Information Fusion 111 (2024) 102498

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

End-to-end multiview fusion for building mapping from aerial images

Qi Chen [a,b], Wenxiang Gan [a], Pengjie Tao [c,d,*], Penglei Zhang [a,b], Rongyong Huang [e], Lei Wang [f]

[a] School of Geography and Information Engineering, China University of Geosciences (Wuhan), China
[b] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[c] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
[d] Hubei Luojia Laboratory, Wuhan, China
[e] Guangxi Laboratory on the Study of Coral Reefs in the South China Sea, School of Marine Sciences, Guangxi University, China
[f] Huawei Cloud Computing Technologies Co., Ltd., China

ARTICLE INFO

Keywords:
Multiview fusion
Building mapping
Deep learning
Aerial imagery
Photogrammetry

ABSTRACT

In the domain of photogrammetry, the fusion of information from multiple views holds the potential to significantly enhance the accuracy and robustness of building mapping. While multiview observation and stereoscopic imaging form the bedrock of photogrammetric projects, current deep learning methodologies predominantly focus on orthophotos and digital surface models (DSMs), often sidelining the rich multiview information inherent in original images. Addressing this gap, we present Multiview Mapper (MVMapper), an end-to-end learning framework explicitly crafted to harness and fuse the rich semantic information from original multiview images with object-space features. MVMapper uses stereo labels for supervised building segmentation in a dual-space, encompassing both image and object domains. Additionally, it incorporates a novel piecewise affine projection method for ensuring a robust image-to-object feature transformation. Experimental results on an aerial photogrammetric dataset with a resolution of 30 cm demonstrate MVMapper's superiority over state-of-the-art multiview data fusion methods, yielding significant improvements in segmentation and contour accuracy. Notably, the proposed piecewise affine projection method mitigates misalignment issues caused by DSM noise, enabling the effective fusion of multiview features with object-space features. Further experimentation on a separate open-source dataset demonstrates MVMapper's substantial advantages in transferability to other contexts.

1. Introduction

Multiview feature fusion has emerged as a pivotal concept across various research domains, including medical image analysis [1], food image recognition [2], and 3D shape understanding [3], underscoring its significance in enhancing model performance and understanding complex data structures. In the realm of photogrammetry and remote sensing, multiview observation and stereoscopic imaging stand as fundamental capabilities. When planning an aerial or satellite photogrammetric project, it typically necessitates ensuring adequate overlap and a specific intersection angle between adjacent images [4]. This configuration establishes a robust multiview observation network, enabling the acquisition of 3D information from the survey area, thus facilitating the production of digital surface models (DSMs), digital elevation models (DEMs), and digital orthophoto maps (DOMs) [5].

Building mapping is a classic research problem in the field of photogrammetry and plays a vital role in applications such as urban planning and disaster management [6]. Most existing approaches for image-based building mapping primarily rely on DOM or true DOM (TDOM) [7–9]. However, it is crucial to recognize that these DOM or TDOM signals are usually derived from multiview images through photogrammetric production. Relying solely on single-view information for building mapping overlooks the valuable insights gained from redundant observations. This raises an intriguing question: can we harness the surplus observational information present in the original multiview images, readily available throughout the photogrammetric workflow, to enhance learning-based models and achieve higher performance for building recognition and segmentation?

While trained human operators have traditionally been generating 3D vector representations of buildings through interpreting stereoscopic image pairs, there is currently a dearth of deep learning models capable of emulating this process. Many learning-based studies have explored using DSMs as auxiliary information for building extraction from DOMs [10–12], making it an easily implementable approach

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.
E-mail addresses: chenqi@cug.edu.cn (Q. Chen), gwx@cug.edu.cn (W. Gan), pjtao@whu.edu.cn (P. Tao), plzhang@cug.edu.cn (P. Zhang), hry@whu.edu.cn (R. Huang), Leiwang4@huawei.com (L. Wang).

https://doi.org/10.1016/j.inffus.2024.102498
Received 31 October 2023; Received in revised form 23 April 2024; Accepted 24 May 2024
Available online 1 June 2024

in typical photogrammetric projects. However, despite DSM elevation being derived from dense multiview matching, it cannot fully represent the texture information captured by the original multiview images from different perspectives. Moreover, considering the errors and noise inherent in DSM production, it can be argued that this approach utilizes limited multiview information with uncertainties for building mapping. Other studies have further attempted to construct a multitask architecture for simultaneously predicting disparity information and semantic segmentation from binocular images [13,14]. However, during segmentation, these approaches remain focused on individual images and do not explore the explicit fusion and utilization of multiview semantics.

The motivation of this study is driven by the goal to achieve effective fusion of semantic features from multiview images for building mapping, which gives rise to two key questions: (i) how can we construct a learnable framework to extract and integrate semantic features from different viewpoints, and (ii) considering the limitations in DSM quality, particularly the presence of noise at building edges, how can we robustly project image-space features to the object space for generating mapping results?

In addressing the above questions, this paper presents Multiview Mapper (MVMapper), an end-to-end learning framework that integrates semantic features from multiview original images with the object-space building segmentation workflow to enhance recognition capability and segmentation accuracy. The framework takes DOM, DSM, and the complete set of original images from a photogrammetric project as inputs, utilizing stereo labels within the training data to construct a dual-space (encompassing both image and object spaces) supervised building segmentation pipeline. Additionally, we propose a piecewise affine projection scheme to robustly transfer the multiview building semantics to the object space. This method estimates an affine transformation for each building instance based on DSM points and their correspondences in image space, facilitating effective fusion of multiview features with object-space features for building segmentation.

In our experiments on an aerial photogrammetric dataset created specifically for this study, with a resolution of 30 cm, covering 15,741 building polygons, we evaluated the performance of MVMapper for building segmentation on the test set. The results demonstrated that the proposed framework outperforms two state-of-the-art multiview data fusion strategies. Compared to scenarios using only TDOM and DSM as input, our framework shows a considerable improvement, with Intersection over Union (IoU) increasing from 83.8% to 86.1% and contour-based IoU increasing from 55.2% to 60.0%. Notably, the application of our proposed piecewise affine projection scheme played a crucial role in observing this positive feedback. When using the traditional collinearity equations for image-to-object feature map projection, the multiview image features were unable to contribute to improving building segmentation accuracy. Additionally, we conducted a small-scale transferability validation experiment on an open-source aerial imagery dataset located in Dortmund city center [15]. The results demonstrate that MVMapper consistently achieves significant accuracy improvements over comparative methods.

The main contributions of our work are summarized as follows:

- We propose MVMapper, an end-to-end feature fusion framework that integrates image-space multiview semantic information with TDOM/DSM features, enabling a more comprehensive utilization of the photogrammetric project's original observation data.
- We propose a piecewise affine projection scheme for robust image-to-object transformation of multiview building features, which achieves better feature fusion and significantly improves the building mapping accuracy compared to traditional projection schemes based on collinearity equations.
- We incorporate 3D annotations to automatically generate ground truths for the multiview original images and optimize the training efficiency of MVMapper by precomputing piecewise mapping tables.

The paper is structured as follows. Section 2 reviews related works. Section 3 introduces the dataset, including data preparation and sample generation. Section 4 presents our MVMapper's methodology. Section 5 discusses the experimental results and provides an in-depth analysis of the framework's performance. Section 6 draws the conclusion.

