the facade of building, and (c) the orientation and roll angle of the building roof.

fitting method. The combined use of direction, roll angle, color, and texture features can enable an accurate, quick, and automatic characterization of the detailed elements of buildings. Figure 4(a–d) shows the schematic diagrams of projection for building facades under multiple-view perspectives. Each building facade is filling with a color.

The point cloud obtained from the sequence image was used to calculate the direction and roll angle semantics of the building; however, the buildings in different scenes had different facade orientations and structural postures. To establish the correlation mapping between 2D and 3D features while considering the overall shape of the building, this study combines the facade orientation of an actual building with the design of multi-angle projection, divides it into different perspectives, and selects multiple perspectives of the building for projection to establish the mapping relationship between the direction and roll angle of different building perspectives and the color and texture semantics (Figure 4(d)), which provides the basic data for multi-task learning.



Figure 2: Different building color and texture semantics: (a) distinct differences in color and texture features, (b) interference with color and texture features, and (c) color and texture of old building surfaces.

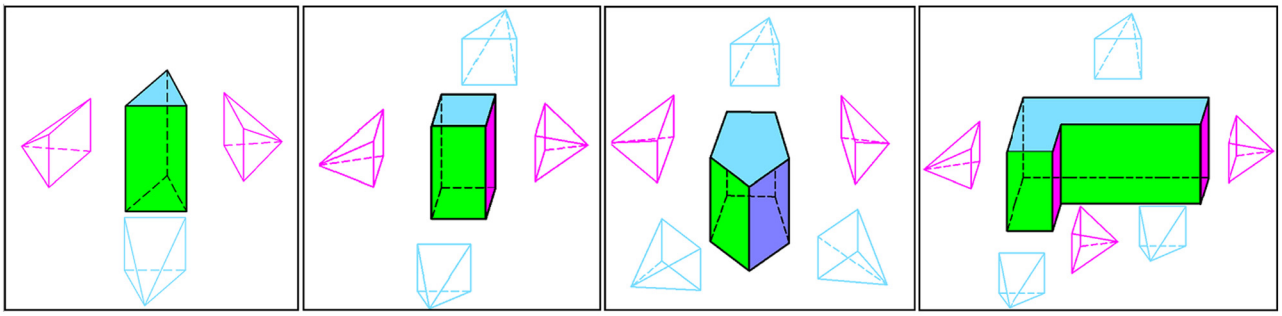


Figure 4: The schematic diagrams of projection of building facades under various view perspectives: (a) three view perspectives, (b) four view perspectives, (c) five view perspectives, and (d) six view perspectives.

2.2 Multi-task learning and segmentation based on MTL-AINet

AINet adopts an encoder–decoder architecture to acquire deeper features and output feature maps with superpixel embeddings as the input image sensory field increases in the encoding stage, which is fed into the decoding stage to generate the pixel–superpixel association maps. Meanwhile, the superpixel and pixel embeddings generated in the encoding and decoding stages, respectively, realize the direct interaction between the pixel and its field through the AI module, i.e., the superpixel is embedded around pixel p so that the network can capture the association between p and its neighboring grids and realize the pixel–superpixel association, which is more in line with the goal of superpixel segmentation. A matrix Q of $h \times w \times 9$ is finally obtained as the relationship between p and the surrounding nine superpixels [33], i.e., the probability that p belongs to the surrounding nine superpixels. The central information of the superpixel $\mathbf{c}_s = (\mathbf{u}_s, \mathbf{l}_s)$ is calculated using the association matrix, and the image is reconstructed based on the superpixel features using the association between the pixel and the superpixel. The specific calculation is as follows:

$$\boldsymbol{u}_s = \frac{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} \boldsymbol{f}(\mathbf{p}) \cdot q_s(\mathbf{p})}{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} q_s(\mathbf{p})}, \quad l_s = \frac{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} \mathbf{p} \cdot q_s(\mathbf{p})}{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} q_s(\mathbf{p})}, \tag{1}$$

$$\boldsymbol{f}'(\mathbf{p}) = \sum_{\mathbf{s}\in\mathcal{N}_\mathbf{p}} \boldsymbol{u}_\mathbf{s} \cdot q_s(\mathbf{p}), \quad \boldsymbol{p}' = \sum_{\mathbf{s}\in\mathcal{N}_\mathbf{p}} l_\mathbf{s} \cdot q_s(\mathbf{p}), \tag{2}$$

where $f(\mathbf{p})$ is the feature of pixel p, $q_s(\mathbf{p})$ is the association of pixel p with superpixel s in $\boldsymbol{Q}$, $\mathcal{N}_\mathbf{p}$ is the number of superpixels around the pixel, $\boldsymbol{u}_\mathbf{s}$ represents the feature information, and $l_\mathbf{s}$ represents the position information.

To make the superpixels fit more closely to the edges of the object, a boundary perception loss was constructed, which was applied to a pixel-by-pixel embedding to improve the boundary accuracy. It includes two main components: the first makes pixels of the same category closer to the mean, i.e., closer to each other, and the second increases the variability between pixels belonging to different categories. The overall concept of the cross-entropy (CE) function was used. Finally, the loss function was constructed based on the CE of the semantic labels and position vectors, L2 reconstruction loss, and boundary-perceiving loss using the following equations:

$$\mathcal{L} = \sum_\mathrm{p} \mathrm{CE}(l_s'(\mathrm{p}), l_s(\mathrm{p})) + \alpha \|p - p'\|_2^2 + \beta \ell_\mathcal{B}, \tag{3}$$

$$\ell_\mathcal{B} = \frac{1}{|\mathcal{B}|} \sum_{B\in\mathcal{B}} \ell_B, \tag{4}$$

where $l_s(\mathrm{p})$ is the truth label, $l_s'(\mathrm{p})$ is the semantic label of the estimated association matrix $\boldsymbol{Q}$ reconstruction, B is the sampling block, l_B is the boundary perception loss, and α and β are the two weighting factors.

Because the AINet superpixel segmentation network is only applicable to 2D color texture features of images, it is difficult to handle 3D feature data, such as roll angle and direction. Therefore, a multi-task learning segmentation model, MTL-AINet, was designed based on texture, color, direction, and roll angle semantics. AINet was used as the basis function, and the mapping association map obtained from the projection of texture and color, direction, and roll angle features was input into the model for multi-task association learning. Therefore, multi-task learning considering the texture, color, direction, and roll angle features could be performed. The specific steps are as follows: for building facade information segmentation in different scenes, the dense point cloud model of the building is first reconstructed using the image sequence of the building according to the MVS algorithm [34], calculating the direction and roll angle semantics, dividing it into different viewpoints based on the overall structure and form to obtain the multi-view, and then projecting it onto the 2D image to obtain the pixel-level association of direction and roll angle maps and color and texture maps to be used as the input of the model. The fused pixel–superpixel association matrix (shown in Figure 5)

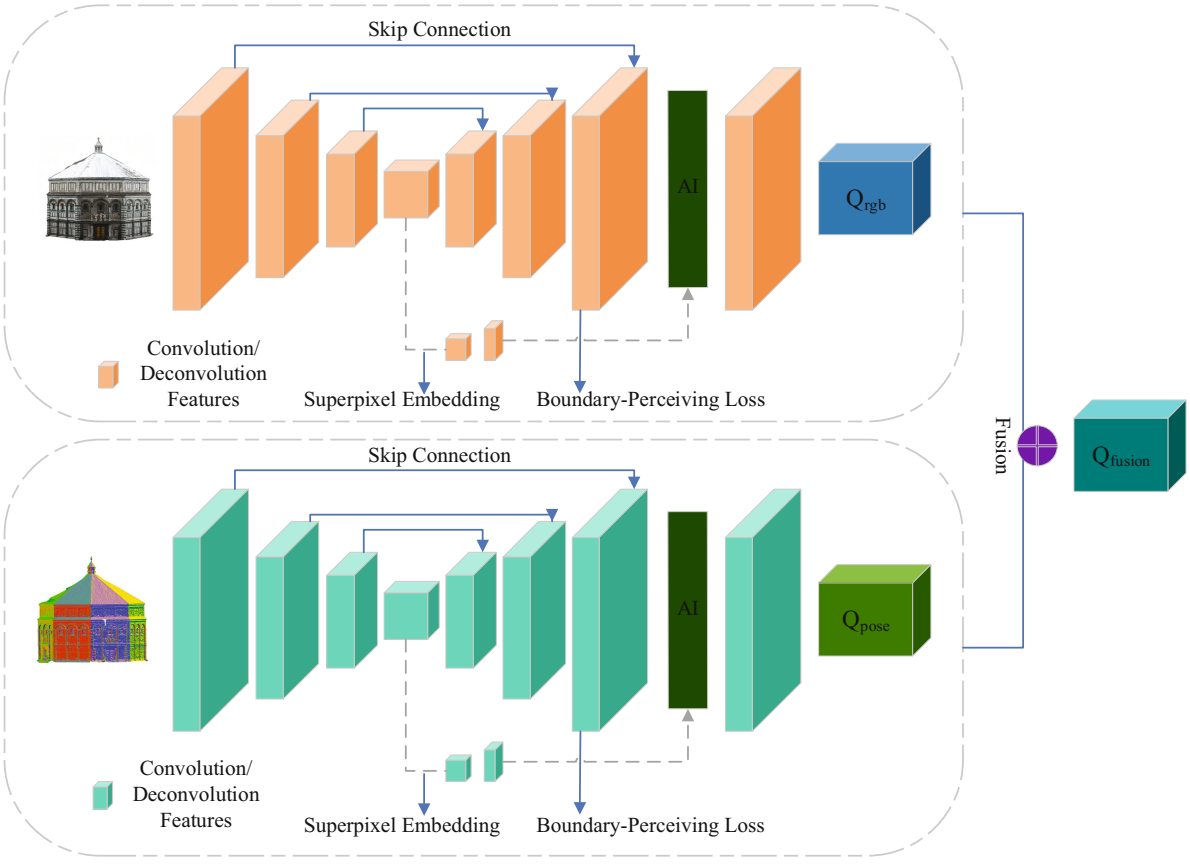


Figure 5: Structure of the MTL-AINet network.

is obtained via multi-task-integrated learning through the AINet network.

In the proposed algorithm, two sets of pixel embeddings corresponding to the semantic features of color and direction and roll angle, respectively, denoted as $E_{\text{rgb}} \in R_{\text{rgb}}^{H\times W\times D}$ and $E_{\text{pose}} \in R_{\text{pose}}^{H\times W\times D}$, are obtained through the deep neural network. The embeddings of pixel p, involving the color and direction and roll angle semantic features, can be denoted as $e_p^{\text{rgb}} \in R_{\text{rgb}}^{\text{D}}$ and $e_p^{\text{pose}} \in R_{\text{pose}}^{\text{D}}$, respectively. Let S be the sampling interval. The input image is compressed through multiple convolution and maximum pooling operations to generate two grid cell feature maps with multidimensional semantics, i.e., $M_{\text{rgb}} \in R_{\text{rgb}}^{h\times w\times D'}$ and $M_{\text{pose}} \in R_{\text{pose}}^{h\times w\times D'}$, where $h = H/S$, and $w = W/S$.