2. Related work

In general, deep learning techniques have now become the prevailing direction for building mapping from remote sensing images. Convolutional neural networks (CNNs) [16,17] and self-attention-based networks [18,19] have demonstrated remarkable performance in addressing building segmentation tasks. Furthermore, there is an emerging trend to conceptualize the mapping process as a vector generation challenge, placing emphasis on delineating accurate and regularized vector boundaries [20–22]. However, we take a different focus by enhancing existing building segmentation frameworks through the incorporation of multiview data. To synthesize relevant literature, we primarily employed a snowballing strategy, starting from key publications in building mapping and multiview-based image interpretation, and in this section, we review relevant research that explores techniques for modeling multiview features and aligning them to optimize the utilization of multiview observations.

2.1. Multiview feature modeling

A common approach to leveraging multiview observations for building mapping has involved generating DSMs through dense matching, which were then combined with DOM for prediction in a neural network [23,24]. A similar strategy was adopted by Yu et al. [10], with the key difference being their exploration of generating the DSM using an end-to-end depth prediction network. However, a shared limitation among these methods was the absence of a direct link between the multiview information and the segmentation task. In such cases, the segmentation learning framework operated on the DSM, which inherently comprised secondary, compressed, or potentially distorted observation information.

Another line of research explored the synergy between binocular disparity prediction and semantic segmentation tasks. These studies established multi-task frameworks using stereo image pairs, yielding positive results for both tasks [14,25,26]. However, they did not address the issue of effectively integrating the segmentation results from the binocular image space into the object space.

Lu et al. [27] attempted to align all images using ground control points and conduct multiview feature modeling to enhance the nitrogen nutrition status estimation in winter wheat. Nevertheless, the same technique might struggle to mitigate the tilt effect in building images; moreover, even the effort to rectify multiview images using DSM could yield distorted or compromised images due to inherent DSM inaccuracies. Similarly, Huang et al. [28] extracted features separately from multiview images and then merged them to feed a neural network, achieving success in image classification. However, this study is also not applicable to building mapping as they bypassed the need to address spatial alignment challenges for semantics extracted from diverse views.

Certain studies have already addressed the advantages of utilizing multiview original images for recognition tasks compared to solely relying on DOM. They improved recognition accuracy by projecting multiview interpretations from image space to object space and then employing a voting mechanism [29,30]. For similar reasons, Kurz et al. [31] applied CNNs for synchronized road markings extraction in both the object and image spaces, subsequently validating and optimizing the results by projecting object-space outcomes to the image space. However, despite their efforts to integrate multiview information, these studies still rely on handcrafted design for multiview feature fusion and do not achieve end-to-end learning.

Table 1
Comparison between our approach and the state-of-the-art multiview fusion strategies for building segmentation.

Approaches	Input of prediction model		Dual-space end-to-end training
	Object-space Products	Image-space Multiviews	
Object-space Fusion [10,11,23]	✓		
Multiview Voting [29,30]	✓	✓	
Ours	✓	✓	✓

Multiview representation learning is a well-explored area in machine learning and data mining, commonly focused on utilizing multimodal measurements to describe the same object [32,33]. The target is typically to enhance the effectiveness of features in various tasks, such as recognition [34], clustering [35], and classification [36], by harnessing redundant information and employing feature fusion. In the context of photogrammetry, "multiview" specifically refers to images captured from different perspectives. Recent applications of end-to-end multiview feature modeling have shown success. For instance, Robert et al. [37] endeavored to project semantically encoded features from multiple images onto a 3D point cloud, resulting in aggregated multiview features that enhance 3D semantic segmentation accuracy; Qi et al. [38] optimized a neural field to capture color and 3D structure from multiview images and then integrated 2D features with 3D priors to enhance semantic segmentation accuracy. However, it is noteworthy that, within the domain of building mapping, there remains a notable absence of end-to-end multiview fusion strategies tailored to the unique characteristics of this field.

In summary, the state-of-the-art works closely related to our research can be classified into two categories, namely Object-space Fusion and Multiview Voting methods, as outlined in Table 1. Object-space Fusion methods typically do not directly input multiview original images into the prediction model. Instead, they use the object-space products generated by the multiview data, such as DOM and DSM, as inputs. On the other hand, Multiview Voting methods explicitly incorporate image-space multiview data alongside their object-space products in the prediction model, segmentation results from different views are projected onto the object space, and the final result is determined by a voting mechanism. The core distinction of our approach from these two lies in not only simultaneously utilizing dual-space data as inputs to the prediction model but also establishing an end-to-end learning framework by bridging these two spaces.

2.2. Feature warping and image transformation

To effectively utilize multiview data for building mapping, it is essential to accurately project multiview image features onto the object space with precise spatial alignment. Deep learning-based multiview stereo frameworks commonly employs homography to project features from target images to reference views for feature modeling [39,40]. However, these methods primarily leverage disparity-based loss functions and depth prediction for establishing correspondence between images; they typically do not explicitly tackle the direct alignment of multiview feature spaces.

Alternatively, collinearity equations offer a theoretically robust point-to-point mapping from object space to image space [41]. Yet, precise collinearity-based projection heavily relies on accurate elevation data from DSM. Unfortunately, building point clouds reconstructed through photogrammetry may exhibit detail omissions and errors, posing challenges in achieving meticulous alignment when projecting multiview features onto the object space. Moreover, rational function models, known for their high-precision fitting to strict geometric standards, can also be utilized for image-to-object feature warping [42]. Nonetheless, they face similar difficulties in mitigating the influence of DSM noises.

One simplified strategy involves neglecting elevation and modeling image-to-object projection as a 2D-to −2D geometric transformation between two images. This approach allows for the adaptation of relevant techniques from the field of image registration. Since globally applying rigid, affine, or polynomial transformations can be challenging to achieve precise alignment due to complex distortions and relief displacement effects, existing studies have demonstrated that employing a piecewise fitting approach and linear estimation with corresponding points can achieve high-precision image registration [43,44]. Furthermore, other research has shown that the piecewise fitting method, when assisted by a sufficient density of correspondences, can effectively alleviate noise interference by employing outlier removal methods [45,46]. While these promising outcomes have established a crucial foundation for the proficient projection and learnable fusion of multiview features, there is currently a lack of research exploring piecewise 2D-to −2D transformation strategies to address DSM noises when projecting image-based features onto the object space.

3. Dataset

3.1. Data acquisition and study area

We constructed an aerial photogrammetry dataset that includes both the original observational data in image space and the geospatial data products in object space. The image-space data consists of original images with resolution of approximately 0.3 m and dimensions of 17,310 × 11,310 pixels. These images were captured using a UCX Camera with a focal length of 70.5 millimeters at an average altitude of 3500 m. The overlap percentages within and between flight strips are 65% and 30%, respectively. For data processing, we leveraged ContextCapture software to execute aerial triangulation and dense matching on the original images. This yielded DSM and TDOM products, each with a ground sample distance of 0.3 m. In addition, we employed TerraScan software to extract ground points from the DSM point cloud and generate a DEM for further use.

As shown in Fig. 1, the study area in Xuzhou, China, encompasses various building types, including commercial areas, low-rise houses, high-rise residential zones, and large industrial areas, providing a diverse range of building types. We extracted 8 original images (Fig. 1b) that intersect with this region for our experiments; to facilitate the integration of RGB and depth information for building segmentation in the image space, we projected the DSM product onto the image coordinate system of these 8 images, creating image-space elevation rasters, which we term *projected DSMs*. The entire study area was divided into 30 subregions, of which 24 were selected for training and validation, while the remaining 6 were designated for testing. By employing skilled personnel wearing stereoscopic glasses to annotate the building outlines in stereo image pairs, we obtained a total of 15,741 3D polygons as ground truth data.

3.2. Label generation and patch grouping

The object-space label maps can be easily generated by rasterizing the 3D building polygons with their planar coordinates. In the image space, we projected the 3D polygons onto the multiview original images for label generation to ensure annotation consistency while minimizing manual efforts. However, projecting the polygons onto the image space without considering occlusion in the original images can result in errors.

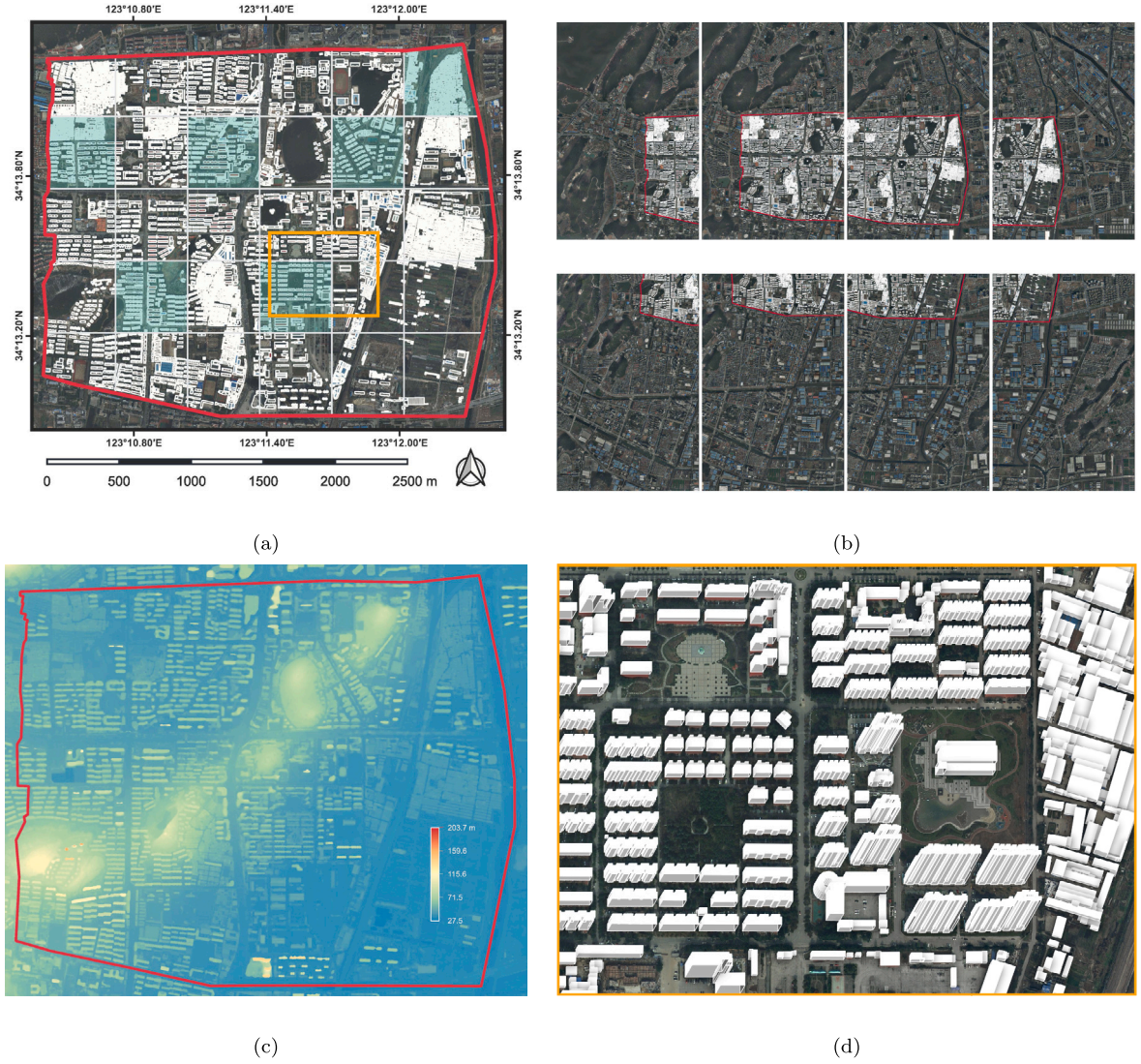


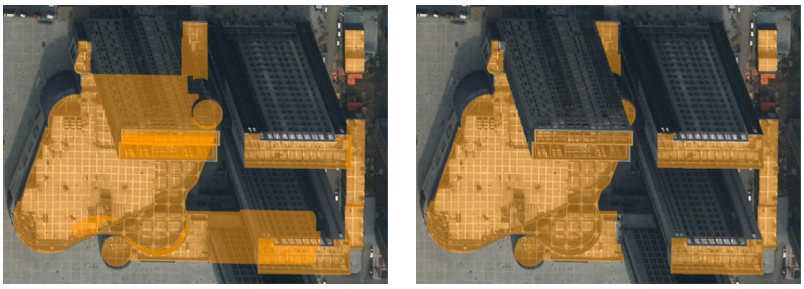
Fig. 1. Overview of the multiview aerial image dataset. In (a), the red polygon delineates the entire area of interest (AOI), within which white vectors represent annotated buildings. The gridlines divide the entire area into 30 subregions, with 6 shaded in light green designated for testing, while the remaining regions are allocated for training and validation. (b) displays 8 original images intersecting with the AOI and the projected building ground truths within the AOI. (c) illustrates the DSM generated by dense matching. (d) presents building annotations from 3D perspectives within the orange box outlined in (a).

To address this, we implemented the following occlusion-aware labeling approach: first, for each 3D polygon, we projected its vertices onto the image space to obtain the roof outline; simultaneously, we replaced the height value of each vertex with DEM elevation and projected the vertices onto image to derive the footprint outline of the building; by connecting the corresponding contour lines between the roof and footprint, we determined the positions of the building's walls; through interpolation we estimated the elevation of each pixel of the walls; after completing the roof/footprint projection, wall extraction and elevation estimation for all buildings, we performed elevation comparisons in areas with multiple pixel coverage to detect occlusions. As illustrated in Fig. 2, this labeling method significantly reduces projection errors in roof labels when applied.

We performed grid-based division in object space to define patches for input into the model. We adopted a dual-space synchronous input approach to meet MVMapper's requirements. This involved matching each object-space patch with corresponding patches in image space. Specifically, we projected the four vertices of the object-space patch onto the multiview images using DSM elevation. However, as the projected vertices in image space did not form fixed-size rectangles, we cropped the image-space multiview images, along with their label maps, using the bounding rectangle of the vertices and resized the crops to match the size of the object-space patch. This treatment allowed us to sort the patches into groups to align with MVMapper's specifications. Each patch group included $1 + N$ image crops, where 1 represents the object-space patch, and N is a variable number representing the count of patches with the same coverage as the object-space patch found in image space. By applying a patch size of 512×512 pixels, Fig. 3 shows a sample with 3 imaging views and provides the statistics on the number of views in all patch groups within the entire dataset.