The feature maps $M_{\text{rgb}} \in R_{\text{rgb}}^{h\times w\times D'}$ and $M_{\text{pose}} \in R_{\text{pose}}^{h\times w\times D'}$ are transformed into new feature maps $\hat{M}_{\text{rgb}} \in R_{\text{rgb}}^{H\times W\times D}$ and $\hat{M}_{\text{pose}} \in R_{\text{pose}}^{H\times W\times D}$, respectively, through a 3 × 3 convolution. Thus, the embedding of the nine grid cells around pixel p is defined using equation (5), which directly associates a pixel with a semantic block:

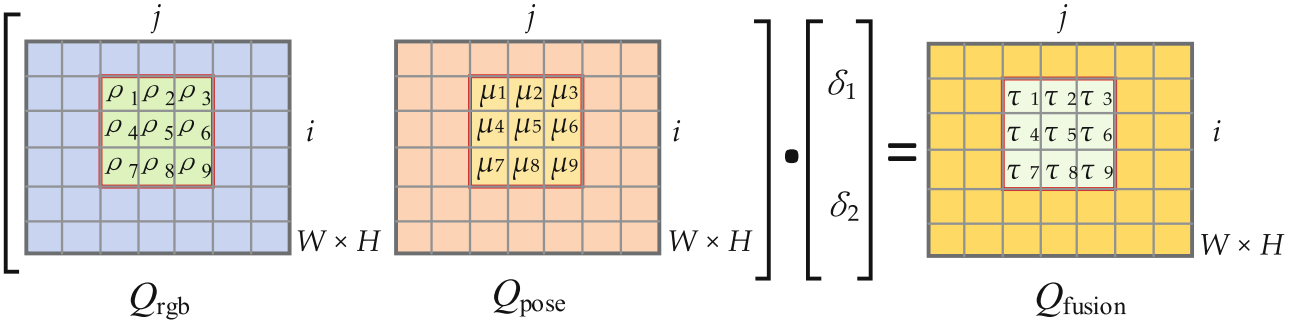


Figure 6: Detailed computation procedure of the soft association graph Q_{Fusion} for pixel $p(i, j)$.

$$\mathbf{SP} = \begin{bmatrix} \hat{\boldsymbol{m}}_{tl} & \hat{\boldsymbol{m}}_{t} & \hat{\boldsymbol{m}}_{tr} \\ \hat{\boldsymbol{m}}_{l} & \hat{\boldsymbol{m}}_{c} + \boldsymbol{e}_{p} & \hat{\boldsymbol{m}}_{r} \\ \hat{\boldsymbol{m}}_{bl} & \hat{\boldsymbol{m}}_{b} & \hat{\boldsymbol{m}}_{br} \end{bmatrix}, \tag{5}$$

where $\mathbf{SP} = [\mathrm{SP}_{\mathrm{rgb}} \quad \mathrm{SP}_{\mathrm{pose}}]$, and $\hat{\boldsymbol{m}}_{|\cdot,\cdot|} = \left[\hat{\boldsymbol{m}}_{|\cdot,\cdot|}^{\mathrm{rgb}} \quad \hat{\boldsymbol{m}}_{|\cdot,\cdot|}^{\mathrm{pose}}\right]$.

The association graph can be predicted through a 3 × 3 convolution and equation (6):

$$\begin{cases} \boldsymbol{e}_{\mathrm{p}}^{\mathbf{rgb}\prime} = \sum\limits_{ij} \mathbf{SP}_{\mathbf{rgb}}^{ij} \times \boldsymbol{\omega}_{ij} + \boldsymbol{b} \\ \boldsymbol{e}_{\mathrm{p}}^{\mathbf{pose}\prime} = \sum\limits_{ij} \mathbf{SP}_{\mathbf{pose}}^{ij} \times \boldsymbol{\omega}_{ij} + \boldsymbol{b} \end{cases}. \tag{6}$$

The proposed method uses the same loss function as the AINet superpixel segmentation method, including the three losses: CE loss, pixel reconstruction loss, and boundary-perceiving loss (as in equation (3)).

A new set of embedded pixels $\boldsymbol{E}' = \{\mathrm{E}'_{\mathrm{rgb}} \quad \mathrm{E}'_{\mathrm{pose}}\}$ can be computed using equations (5) and (6), which directly reflect the pixel–superpixel semantic association regarding the color and roll angle. In the proposed method, AINet is used as the base learner, and the multi-feature semantic association projection maps, including the color and texture feature map and the direction and the roll angle map, are used as multiple inputs for MTL-AINet. The soft association maps $\boldsymbol{Q}_{\mathrm{rgb}}$ and $\boldsymbol{Q}_{\mathrm{pose}}$ are obtained through MTL-AINet, which are, respectively, used to describe the probabilities each pixel belonging to its adjacent superpixels. Furthermore, $\boldsymbol{Q}_{\mathrm{rgb}}$ and $\boldsymbol{Q}_{\mathrm{pose}}$ are integrated to calculate the fusion association map $\boldsymbol{Q}_{\mathrm{Fusion}}$ by equation (7), which is the output considering multi-feature semantics of the MTL-AINet. Subsequently, a set of semantic blocks is extracted according to the soft association map $\boldsymbol{Q}_{\mathrm{Fusion}}$:

$$\boldsymbol{Q}_{\mathrm{Fusion}} = \delta_1 \boldsymbol{Q}_{\mathrm{rgb}} + \delta_2 \boldsymbol{Q}_{\mathrm{pose}}, \tag{7}$$

where δ_1 and δ_2 denote the weight factors of association maps $\boldsymbol{Q}_{\mathrm{rgb}}$ and $\boldsymbol{Q}_{\mathrm{pose}}$, respectively, satisfying $\delta_1 + \delta_2 = 1$.

Figure 6 shows the detailed computation process of the soft association graph $\boldsymbol{Q}_{\mathrm{Fusion}}$. In Figure 6, $\rho_1 \sim \rho_9$ represents a probability attribution distribution on color and texture of pixel $p(i, j)$, and similarly, $\mu_1 \sim \mu_9$ represents a probability attribution distribution of pixel p on the direction and roll angle. This pixel attribution reflects the similarity between pixel $p(i, j)$ and its nine adjacent grid cells in 2D and 3D features, respectively; τ_1–τ_9 represents the fusion probability distribution of pixel $p(i, j)$ on multi-feature semantics. The maximum of the nine multi-feature probabilities for each pixel yields a mapping of the label, which corresponds to the result of the semantic block segmentation. In summary, by calculating the feature maps of color and texture, and direction and roll angle, and using the AINET basis function to calculate their soft association mappings $\boldsymbol{Q}_{\mathrm{rgb}}$ and $\boldsymbol{Q}_{\mathrm{pose}}$, respectively, the optimal association matrix $\boldsymbol{Q}_{\mathrm{Fusion}}$ is obtained using a multi-task learning strategy, and the value with the highest probability is considered the final homogeneous semantic block segmentation result.

2.3 Building information extraction based on semantic clustering

On the basis of obtaining the initial segmentation results of each building facade, the lowest level of homogeneous regions is obtained. For recognition tasks with different requirements, a region adjacency graph based on 2D color and texture features and 3D direction and roll angle features is constructed. Region merging is considered as an approximation problem of the image, and the final clustering results are obtained using a stepwise iterative optimization method based on the nearest neighbor graph for the nodes with the smallest edge weights for region merging [35].

The texture characteristics of the image are measured using the joint probability distribution histogram of LBP (local binary pattern) and LC (local contrast) [36] by combining the structure and intensity of image texture. The *G*-statistic method is used to perform the analysis. Let x and y, respectively, represent a random variable set, and let g denote the probability density function, then the *G*-statistic formula is as follows [37]:

$$\begin{aligned} G(x,y) = {} & 2\sum_{x,y}\sum_{i=1}^{t} g_i \log g_i - 2\sum_{x,y}\left(\sum_{i=1}^{t} g_i\right)\log\left(\sum_{i=1}^{t} g_i\right) \\ & - 2\sum_{i=1}^{t}\left(\sum_{x,y} g_i\right)\log\left(\sum_{x,y} g_i\right) \\ & + 2\left(\sum_{x,y}\sum_{i=1}^{t} g_i\right)\log\left(\sum_{x,y}\sum_{i=1}^{t} g_i\right), \end{aligned} \tag{8}$$

where $G(x, y)$ represents the value of the pixel at position (x, y) on the image. g_i represents the number of pixels (pixel frequency) with gray level i in the image, and t represents the number of gray levels in the image. The results are highly capable of texture description and differentiation.

Buildings are geographic entities with certain shape information, and in the pixel merging process, it is often difficult to appropriately distinguish the contour edges of buildings using only texture information [38]; therefore, the geometric orientation constraint is adopted to obtain the object with a good edge fit. During merging, neighboring regions with longer common edges are prioritized to obtain more compact objects. The influence of the common edge length is introduced based on equation (9), and the geometric directional heterogeneity property is

defined as H, where ℓ^i_{xy} is the common edge length of neighboring regions x and y, and i is the influence factor of the common edge. When $i = 0$, $\ell^i_{xy} = 1$, i.e., the common edge has no influence on the regional heterogeneity, whereas when $i \neq 0$, the longer the common edge, the smaller the heterogeneity [39].

$$H(x, y) = \frac{G(x, y)}{\ell^i_{xy}}, \tag{9}$$

In the region adjacency graph- and the nearest neighbor graph-based area merging, the problem of minimizing the adjacent-area approximation error is gradually transformed into that of calculating the area with the smallest merging cost. The merging cost is the weight of the edge of an adjacent area with a common edge and is defined as:

$$Q(x, y) = \frac{S_x S_y}{(S_x + S_y)} \cdot H(x, y), \tag{10}$$

where $Q(x, y)$ denotes the combined generational values of regions x and y; S_x and S_y denote the areas of x and y, respectively; and $H(x, y)$ denotes the heterogeneous properties of the two regions.

Using the strategy of building element extraction based on the aforementioned clustering strategy, we combined feature maps with hierarchical clustering to extract building elements in different scenes.

3 Experiments and results

Real buildings were selected as experimental objects, and datasets comprised of RGB samples and 3D semantic samples first need to be made for model training. The RGB and 3D semantic datasets are divided equally and with a one-to-one match. The training dataset contains 500 samples, and the testing dataset contains 200 samples. Semantic segmentation and hierarchical clustering on different facades of buildings in different scenarios are performed. In order to validate the accuracy and completeness of the proposed MTL-AINet method, three traditional superpixel segmentation methods, namely, SLIC-based, LSC-based, and AINet-based methods, are compared and analyzed with the proposed method. Moreover, superpixel evaluation metrics [40,41], including boundary recall (REC), undersegmentation error (UE), achievable segmentation accuracy (ASA), compactness (CO), and intra-cluster variation (ICV), are used to evaluate the reliability and robustness of the experimental segmentation methods for each building facade.

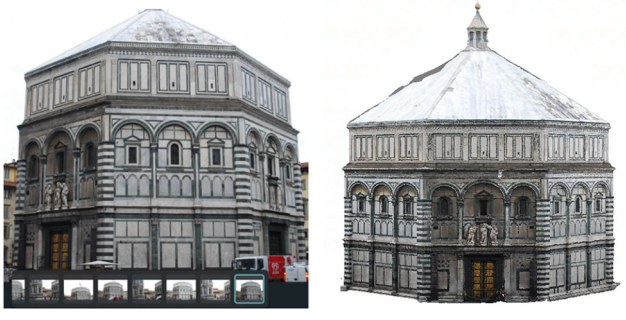
3.1 Segmentation extraction for building facades with similar texture

The Florentine Cathedral multi-view sequence image dataset (Figure 7(a)) was used in the experiment. A total of 105 images were selected with a resolution of 1,296 × 1,936 and a focal length of 29 mm. The sparse point cloud and camera parameters were first obtained using the structure from motion (SfM) technique, and the dense point cloud was then obtained using the MVS method. In total, 39.01 million points were obtained in the experiment, and the dense reconstruction results are shown in Figure 7(b).