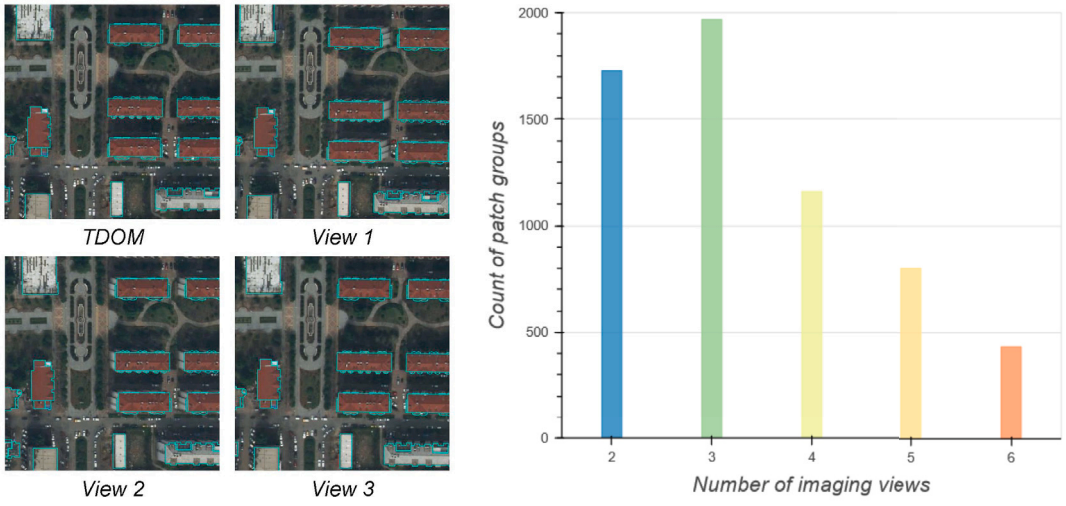
4. Methods

Observing the same target from multiple angles often aids humans in more accurately identifying the target, particularly its edge contours. With this principle in mind, our approach's core objective is to leverage the surplus observational information in original multiview images to elevate the performance of building segmentation within the object space. To achieve this goal, we focus on addressing three key technical challenges: firstly, injecting image-space multiview observation information into the object-space segmentation workflow in a learnable manner; dealing with the inherent errors in the photogrammetric DSM



(a) (b)

Fig. 2. Examples of automatically correcting erroneous image-space projected labels based on occlusion detection. The orange masks indicate the building roof labels. (a) shows incorrect annotations that occur after projecting object-space 3D labels to the image space. (b) demonstrates annotations after automatic occlusion detection and correction of projection errors.



(a) (b)

Fig. 3. (a) A patch group comprising TDOM images and original images from three different viewpoints, in which the ground truth building vectors are delineated in cyan polygons. (b) The statistics on the number of views in all patch groups within the entire dataset.

to robustly project image-space semantics onto the object space; and finally, efficiently implementing end-to-end training for the model. Following we provide the details of our methods and strategies.

4.1. The multiview fusion framework

Our proposed MVMapper is a learning architecture that integrates semantic features from both image and object spaces for segmentation, as depicted in Fig. 4. It includes semantic segmentation modules in both spaces, operating first in image space and then in object space. The framework processes data in patch groups, with each group consisting of N image-space patches and 1 object-space patch. In image space, we concatenate each of the N multiview image crops with its corresponding projected DSM, and these concatenated pairs are then separately input into a feature encoder. This encoder may take the form of either a CNN-based structure or a self-attention-based design, resulting in the generation of multiview semantic feature maps. Subsequently, these feature maps are directed into fully convolutional head structures to generate building segmentation results. Finally, we apply an edge tracing method [47] to extract building contours. All image-space patches share weights for the feature encoder and segmentation head.

In object space, the TDOM image crop belonging to same patch group is concatenated with the DSM data and fed into a network structure that aligns with the one designed for image space. This process produces the object-space semantic feature map. By applying their respective piecewise mapping tables, the N multiview feature maps from the image space are projected to the object space. Subsequently, element-wise averaging of the projected feature maps is performed, followed by addition to the object-space feature map. This step yields a fused feature map, which is then input into the segmentation head to obtain the final building mapping results. Given the potential distortion and noise in TDOM images resulting from DSM errors, we use separate weights for the object-space segmentation network from those in the image space.

4.2. Piecewise affine projection using DSM points

The photogrammetric DSM production is still faces challenges in accurately preserving the sharp edges of ground objects [48,49], particularly when dealing with low-resolution aerial and satellite imagery. When applying the collinearity equation for projecting 2D image-space points onto the object space, errors within the DSM can complicate the accurate localization of their corresponding positions. Consequently, this misalignment often affects the accurate projection of image-space semantic features onto the object space.

To tackle this issue, we propose a piecewise affine projection method based on DSM corresponding points to robustly transform image-space feature maps into the object space. The method's design hinges on a notable observation: while individual DSM point projection may be biased, when considering all the DSM points within a building

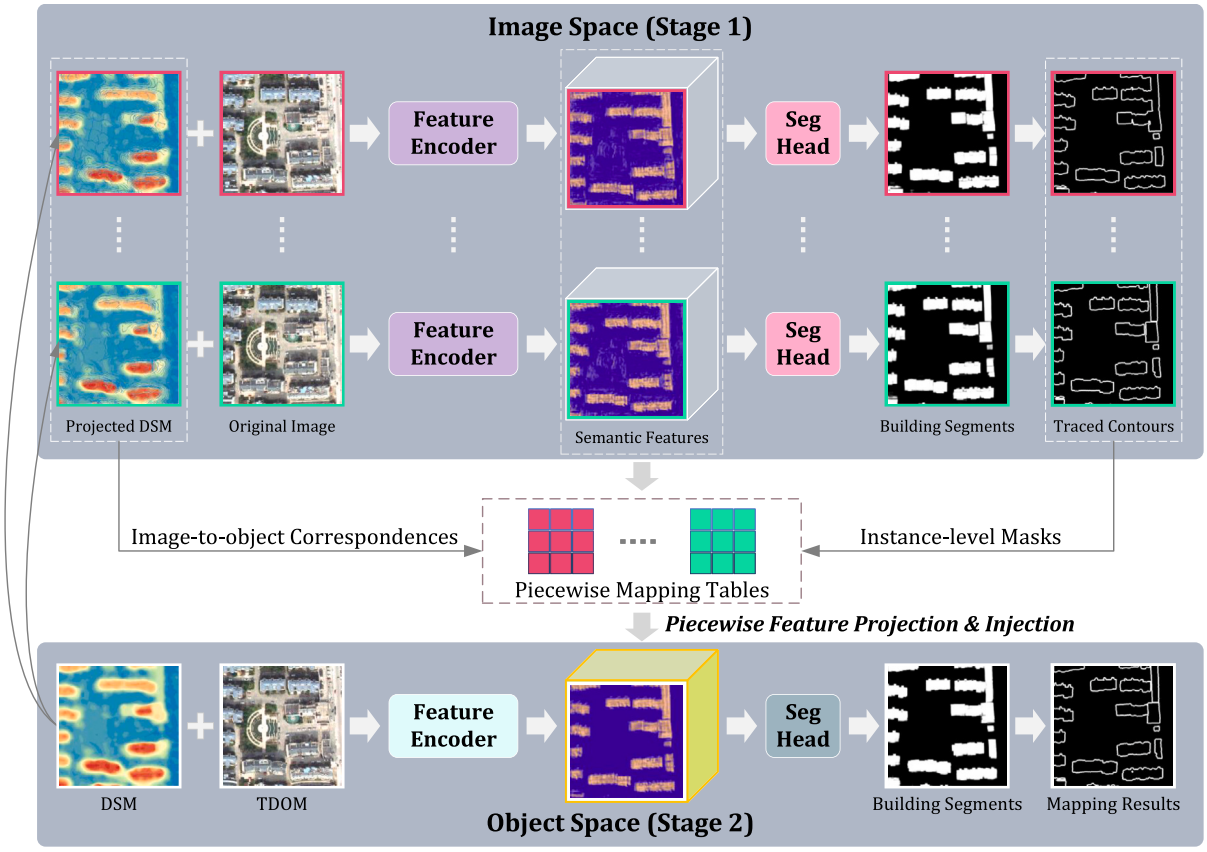


Fig. 4. The architecture of the proposed MVMapper framework, which includes Stage 1 in image space and Stage 2 in object space. In comparison, the Object-space Fusion strategy comprises only the design of MVMapper Stage 2, while the Multiview Voting method lacks the establishment of an end-to-end learning connection between Stage 1 and 2, instead generating segmentation results within their respective pipelines and aggregating them through pixel-level voting.

instance, most of them exhibit reliable image-to-object corresponding relationships. Therefore, as shown in Fig. 4, after completing the image-space stage of our MVMapper framework, we extract DSM points within the predicted contour of each view's building segment. These points, in addition to their image-space coordinates, also possess known object-space coordinates. We leverage these corresponding coordinates to estimate a set of affine parameters for a 2D image-to-object transformation. Subsequently, we apply the estimated parameters to the image-space building contour to obtain an object-space 2D contour.

Fig. 5 illustrates a comparison between projecting image-space predicted building contours onto the TDOM using collinearity equations and the results obtained by our proposed method. It is evident that DSM errors lead to a significant distortion in the projected object-space contours. In contrast, our method demonstrates a more robust ability to maintain the inclusion relationship of contours with respect to the buildings.

Furthermore, it is worth noting that applying the image-to-object affine parameters to each pixel within the image-space building contour does not ensure that every pixel within the object-space contour will be properly filled. Therefore, we compute the inverse transformation associated with the affine parameters, enabling us to determine the corresponding image-space coordinates for every pixel within the object-space contour. We then achieve a seamless projection of the multiview feature maps through interpolation.

4.3. End-to-end training

The piecewise affine projection of multiview feature maps introduces complexity into MVMapper's workflow, prompting us to take two measures to improve the efficiency of end-to-end training. Firstly, we precompute the piecewise transformation parameters and construct a mapping table for each original image before training initiation. This table, aligned with the resolution of MVMapper's feature encoder output, records image-space floating coordinates for every object-space pixel requiring piecewise projection of image-space features. This approach eliminates the necessity for recurrently estimating and applying transformation parameters during training.

Secondly, we address the challenge posed by the dynamic generation of image-space building contours during each iteration of the training process. Utilizing these updated contours to calculate piecewise projection parameters would necessitate frequent updates to the mapping table. To avoid this computational overhead, we employ ground truth data to compute a stable mapping table for each building polygon, keeping it constant throughout MVMapper's training. Furthermore, to optimize the training efficiency, we implement a phased training strategy. Initially, we train the image-space stage until it converges to an optimal state. Following this, we integrate the object-space workflow for end-to-end training of the entire framework. We have found that this phased strategy yields significantly better results compared to direct end-to-end training.

As a framework, the choice of loss function is not the primary focus of MVMapper's design. We employ the same loss function for both the segmentation maps in the image and object spaces, leading to L_{image} and L_{object}, respectively. To keep things simple, we do not allocate specific loss weights to the two segmentation workflows.

5. Results and discussion

5.1. Implementation details

We implemented and tested MVMapper in PyTorch on a 64-bit Ubuntu system equipped with 6 NVIDIA GeForce GTX 2080 Ti GPUs.

Fig. 5. Comparison between collinearity-based projection and piecewise affine projection applied to image-space predicted building contours.

Recognizing the effectiveness of High-Resolution Networks (HRNet) in retaining fine-grained features [50], we adopted its design as our segmentation backbone structure. Following the original HRNetV2p-W32 implementation, we produced feature maps at resolutions of 1/4, 1/8, 1/16, and 1/32 of the input. These feature maps were then upsampled to 1/4 resolution and concatenated. The concatenated feature map underwent convolution processing and was further upsampled to match the input resolution for segmentation inference.

In image space, we employed the piecewise affine projection method to convert the concatenated feature map generated by each view into the object space for feature fusion. During training, we utilized AdamW optimizer [51] with a learning rate of 0.003. To ensure thorough model training, we conducted a total of 300 epochs, implementing a cosine decay learning rate adjustment strategy with a cycle of 300 epochs and a minimum learning rate of 0.

We set the crop size to 512×512 pixels and generated the crops using a sliding window with a 30% overlap between adjacent crops. During the image-space training stage, the batch size was set to 16 per GPU. In the second training stage, the batch size was set to 8 per GPU, with this number referring to the patch groups in object space. The number of input samples in image space for each group depended on the number of image views. To address class imbalance issues, we combined the dice loss [52] and focal loss [53] to construct the loss function, applying it simultaneously to both image and object spaces.

The images within the subregions for training and validation were initially cropped and organized into patch groups. Subsequently, a random selection of 90% of these patch groups was used for training, while the remaining 10% was reserved for validation. After completing hyperparameter tuning, these two subsets were merged, and training continued with a reduced learning rate. The resulting model was ultimately applied to the test dataset for method comparison.

We completed the training of MVMapper in a total of 9.58 h. It is noteworthy that the computation time for calculating projection parameters for one image-space patch was 0.045 s, and there were a total of 5112 image-space patches in all training data. If we did not adopt our proposed precomputing mapping tables approach and instead estimated and applied transformation parameters during each iteration, it would result in an additional time of approximately 230 s per epoch. This would require an additional 19.17 h for model training. At inference stage, MVMapper required an average time of 0.620 s per patch group.

5.2. Overall comparison

We used metrics of IoU [54], F1-score, Precision and Recall [55] to evaluate the overall performance of building segmentation. Additionally, we evaluated the accuracy of the generated building contours using contour-based metrics proposed by Perazzi et al. [56]. This method entails applying mathematical morphology operations, specifically dilation, to both the generated building contours and the ground truth polygons. By dilating the contours, we created buffer zones around each contour, serving as regions of interest for similarity measurement. We then computed IoU, F1-score, Precision, and Recall over five different buffer sizes, ranging from 1 to 5 pixels, and used their averages as contour-based metrics.

We compared MVMapper with two typical multiview fusion strategies from Table 1, and for a fair comparison, we made the following adaptations:

- **Object-space Fusion:** We adopted MVMapper's object-space structure and maintained consistent implementation settings to represent this strategy for comparison.
- **Multiview Voting:** We implemented a comparable version of this method. We separately trained and applied MVMapper's image-space and object-space structures to perform building segmentation. We used our piecewise affine projection method to convert the predicted probability map from each view onto the object space, and then applied Hu et al. [30]'s voting method to combine these projected maps with the object-space prediction to obtain the final result.

We maintained consistent network backbone and segmentation head structures across different comparison methods to focus on evaluating the fusion strategies.

Table 2 displays the overall performance of the three methods on the test dataset. MVMapper shows improvements compared to the Object-space Fusion strategy. Specifically, the Segmentation IoU increased by 2.3%, indicating the successful utilization of semantic information from multiview original images to enhance building segmentation performance. Moreover, the contour-based IoU saw an even more substantial improvement of 4.8%, highlighting MVMapper's strength in producing more accurate edge results. Surprisingly, the Multiview Voting method not only failed to improve accuracy but also exhibited a slight overall performance decline compared to the Object-space Fusion strategy. This suggests that without the benefit of end-to-end learning for network refinement, a simple voting approach struggles to effectively leverage information from multiview original observations.

The evaluation details of building contour results based on different buffer sizes ranging from 1 to 5 pixels are depicted in Fig. 6. It can be observed that the performance metrics of all three methods improve with increasing buffer size. The enhancement of our approach relative to the comparative methods is slightly more pronounced at smaller buffer sizes (e.g., 1 pixel).