Based on the geometric orientation of the building, the color and texture, and direction and roll angle projections were selected from the left, front, and right views of the building, and directional feature maps were obtained. Figure 8(a–c) shows the color and texture projections of the left, front, and right views of the building, respectively. Figure 8(d–f) shows the direction and roll angle projections of the left, front, and right views of the building, respectively. The figures show that the color and texture and the direction and roll angle projections express different details of the building facades.

Figures 9–11, respectively, show the segmentation results of the four segmentation methods under different viewing angles. From regions marked with the two red rectangles, it can be seen that the proposed MTL-AINet method performs better on those regions with similar color and texture but different façade orientations than the other three traditional methods; different facades were not be segmented successfully with the three traditional methods, while the proposed method can effectively distinguish different facades with similar textures, which facilitates the clustering task.

The clustering results of each facade shown in Figures 12–14 indicate that the AINet method can easily identify the roofs as the same facade, which cannot be separated by the AINet



(a) (b)

Figure 7: Building image models with similar texture features: (a) single image, (b) dense point cloud.

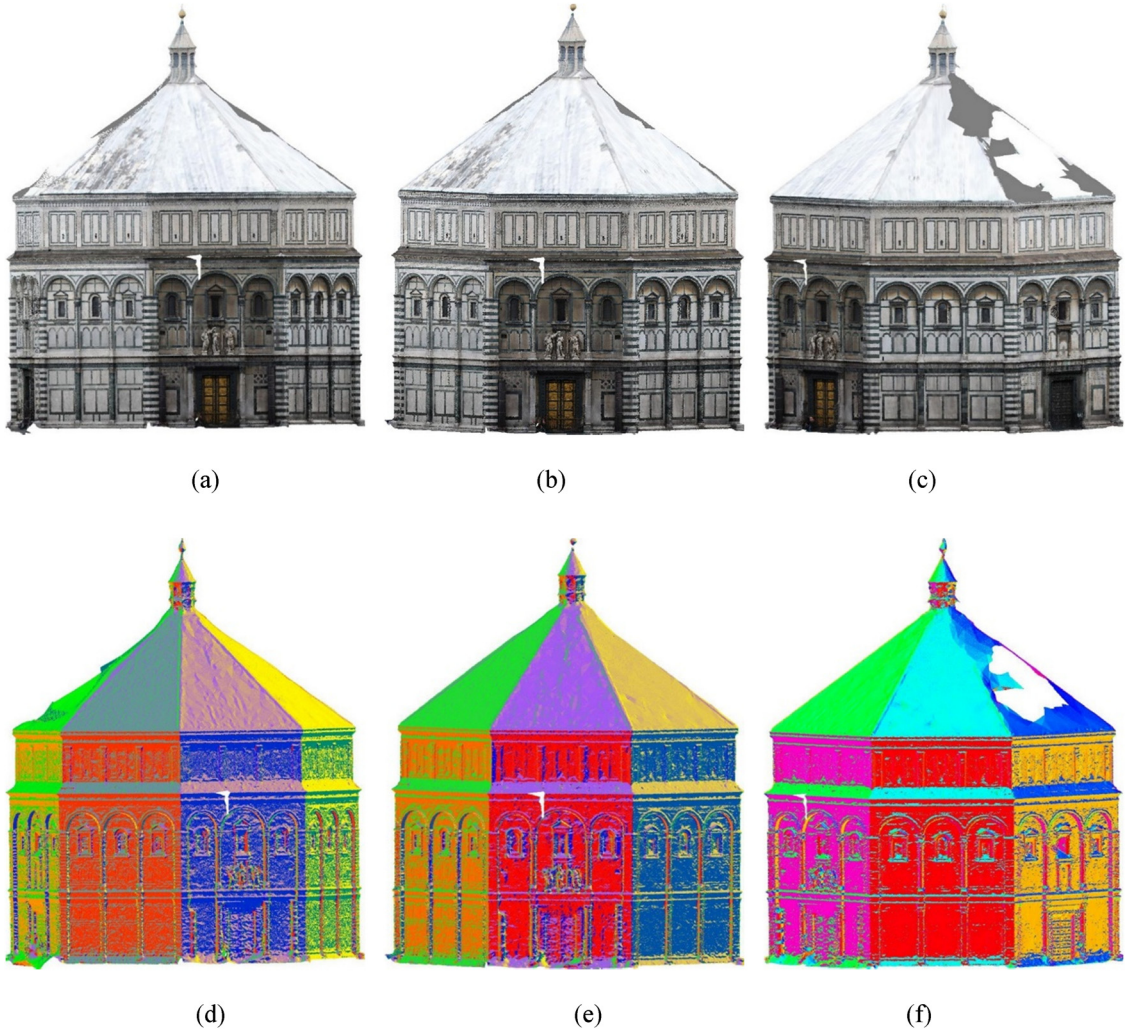


Figure 8: Different view images of buildings and the corresponding directional semantic features: (a and d) left-view, (b and e) front-view, (c and f) right-view.

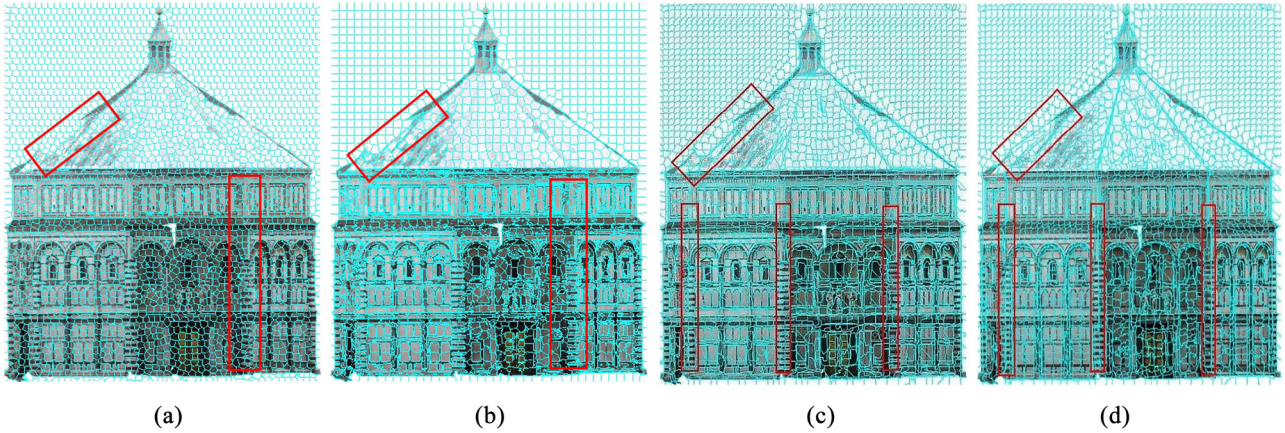


Figure 9: Left-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

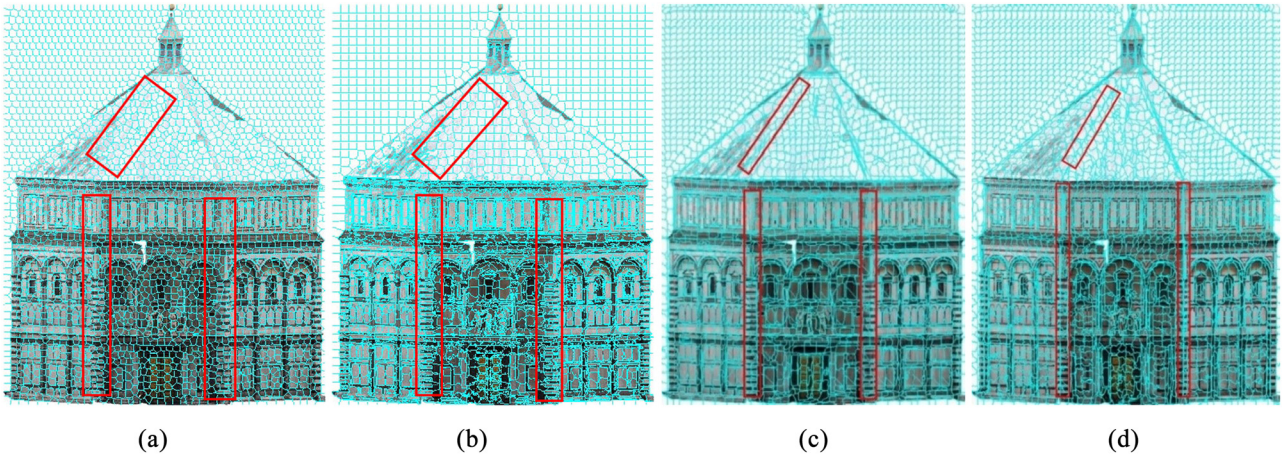


Figure 10: Front-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

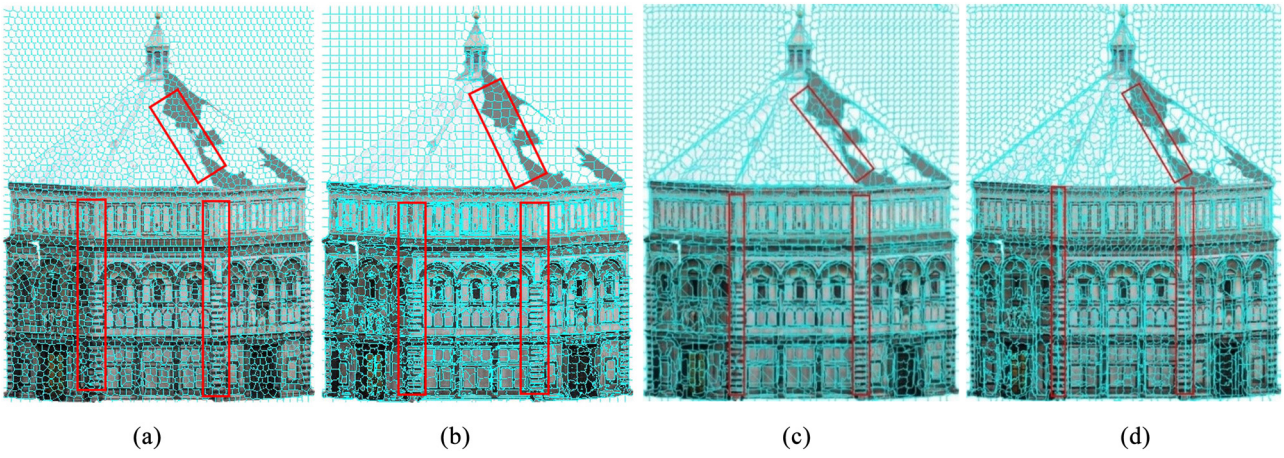


Figure 11: Right-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

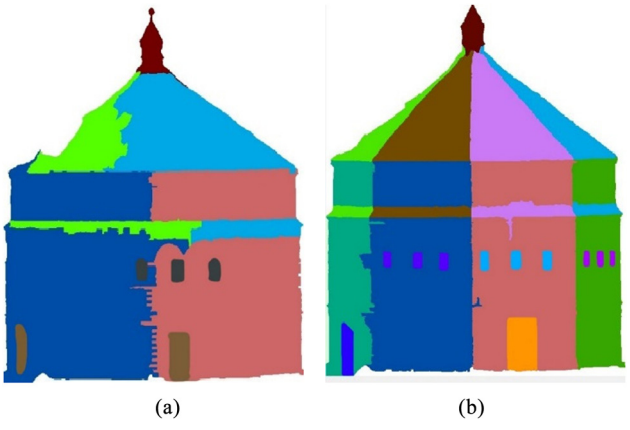


Figure 12: Left-view clustering result maps: (a) AINet and (b) MTL-AINet.

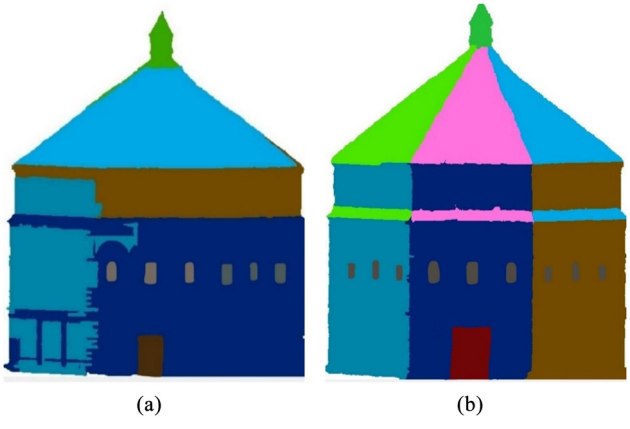


Figure 13: Front-view clustering result maps: (a) AINet and (b) MTLAINet.

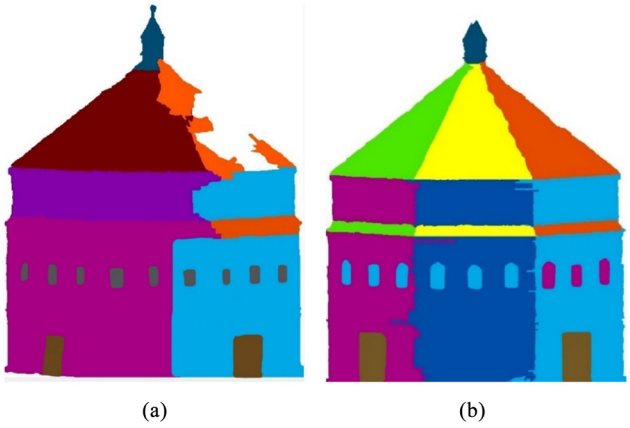


Figure 14: Right-view clustering result maps: (a) AINet and (b) MTL-AINet.

method due to the overly similar color texture. The proposed MTL-AINet method can effectively utilize the directional features of each facade, distinguish different facades with similar textures (such as the roofs and edges of each facade), and can accurately extract the detailed structure of each building. Thus, the details of the facade structure can be extracted accurately. Therefore, the MTL-AINet method is robust for buildings with similar color and texture features but large facade differences.

The experimental results show that for buildings with different facades but very similar texture features, the segmentation method relying only on single-texture semantics is prone to misclassification, and it is difficult to extract sufficient detail information. The segmentation method based on color and texture and direction and roll angle semantics considerably overcomes this limitation and can efficiently extract fine details of each building facade.

3.2 Segmentation extraction for complex buildings with occlusions and shadows

The multi-view sequence image dataset of Örebro Castle (Figure 15a) was used in this experiment. A total of 136 images were selected with an image resolution of 1,936 × 1,296 and a focal length of 29 mm. The sparse point cloud and camera parameters were first obtained through SfM, and the dense point cloud was then obtained through the MVS method. A total of 77.72 million points were obtained in the experiment, and the dense point cloud reconstruction result is shown in Figure 15(b).

The building was divided into four views (front, back, right, and left) according to its architectural characteristics, and a directional semantic feature map was constructed, as shown in Figure 16. In this experiment, the four views of the building were selected for the color and texture and direction and roll angle projections, and directional feature maps were obtained. Figure 16(a–d) shows the color and texture projections of the front, right, back, and left views of the building, respectively. Figure 16(e–h) shows the direction and roll angle projections of the front, right, back, and left views of the building, respectively. From the figures, it is evident that the texture and color and the direction and roll angle projection maps can express different details of the building facades. In particular, the direction and roll angle projection maps can significantly reduce the influence of shadows.

Because all the facades of the building are affected by shadows and occlusions to different degrees, the three traditional superpixel segmentation methods based on texture semantics can only segment the areas with obvious color–texture features and does not sufficiently distinguish



Figure 15: Complex building image model with occlusion and shadow: (a) single image and (b) dense point cloud reconstruction result.

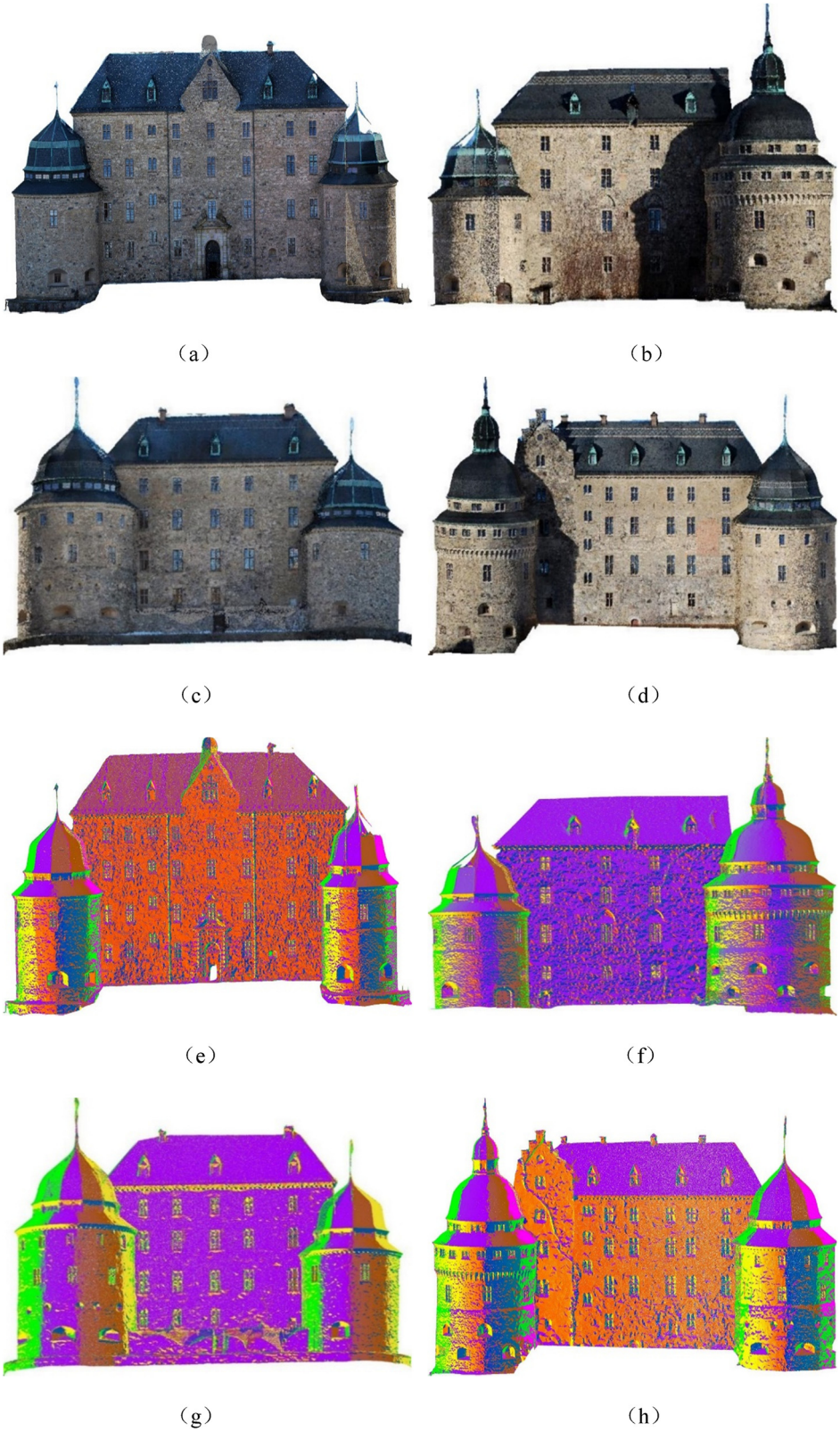


Figure 16: Different view images of buildings and the corresponding directional semantic features: (a) the front view image, (b) the right view image, (c) the back view image, (d) the left view image, (e) the front view semantic features, (f) the right view semantic features, (g) the back view semantic features, and (h) the left view semantic features.