Since the three comparative methods exhibit similar performance in identifying the main body of buildings' rooftops, their distinctions mainly lie in their ability to reproduce fine details in individual building edges. To highlight these differences, we select typical areas within the 6 test subregions and magnified the results for comparison. It is worth noting that in TDOM, due to DSM errors, buildings often exhibit

Table 2
Overall quantitative evaluation results of the comparative methods.

Method	Segmentation metrics				Contour-based metrics			
	IoU	F1-score	Precision	Recall	IoU	F1-score	Precision	Recall
Object-space Fusion	0.838	0.912	0.923	0.901	0.552	0.712	0.728	0.696
Multiview Voting	0.832	0.909	0.926	0.892	0.519	0.683	0.706	0.662
MVMapper (Ours)	**0.861**	**0.925**	**0.936**	**0.915**	**0.600**	**0.750**	**0.757**	**0.743**

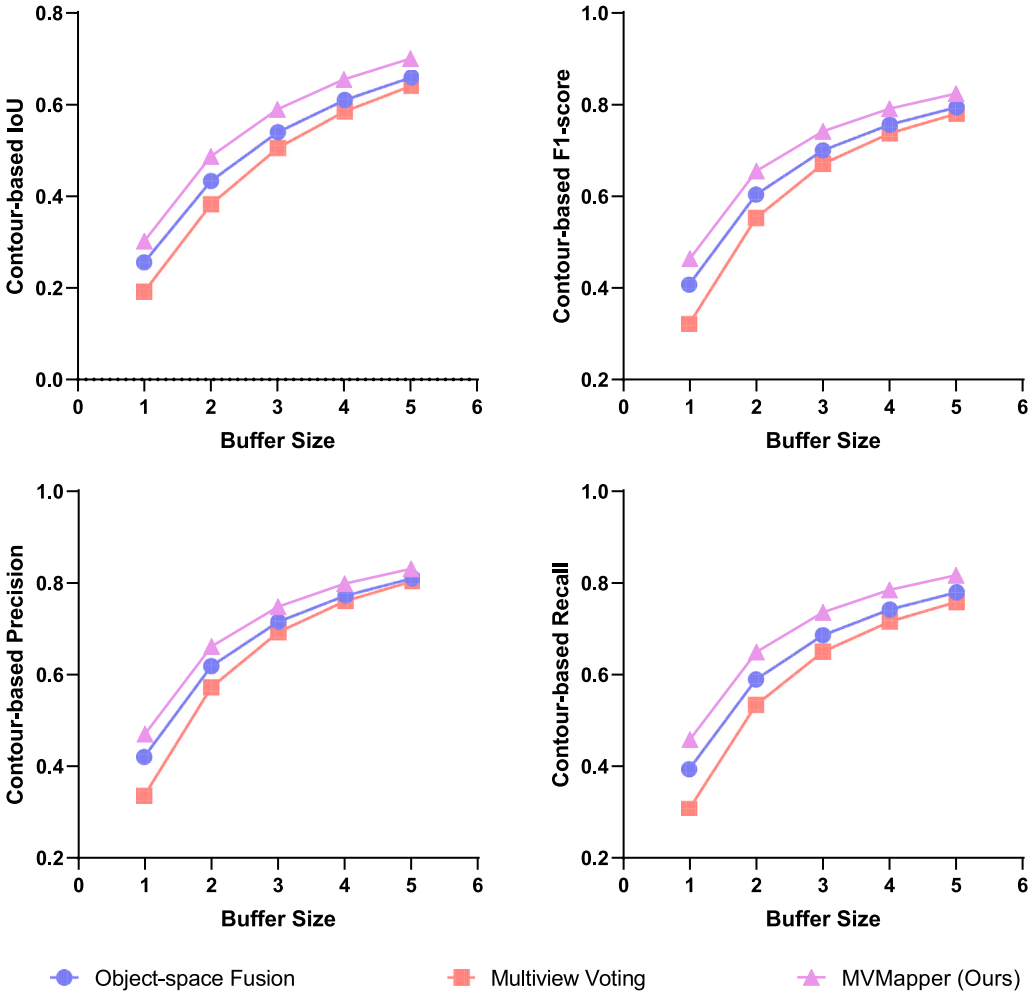


Fig. 6. Contour-based evaluation results under varying buffer sizes in the overall comparison.

certain distortions along their edges. As a result, many of them do not perfectly align with the ground truth produced by stereo mapping.

As shown in Fig. 7, Object-space Fusion exhibited more false detections around certain buildings (see Fig. 7c, e). These inaccuracies primarily stemmed from elevation errors in the DSM, despite the DSM's overall contribution to building segmentation accuracy. Multiview Voting improved these misclassification issues after incorporating multiview results projected from image space. However, upon closer examination, we noticed that it did not consistently make optimal decisions at building edges. This resulted in systematic edge misalignment (as indicated by the orange circle) and relatively more noticeable edge omissions in some instances (as indicated by the purple arrows). In comparison, MVMapper not only avoided pronounced errors through its multiview fusion approach but also demonstrated superior accuracy in recovering edge details.

Fig. 8 showcases segmentation results for several typical samples obtained through the three methods. It is important to note that, in our pursuit of comparing segmentation performance, we intentionally refrained from simplifying or regularizing the generated contours. We observed that when TDOM introduces substantial distortions (see Fig. 8a), Object-space Fusion exhibits some vulnerability to these distortions. Multiview Voting shows an ability to rectify the obvious errors but struggles to faithfully reproduce better edge details. Furthermore, where roof textures closely resemble the background and the DSM fails to provide precise height differentiation (e.g., in Fig. 8b, the roof and background present a height difference of about 5 meters in the input DSM), Object-space Fusion might erroneously classify buildings as part of the background. Multiview Voting can alleviate such misclassifications to some extent. However, due to the limitations in the accuracy of image-to-object projection, notable edge errors may still persist. MVMapper performs better in both of these scenarios.

More often than not, MVMapper's advantages over the other two methods are primarily evident in the subtle improvements it makes to building edge structures. For instance, in Fig. 8c, MVMapper avoids the false detection in the top left corner seen in Object-space Fusion, while also correcting the leftward misalignment issue observed in Multiview Voting; in Fig. 8d, MVMapper excels at accurately reproducing the sawtooth-like structure along the roof boundaries.