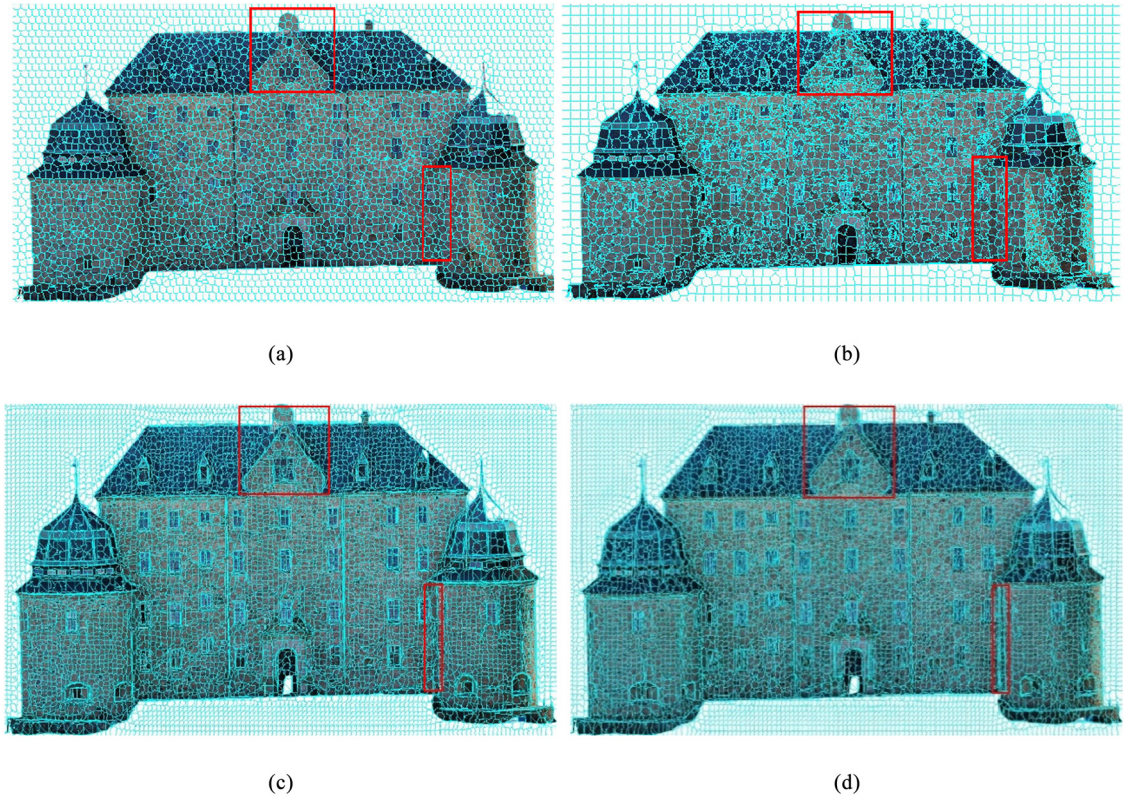


Figure 17: Front-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

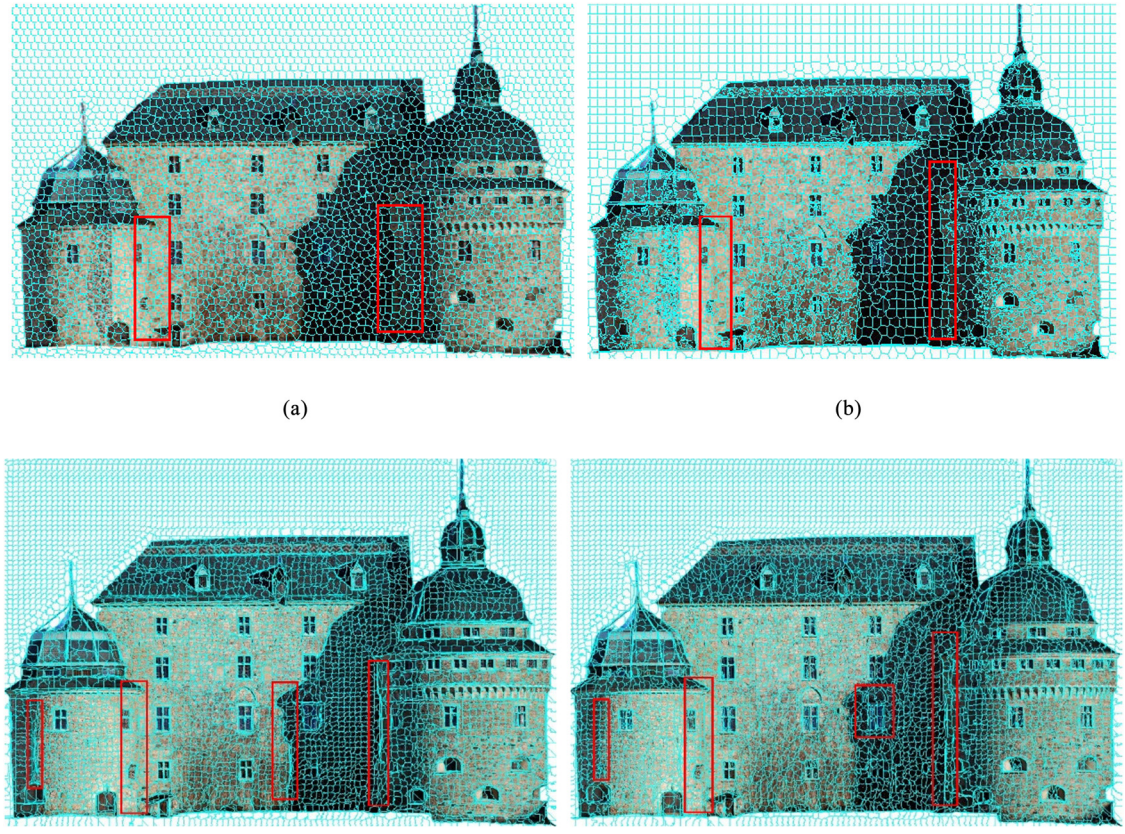


Figure 18: Right-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

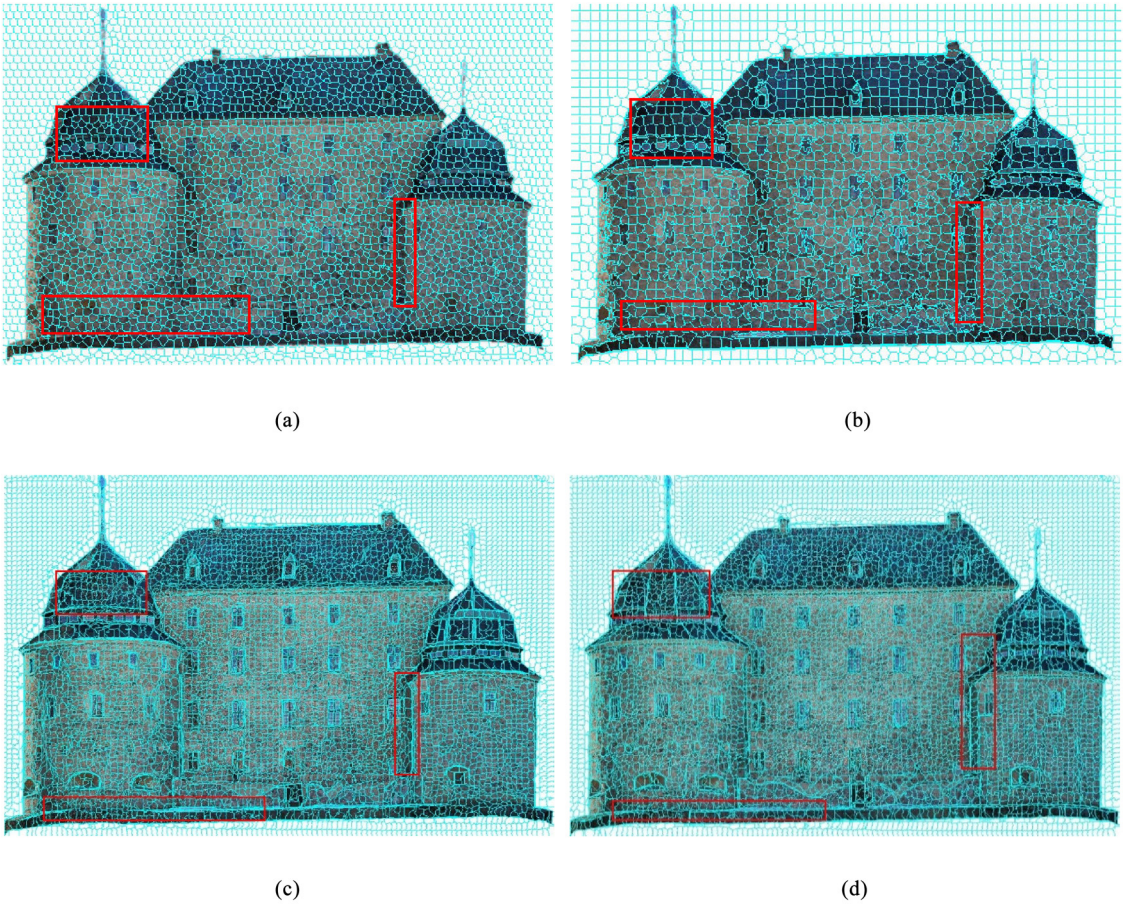


Figure 19: Back-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, (d) MTL-AINet.

between the shadowed parts and occlusions, such as the boundaries of the column-shaped buildings in Figures 17(a–c) and 18(a–c) and the roof and bottom fences in Figures 19(a–c) and 20(a–c). The proposed MTL-AINet method compensates for these shortcomings by weakening the influence of shadows and occlusions and improving the segmentation accuracy; for example, the top protrusion in Figure 17(d) is not mixed with the roof, the shadows and edges in Figures 18(d) and 20(d) are not divided separately, and the bottom fence in Figure 19(d) is not divided with the background but into a homogeneous surface.

From Figures 21(a) and 22(a), it is evident that the clustering result maps based on the superpixel segmentation results generated by AINet erroneously divides the shadows into a separate facade. Moreover, because the shaded parts present similar texture features with other facades, they can be easily mixed into the same facade, as shown in Figures 23(a) and 24(a), which considerably impacts the extraction of building elements. In contrast, the proposed MTL-AINet method can effectively avoid effects of occlusions and shadows, and improve the accuracy of facade structure extraction, as shown in Figures 21(b), 22(b), 23(b) and 24(b), where the large areas of shadows and shaded areas are efficiently distinguished and extracted. This significantly weakens the influence of shadows and facilitates subsequent tasks.

The experimental results show that because of shadows and occlusions, building facades cannot be segmented accurately based only on texture. However, a segmentation method based on color and texture, and direction and roll angle semantics can efficiently extract detailed elements of building facades.

3.3 Segmentation extraction for old buildings with inconsistent textures

The Martenstroget multi-view sequence image dataset (Figure 25(a)) was used in this experiment. Twelve images with an image resolution of 2,592 × 3,872 and a focal length of 30 mm were selected. The sparse point cloud

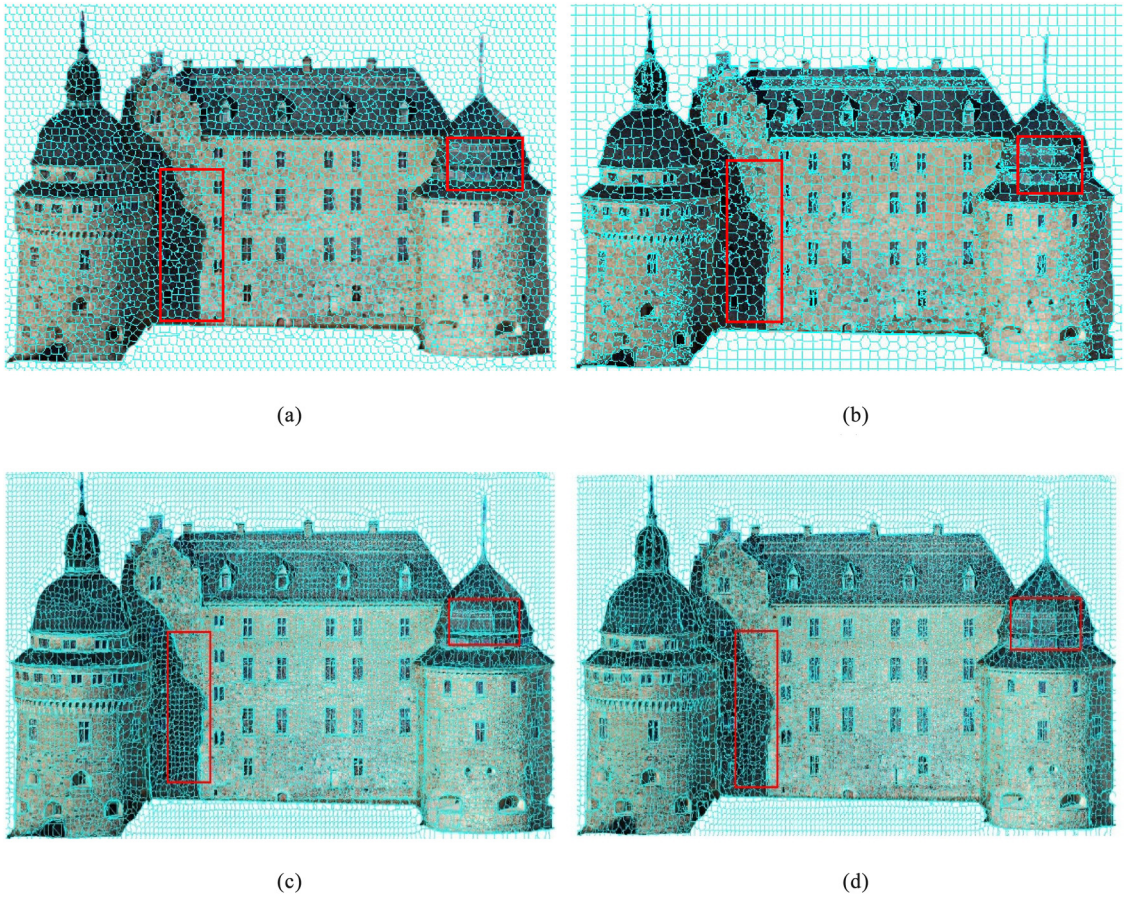


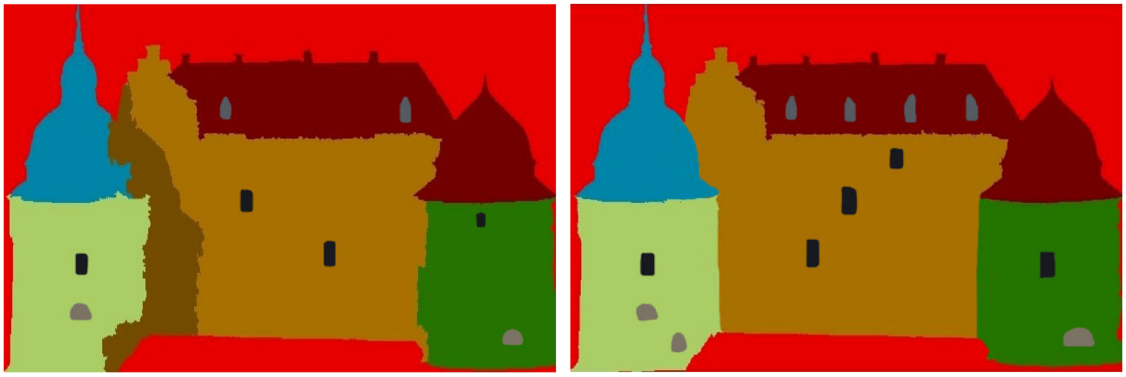
Figure 20: Left-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, (d) MTL-AINet.

and camera parameters were first obtained through SfM, and the dense point cloud was then obtained through the MVS method. A total of 23.2 million points were obtained in the experiment, and the dense point cloud reconstruction results are shown in Figure 25(b).

The building was divided into three views according to its architectural characteristics and a directional semantic feature map was constructed, as shown in Figure 26. In this experiment, three views of the building – front, right, and left – were selected for the color and texture and direction



Figure 21: Right-view clustering result maps: (a) AINet and (b) MTL-AINet.



(a) (b)

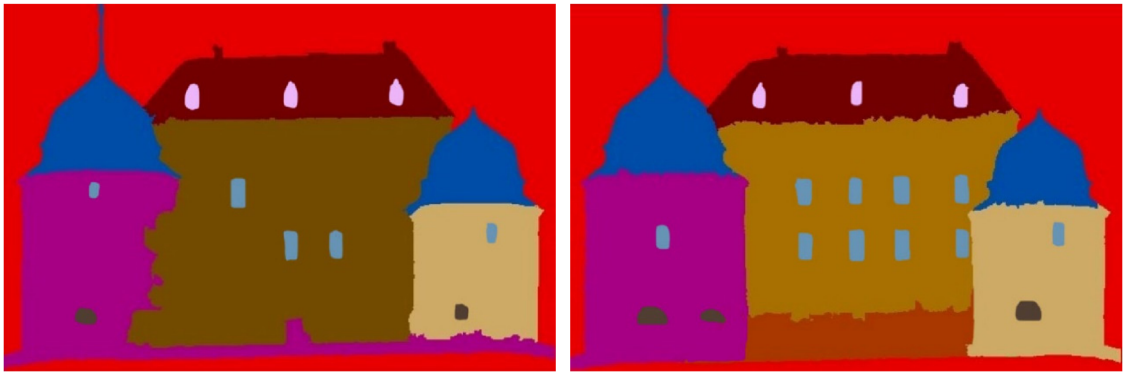
Figure 22: Left-view clustering result maps: (a) AINet, (b) MTL-AINet.



(a) (b)

Figure 23: Front-view clustering result maps: (a) AINet and (b) MTL-AINet.

and roll angle projections, and directional feature maps were obtained. Figure 26(a–c) shows the color and texture projections of the front, right, and left views of the building, respectively. Figure 26(d–f) shows the direction and roll angle projections of the front, right, and left views of the building, respectively. From the figures, it is evident that the texture and color and direction and roll angle projection maps can reveal protrusions and depressions.



(a) (b)

Figure 24: Back-view clustering result maps: (a) AINet, (b) MTL-AINet.

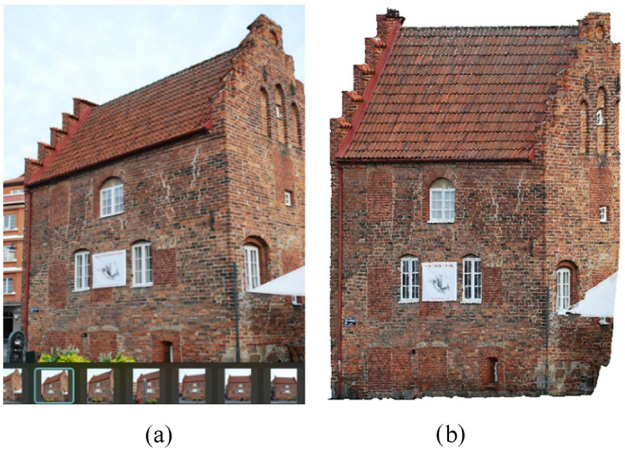


Figure 25: Image sequence and point cloud model of an old building: (a) single image, (b) dense point cloud reconstruction result.

The color texture of the old building surface interferes with the recognition of building detail elements, and it is obvious from Figures 27(a–c) and 28(a–c) that the three traditional superpixel segmentation methods based on texture semantics mistakenly segment walls that have similar color textures but do not belong to the same facade into the same blocks. From Figures 28(a–c) and 29(a–c), it can be seen that the three traditional methods will mistakenly divide the regions belonging to the same facade but with dissimilar color textures into different categories. As seen from Figures 27(d), 28(d) and 29(d), the proposed

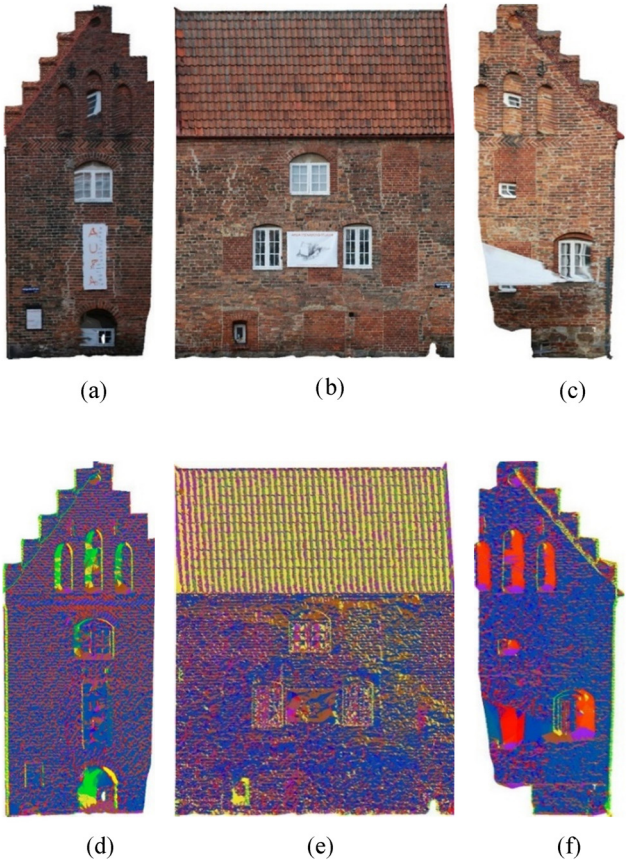


Figure 26: Different view images of buildings and the corresponding directional semantic features: (a) the left view image, (b) the front view image, (c) the right view image, (d) the left view semantic features, (e) the front view semantic features, and (f) the right view semantic features.