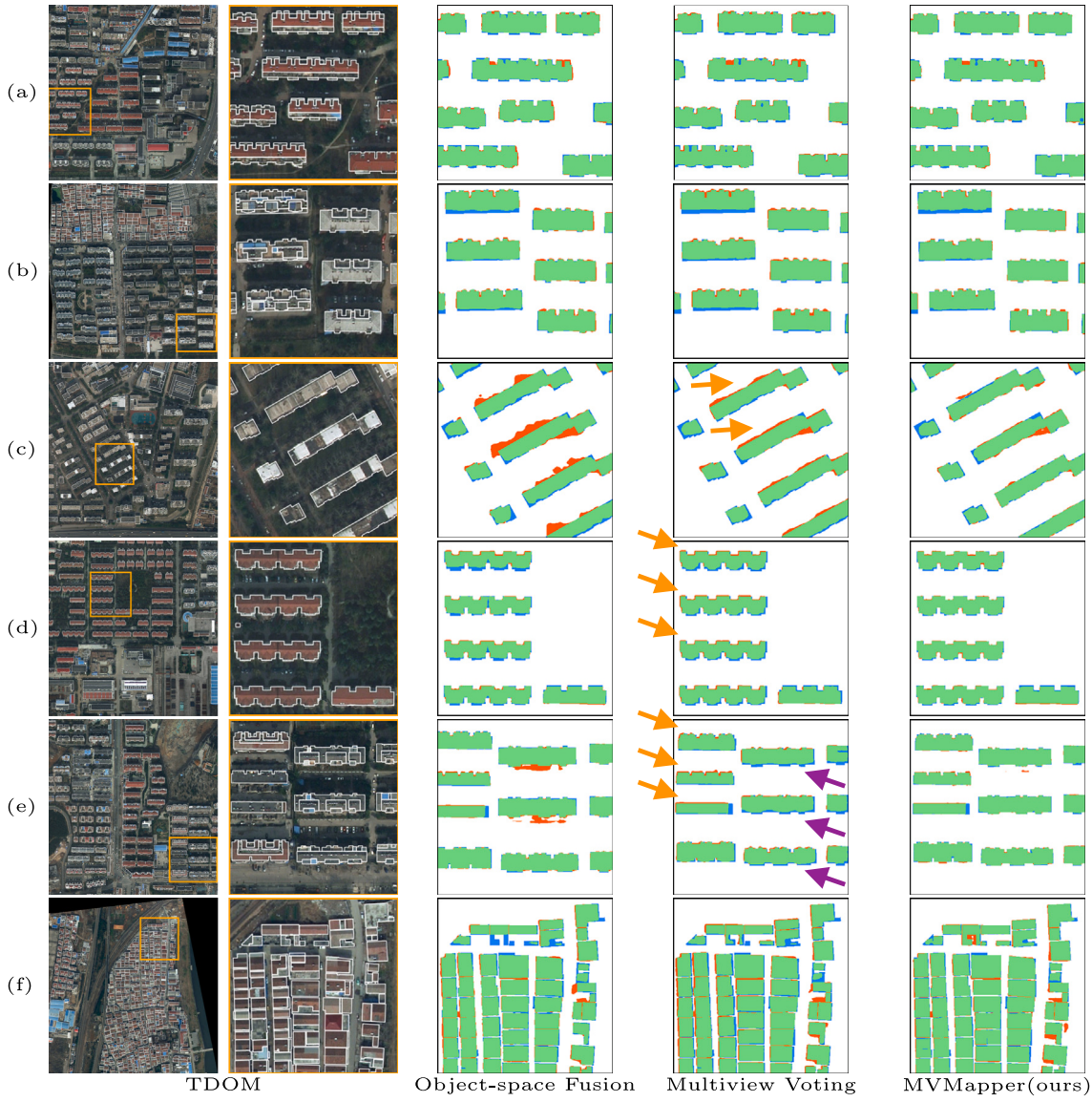


Fig. 7. Building segmentation results of typical scenes using comparative methods in the 6 test subregions. Ground truths are depicted by white polygons on the TDOM. Evaluation maps are color-coded, with true positives in green, false positives in red, and false negatives in blue.

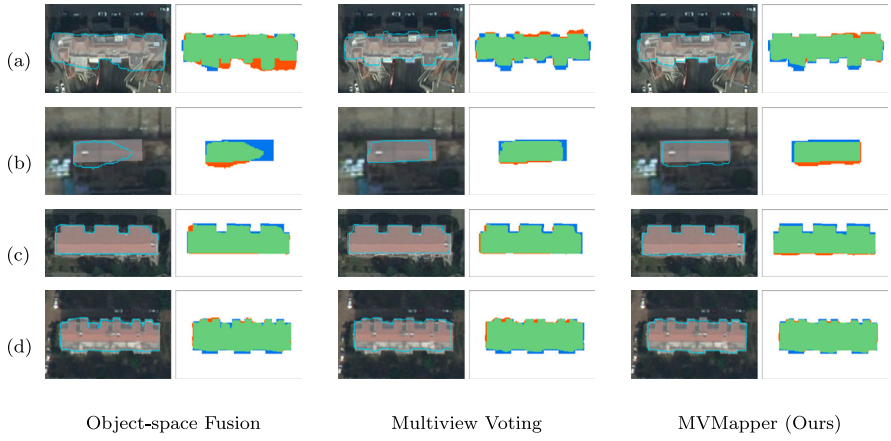


Fig. 8. Results of typical building instances using comparative methods. The cyan polygons represent the building contours generated by the models, while the ground truths are overlaid on the TDOM as white transparent masks. Evaluation maps are color-coded, with true positives in green, false positives in red, and false negatives in blue.

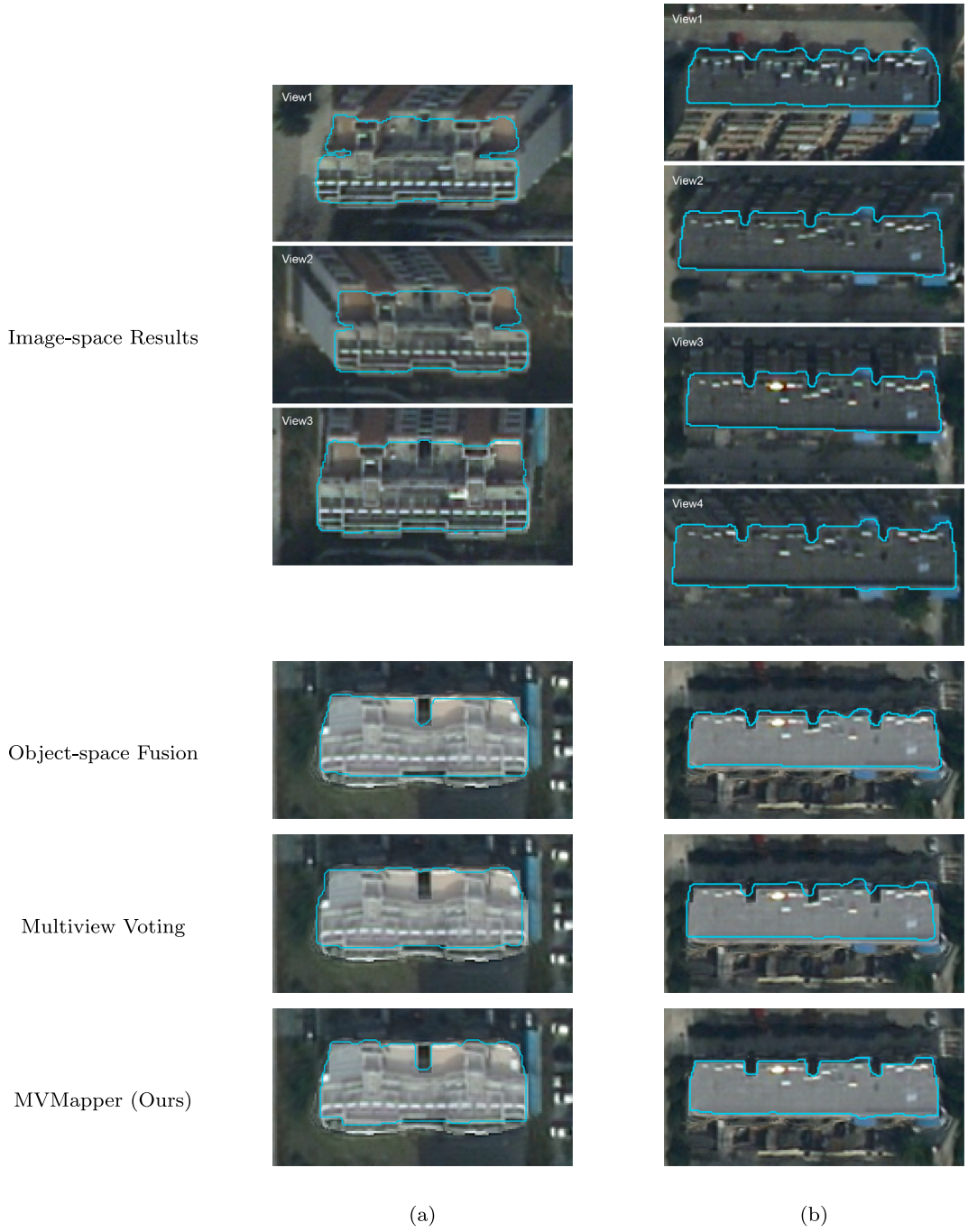


Fig. 9. Contour Results generated by different segmentation modules/methods for two buildings, indicating the benefits of MVMapper's end-to-end learning. The cyan polygons represent the building contours generated by the models, while the ground truths are overlaid on the TDOM as white transparent masks.