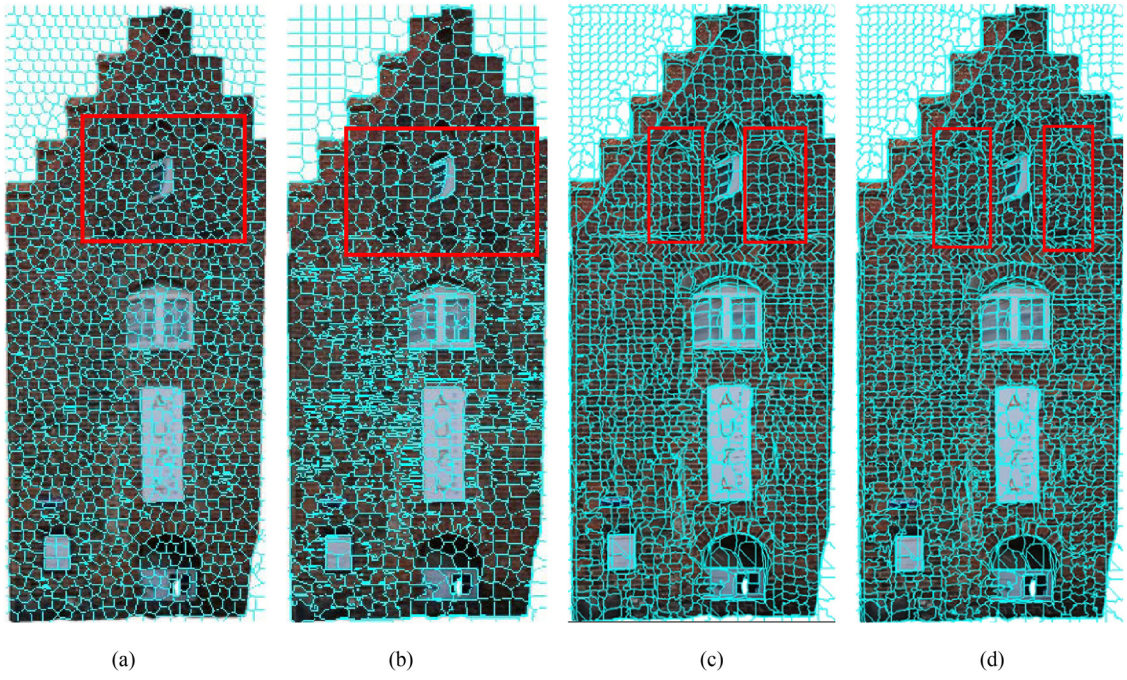


Figure 27: Left-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

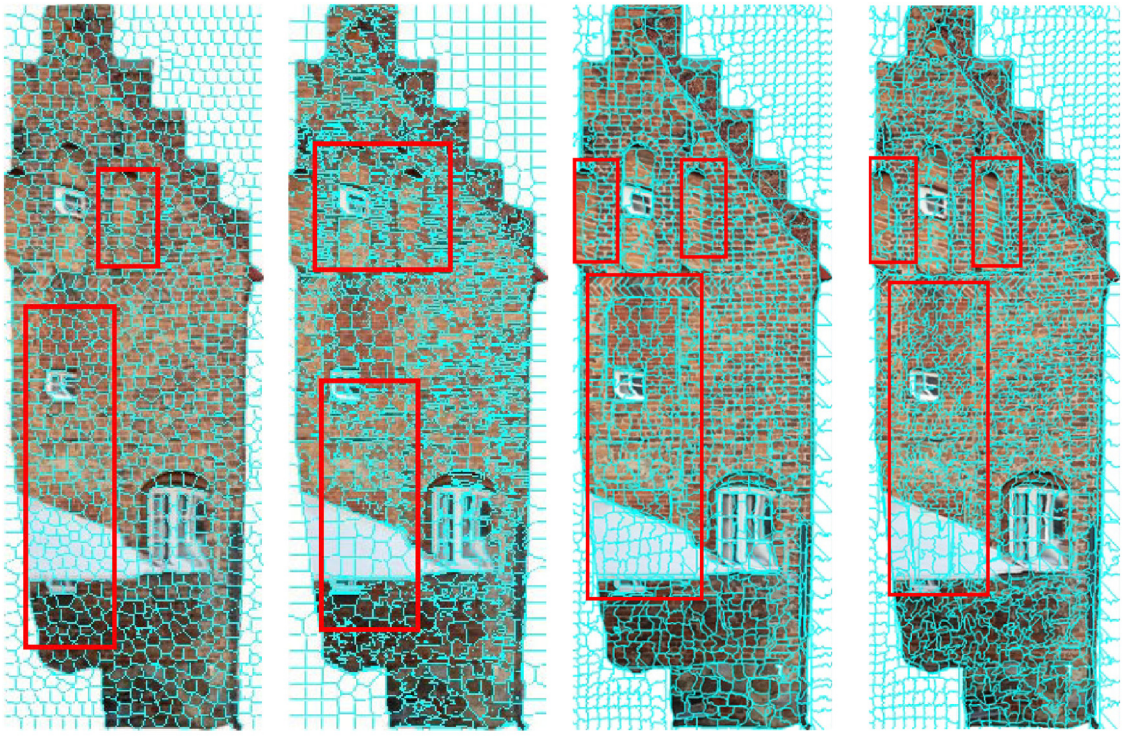


Figure 28: Right-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

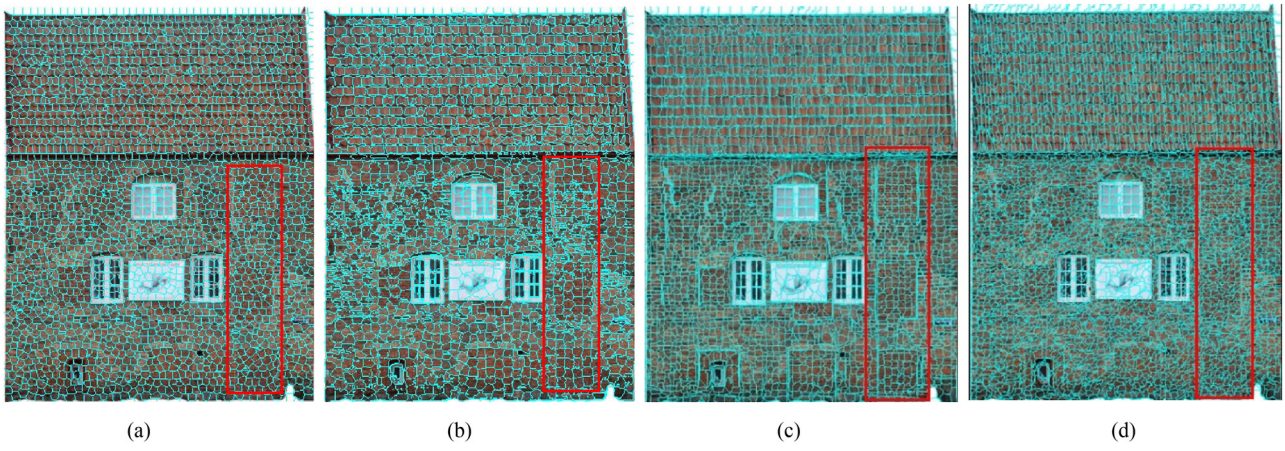


Figure 29: Front-view semantic segmentation result maps: (a) SLIC, (b) LSC, (c) AINet, and (d) MTL-AINet.

MTL-AINet method effectively overcomes this limitation by dividing the regions that belong to the same facade but are not similar in color and texture into the same similar semantic facets, and distinguishing the regions that do not belong to the same facade but are more similar in color and texture, which provides a better basic data for the subsequent task of hierarchical clustering.

From the clustering results shown in Figures 30 and 31(a), it can be concluded that the traditional clustering method can only extract areas with more evident color textures and performs more false extractions. In contrast, the proposed method incorporates direction and roll angle features and can achieve a finer extraction of detailed information on old buildings, such as the protrusions and depressions, as shown in Figures 30 and 31(b).

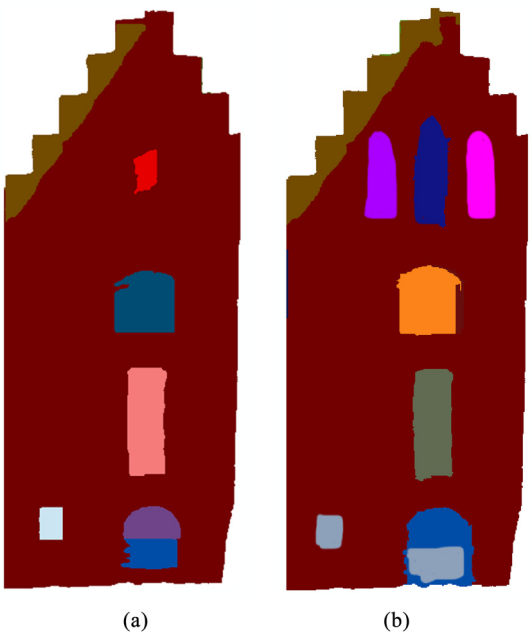


Figure 30: Left-view clustering result maps: (a) AINet and (b) MTL-AINet.

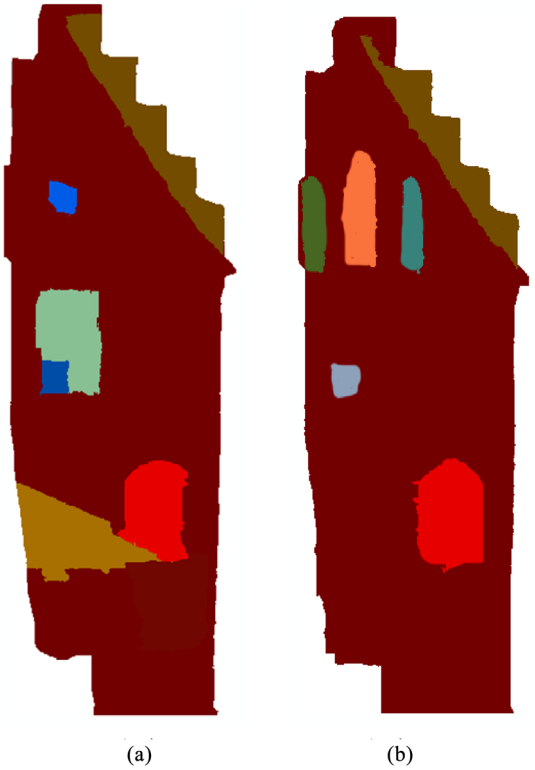


Figure 31: Right-view clustering result maps: (a) AINet and (b) MTLAINet.

Figure 28 shows the segmentation and clustering results of the front view of the building, wherein it is evident that the surface textures of the old buildings have large differences. Figure 28(a–c) shows the texture-based segmentation results of the three traditional methods, and it can be seen that segment regions belonging to the same facade but with variations in brightness and colors were wrongly segmented into separate facets. Better segmentation results are obtained by the proposed MTL-AINet method because it considers the direction and roll angle semantics, as shown in Figure 28(d). From Figure 32(a) and (b), it is evident that the clustering results based on AINet superpixel segmentation still cannot distinguish the edge details of old buildings very well. However, the proposed MTL-AINet performs better because it considered both direction and roll angle constraints.

Thus, the proposed MTL-AINet method obtains a higher accuracy and more accurately extracts details in depressions and protrusions. The experimental results show that for building facades with inconsistent color and texture dilapidation and similar textures, but with geometric depressions on the facade, the MTL-AINet superpixel segmentation method based on color and texture and direction and roll angle semantics has better extraction accuracy.

3.4 Evaluation and discussion

To verify the reliability of the proposed method, the results of building facade segmentation obtained in the experiment were evaluated using the super-pixel evaluation metrics REC, UE, CO, ASA, and ICV, and compared with

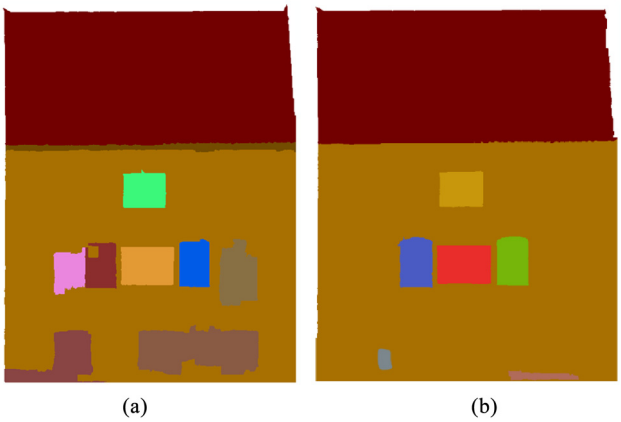


Figure 32: Front-view clustering result maps: (a) AINet and (b) MTL-AINet.

Table 1: Quantitative evaluation of segmentation results of building facades with similar textures

		UE	REC	ASA	CO	ICV
Front	SLIC	0.379947	0.889174	0.620053	0.001069	3.93211
	LSC	0.440799	0.938966	0.559201	0.001074	28.0515
	AINet	0.021018	0.970707	0.978982	0.334899	34.5196
	MTL-AINet	**0.019271**	**0.943104**	**0.980729**	**0.357829**	**29.3193**
Right	SLIC	0.404075	0.892174	0.595925	0.001069	4.13724
	LSC	0.466343	0.943412	0.533657	0.001080	28.3137
	AINet	0.027305	0.963469	0.972695	0.342162	34.0619
	MTL-AINet	**0.017819**	**0.964551**	**0.982181**	**0.358969**	**29.3033**
Left	SLIC	0.419562	0.888362	0.580438	0.001051	4.01667
	LSC	0.461331	0.967186	0.538669	0.001065	28.0491
	AINet	0.021658	0.970996	0.978342	0.336289	35.9970
	MTL-AINet	**0.016265**	**0.988503**	**0.983735**	**0.363060**	**29.6183**

The bold values refer to the evaluation results for the MTL-AINet method proposed in this research.

three traditional methods based on SLIC, LSC, and AINet. The quantitative evaluation of building facade segmentation results across different scenes is shown in Tables 1–3.

As indicated by the results presented in Tables 1–3, in the detail segmentation experiments of each building facade in different scenes and clustering for different recognition extraction tasks, the indicators based on the proposed MTL-AINet method have been greatly improved in comparison with the other three traditional methods. From the quantitative comparative evaluation results, it can be concluded that AINet and MTL-AINet have greatly improved the CO and under-segmentation of different building facades compared with traditional SLIC and LSC. Comparing with AINet, the average BR and UE rates of 11 building facades segmentation results from MTL-AINet were, respectively, improved by approximately 26 and 3%. Building facade segmentation results from MTL-AINet achieve the highest average REC rate and ASA. For MTL-AINet, as it considered multiple semantic features for superpixel segmentation, the ICV has a relatively moderate increase in comparison of AINet, but obviously better than SLIC and LSC methods. In conclusion, compared with the three traditional methods, the proposed MTL-AINet method has achieved better results, which reflects the superiority.

According to these experimental results, compared with the single semantic features that cannot adequately consider the detailed features of different facades of buildings, there are many problems of over- or under-segmentation. The MTL-AINet method proposed in this study integrated color and texture and direction and roll angle semantics of buildings, which can achieve an effective and

Table 2: Quantitative evaluation of segmentation results of complex facades with occlusions and shadings

		UE	REC	ASA	CO	ICV
Front	SLIC	0.236471	0.830849	0.763529	0.000707	10.3795
	LSC	0.257331	0.953757	0.742669	0.000724	24.1789
	AINet	0.023097	0.967408	0.976903	0.339097	41.4140
	MTL-AINet	**0.013739**	**0.993024**	**0.986261**	**0.346239**	**41.2862**
Back	SLIC	0.300599	0.839113	0.699401	0.000721	9.75155
	LSC	0.329494	0.958646	0.670506	0.000737	28.2231
	AINet	0.019018	0.997625	0.980982	0.337187	52.2021
	MTL-AINet	**0.014404**	**0.998875**	**0.985596**	**0.333284**	**46.3717**
Left	SLIC	0.290049	0.831543	0.709951	0.709951	9.61799
	LSC	0.29349	0.953549	0.70651	0.000728	23.3771
	AINet	0.014446	0.998225	0.985554	0.333083	46.3417
	MTL-AINet	**0.010683**	**0.999846**	**0.989317**	**0.348731**	**41.6964**
Right	SLIC	0.272069	0.832898	0.727931	0.000724	8.72767
	LSC	0.304864	0.975086	0.695136	0.000740	27.2945
	AINet	0.014009	0.999728	0.985991	0.343048	44.7818
	MTL-AINet	**0.012982**	**0.996969**	**0.987018**	**0.342476**	**41.3360**

The bold values refer to the evaluation results for the MTL-AINet method proposed in this research.

Table 3: Quantitative evaluation of segmentation results of old buildings with inconsistent textures

		UE	REC	ASA	CO	ICV
Front	SLIC	0.236471	0.830849	0.763529	0.000707	10.3795
	LSC	0.257331	0.953757	0.742669	0.000724	24.1789
	AINet	0.023097	0.967408	0.976903	0.339097	41.4140
	MTL-AINet	**0.013739**	**0.993024**	**0.986261**	**0.346239**	**41.2862**
Back	SLIC	0.300599	0.839113	0.699401	0.000721	9.75155
	LSC	0.329494	0.958646	0.670506	0.000737	28.2231
	AINet	0.019018	0.997625	0.980982	0.337187	52.2021
	MTL-AINet	**0.014404**	**0.998875**	**0.985596**	**0.333284**	**46.3717**
Left	SLIC	0.290049	0.831543	0.709951	0.709951	9.61799
	LSC	0.29349	0.953549	0.70651	0.000728	23.3771
	AINet	0.014446	0.998225	0.985554	0.333083	46.3417
	MTL-AINet	**0.010683**	**0.999846**	**0.989317**	**0.348731**	**41.6964**
Right	SLIC	0.272069	0.832898	0.727931	0.000724	8.72767
	LSC	0.304864	0.975086	0.695136	0.000740	27.2945
	AINet	0.014009	0.999728	0.985991	0.343048	44.7818
	MTL-AINet	**0.012982**	**0.996969**	**0.987018**	**0.342476**	**41.3360**

The bold values refer to the evaluation results for the MTL-AINet method proposed in this research.

accurate segmentation of target objects. In particular, for building objects with similar colors and textures but different directions and angles, building objects with severe shadow and occlusion effects, and old and new aging building objects with different colors and textures, the proposed method can accurately extract building detail elements with higher robustness.

4 Conclusion

This study proposed the MTL-AINet algorithm considering color and texture as well as direction and roll angle semantics to extract a detailed information on building facades. First, a dense point cloud model of the building is generated using multi-view images, and then 3D direction and roll angle features are computed. On this basis, these features are projected onto a 2D plane to generate color and texture as well as direction and roll angle feature maps. Subsequently, multi-task learning is used to obtain the fusion soft association map of multi-semantic features of each facade, according to which each facade can be divided into a series of semantic blocks. Finally, the detailed elements of the building are extracted using a semantic hierarchical clustering method. In this study, three kinds of building facades with similar colors and textures, with shadows and occlusions, and with inconsistent distribution of textures and dilapidated were selected as research objects for experiments. The experimental results show that the proposed MTL-AINet method has achieved the best results in REC, UE, CO, ASA, and the ICV and is also superior to SLIC and LSC methods. Therefore, the MTL-AINet has higher reliability and robustness, and its superpixel segmentation results provide a more accurate data basis for further cluster extraction.

For regions with similar or scarce textures, shadows, and occlusions, the proposed method considered multiple feature semantics including direction and roll angle, so it can not only improve the accuracy of 2D segmentation, but also meet different 3D clustering tasks and requirements. Besides, compared with those methods based on 3D point cloud, this study projected 3D information onto a 2D plane for building elements extraction, which can significantly improve the segmentation efficiency. The proposed method provides an idea and data support for the extraction of building details in smart city construction. The multimodal features of superpixels in this study refer to multiple 2D and 3D features. Future work will focus on exploring the intrinsic correlation of multimodal superpixels based on deep neural networks to achieve multimodal superpixel clustering automatically, quickly, and accurately.

Acknowledgments: The authors would like to thank the reviewers and editors for valuable comments and suggestions. The authors would also like to acknowledge Editage (www.editage.com) for English language editing.

Funding information: This research was funded by Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China (Grant No. KLSMNR-G202213, KLSMNR-G202214), the National Natural Science Foundation of China (Grant No.

41901401, 42271482, 42101070), the China Postdoctoral Science Foundation (Grant No. 2021M691653), the Natural Science Foundation of Jiangsu Province (Grant No. BK20190743), and the Knowledge Innovation Program of Wuhan-Shuguang Project (Grant No. 2022010801020284).

Author contributions: Conceptualization: R.Z., G.L., and L.L.; data curation: M.J.; methodology: R.Z., X.Y., G.L., S.S., and L.L.; validation: M.J., S.S., and Y.H.; writing – original draft: R.Z.; writing – review & editing: R.Z., M.J., G.L., X.Y., S.S., Y.H., and L.L. All authors have read and agreed to the published version of this manuscript.

Conflict of interest: The authors state no conflict of interest.

References

[1] Abdul MH, Ghulam MB. A survey on instance segmentation: state of the art. Int J Multimed Inf Retr. 2020;9(3):171–89.
[2] Su H, Maji S, Kalogerakis E. Multi-view convolutional neural networks for 3d shape recognition. Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 945–53.
[3] Mostajabi M, Yadollahpour P, Shakhnarovich G. Feedforward semantic segmentation with zoom-out features. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3376–85.
[4] Chen LC, Yang Y, Wang J. Attention to scale: Scale-aware semantic image segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 3640–9.
[5] Yin L, Ji X, Wu D. A building extraction method based on semantic segmentation and efficient conditional random fields optimization. Remote Sens. 2018;10(5):788.
[6] Wang R, Du Q, Tao J, Yuan Z, Li T. Semantic segmentation of high-resolution remote sensing images based on joint feature learning and graph cut. Remote Sens. 2019;11(18):2152.
[7] Meng X, Liu Y, Zhang YD. A region-based convolutional neural network for building extraction from remote sensing images. Remote Sens. 2018;10(2):189.
[8] Huang T, Shengyong Y, Zhiqiang Z, Hongyun L. Model analysis of intelligent data mining based on semantic segmentation technology. Proceedings of the 2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering; 2015.
[9] Zheng C, Zhang Y, Wang L. Multilayer semantic segmentation of remote-sensing imagery using a hybrid object-based Markov random field model. Int J Remote Sens. 2016;37(23):5505–32.
[10] Jampani V, Sun D, Liu MY. Superpixel sampling networks. Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 352–68.
[11] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3431–40.
[12] Badrinarayanan V, Kendall AC, Ipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95.
[13] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2015. p. 234–41.
[14] Feng Y, You H, Zhang Z. Hypergraph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33(1):3558–65.
[15] Te G, Hu W, Zheng A. RGCNN: Regularized graph CNN for point cloud segmentation. Proceedings of the 26th ACM International Conference on Multimedia; 2018. p. 746–54.
[16] Li R, Wang S, Zhu F. Adaptive graph convolutional neural networks. Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32, 2018.
[17] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017;1(2):4.
[18] Li Z, Zhong Y, Yang B. Building extraction from airborne LiDAR data using local structural similarity matching. ISPRS J Photogramm Remote Sens. 2020;161:120–33.
[19] Liu Y, Huang X, Zhang L, Qiao Y. Extraction of buildings from LiDAR data with a rectangle model. ISPRS J Photogramm Remote Sens. 2015;101:89–98.
[20] Maturana D, Scherer S. Voxnet: A 3d convolutional neural network for real-time object recognition. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE; 2015. p. 922–8
[21] Klokov R, Lempitsky V. Escape from cells: Deep Kd-networks for the recognition of 3D point cloud models. 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 863–72. doi: 10.1109/ICCV.2017.99.
[22] Riegler G, Ulusoy AO, Geiger A. OctNet: Learning deep 3D representations at high resolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 6620–9. doi: 10.1109/CVPR.2017.701.
[23] Zhang Y, Rabbat M. A graph-cnn for 3d point cloud classification. 2018 IEEE International Conference on Acoustics, Speech Signal Processing (ICASSP), IEEE; 2018. p. 6279–83.
[24] Qi CR, Su H, Mo K. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 652–60.
[25] Jiao Y, Wang W, Li S. Image semantic segmentation fusion of edge detection and AFF attention mechanism. Appl Sci. 2022;12:11248.
[26] Khan MZ, Gajendran MK, Lee Y, Khan MA. Deep neural architectures for medical image semantic segmentation: review. IEEE ACCESS. 2021;9:83002–24.
[27] Giraud R, Ta VT, Papadakis N. Robust superpixels using color and contour features along linear path. Comput Vis Image Underst. 2018;170:1–13.
[28] Giraud R, Ta VT, Papadakis N. Texture-aware superpixel segmentation. 2019 IEEE International Conference on Image Processing (ICIP), IEEE; 2019. p. 1465–9.
[29] Haris K, Efstratiadis SN, Maglaveras N, Katsaggelos AK. Hybrid image segmentation using watersheds and fast region merging. IEEE Trans Image Process. 1998;7(12):1684–99. doi: 10.1109/83.730380.
[30] Yang F, Sun Q, Jin H. Superpixel segmentation with fully convolutional networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 13964–73.

[31] Guo Y, Liu Y, Georgiou T, Lew MS. A review of semantic segmentation using deep neural networks. Int J Multimed Inf Retr. 2018;7:87–93.
[32] Gao S, Li ZY, Yang M, Cheng M, Han J, Torr P. Large-scale unsupervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence; 2022.
[33] Wang Y, Wei Y, Qian X. AINet: Association implantation for superpixel segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 7078–87.
[34] Hu Z, Wu Z, Zhang Q, Fan Q, Xu J. A spatially-constrained color–-texture model for hierarchical VHR image segmentation. IEEE Geosci Remote Sens Lett. 2013;10(1):120–4. doi: 10.1109/LGRS.2012.2194693.
[35] Yao Y, Luo Z, Li S, Fang T, Quan L. MVSNet: Depth Inference for Unstructured Multi-view Stereo," Computer Vision; 2018.
[36] Wang J, Luan Z, Yu Z. Superpixel segmentation with attention convolution neural network. 2021 International Conference on Image, Video Processing, and Artificial Intelligence. Vol. 12076. SPIE; 2021. p. 74–9.
[37] Wu ZC, Hu ZW, Zhang Q, Cui WH. Remote sensing image segmentation method combining spectral, texture and shape structural information. J Surveying Mapp. 2013;1:44–50 (in Chinese).
[38] Bai X, Wang C, Tian Z. Self-adaptive superpixels based on neural network models. IEEE Access. 2020;8:137254–62.
[39] Gaur U, Manjunath BS. Superpixel embedding network. IEEE Trans Image Process. 2019;29:3199–212.
[40] Achanta R, Susstrunk S. Superpixels and polygons using simple non-iterative clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 4651–60.
[41] L Chen, L Shao, Q Bai, J Yang, S Jiang, Y Miao. Review of image classification algorithms based on convolutional neural networks. Remote Sens. 2021;13(22):4712.