5.3. The advantage of end-to-end learning

In order to demonstrate the advantages of MVMapper's end-to-end training in comparison to straightforward voting strategies, we present additional intermediate results using typical samples. Fig. 9 illustrates the contours corresponding to the segmentation outcomes of two buildings, each captured from 3 and 4 different views in the image space.

It is evident that the segmentation results in the image space, obtained through multiviews, exhibit variations in quality. Subsequently, when employing piecewise affine projection, the resulting contours often contain some degree of error. While the Multiview Voting method can mitigate errors that may occur in a few image views, it encounters difficulties in effectively addressing projection errors, particularly at the edges.

In contrast, despite the observed misalignment of image-space features when projected onto the object space, it is clear that MVMapper, through end-to-end learning, achieves adaptive reordering and efficient fusion of multisource features. Consequently, this adaptive feature recombination successfully enhances mapping performance by integrating semantic features from both multi-view image space and object space input data.

Table 3
The evaluation results using different feature fusion strategies.

Method	Segmentation metrics				Contour-based metrics			
	IoU	F1-score	Precision	Recall	IoU	F1-score	Precision	Recall
Multiview Fusion w/o TDOM	0.839	0.913	**0.944**	0.883	0.550	0.710	0.722	0.698
Nadir-TDOM Fusion	0.850	0.919	0.928	0.910	0.575	0.730	0.730	0.731
Multiview Fusion (ours)	**0.861**	**0.925**	0.936	**0.915**	**0.600**	**0.750**	**0.757**	**0.743**

5.4. Effectiveness of design options

By evaluating the performance of alternative design options, we conducted additional experiments to validate the necessity of MVMapper's two key designs: multiview feature fusion and piecewise affine projection.

5.4.1. Multiview feature fusion

We compared MVMapper's multiview fusion design with two ablated versions: the first one, referred to as Multiview Fusion without TDOM, removed the object-space input from MVMapper but retained the projection of image-space multiview features into the object space for prediction; the second one, called Nadir-TDOM Fusion, utilized only the features from the nadir view in the image space for fusion with TDOM, excluding all other original images.

Table 3 presents the evaluation results obtained using different feature fusion strategies. We observed that when we exclude object-space input from the fusion process and rely solely on image-space multiview features, the segmentation performance is limited, similar to that of Object-space Fusion (0.839 vs. 0.838 in IoU). Nadir-TDOM Fusion demonstrates improved segmentation accuracy by fusing both image-space and object-space features. However, the full multiview fusion approach, which utilizes all original images, further enhances segmentation performance, achieving a 1.1% increase in segmentation IoU and a 2.5% boost in contour-based IoU.

Fig. 10 presents detailed results of building contour evaluation using different feature fusion strategies across varying buffer sizes. It can be observed that our fusion approach consistently improves contour accuracy across different evaluation buffers.

The examples in Fig. 11 indicate that Multiview Fusion without TDOM tends to miss small structures at building edges or produce errors at the edges. Nadir-TDOM Fusion, on the other hand, can result in discontinuities in the segmentation results (Fig. 11b, c). In comparison, the full multiview fusion approach exhibits stronger capabilities in structural reconstruction and detail extraction.

5.4.2. Piecewise affine projection

We constructed two variants by replacing MVMapper's piecewise affine projection with alternative methods for comparative experiments: the first variant directly transfers image-space features to object space using the collinearity equation; the second variant, for each original image, estimates a global 2D-to −2D affine transformation using all DSM points within its extent and applies this transformation for feature projection.

From the results in Table 4, it can be observed that using collinearity-based projection significantly reduces the model's performance, especially negatively affecting contour-based accuracy. This indicates that the projection error caused by DSM noise hampers the effectiveness of multiview fusion. The adoption of global affine projection notably improves model performance. It is easy to predict that this global estimation method often leads to misalignment when projecting roof features from image space to object space. The reason for the significant improvement may be attributed to the ability of end-to-end learning to effectively reorganize and fuse features spatially. Piecewise affine projection, building upon a better initial feature projection, achieves superior feature fusion effects. Notably, it leads to a 3.6% improvement in contour-based IoU compared to global projection.

Fig. 12 illustrates building contour evaluation details using different feature projection strategies across varying buffer sizes, showcasing the superior performance of the proposed piecewise affine projection method across all contour-based metrics at different buffer sizes.

The examples in Fig. 13 further illustrate the issues with collinearity-based projection. The impact of projection errors can lead to considerable fluctuations at the contours and misjudgment of building extents in the segmentation results. The end-to-end learning of feature fusion appears to struggle with such significant errors. Another interesting observation is that the results obtained using collinearity-based projection sometimes align well with the distorted building contours in TDOM (Fig. 13c). However, this alignment is not the desired outcome in practical building mapping production. Results from global affine projection are quite comparable to ours, with the primary distinction being the level of detail in the edge contour.

5.5. Cross validation and statistical test

To ensure a comprehensive evaluation, we conducted a 5-fold cross-validation. Following the initial experiment involving 30 subregions, four additional trials were carried out. In each trial, 24 subregions were randomly selected for training, while a distinct set of 6 subregions was chosen for testing. This sampling process was performed without replacement to guarantee a diverse assessment across different data subsets. Fig. 14 illustrates the data splits for these folds. Table 5 presents the quantitative evaluation results from the 5-fold cross-validation experiments.

To assess the statistical significance of our approach compared to the other two methods, we employed segmentation IoU and contour-based IoU as primary performance metrics for analysis. Following 5-fold cross-validation, we conducted the non-parametric Wilcoxon paired tests [57] for MVMapper and the two comparative methods. Each test involved formulating the following hypotheses:

- **Null Hypothesis** H_0: $\mu_1^i - \mu_2^i$ comes from a distribution with median no greater than 0.
- **Alternative Hypothesis** H_A: $\mu_1^i - \mu_2^i$ comes from a distribution with median greater than 0.

Here, i ranges from 1 to 4, representing distinct hypothesis tests. For instance, (μ_1^1, μ_2^1) and (μ_1^2, μ_2^2) denote the paired sample medians of Segmentation IoU values between MVMapper and Object-space Fusion, and MVMapper and Multiview Voting, respectively. Similarly, (μ_1^3, μ_2^3) and (μ_1^4, μ_2^4) indicate the paired sample medians of Contour-based IoU values between MVMapper and Object-space Fusion, and MVMapper and Multiview Voting, respectively.

The results show that all one-tailed p-values for the four Wilcoxon paired tests are 0.03125, leading to the rejection of the null hypothesis at the 0.05 significance level but not at the 0.01 level based on the current 5-fold cross-validation. This indicates that our method exhibits statistically significant improvements over the two comparative methods in terms of the two primary metrics, albeit with a relatively large p-value at the 0.05 level.

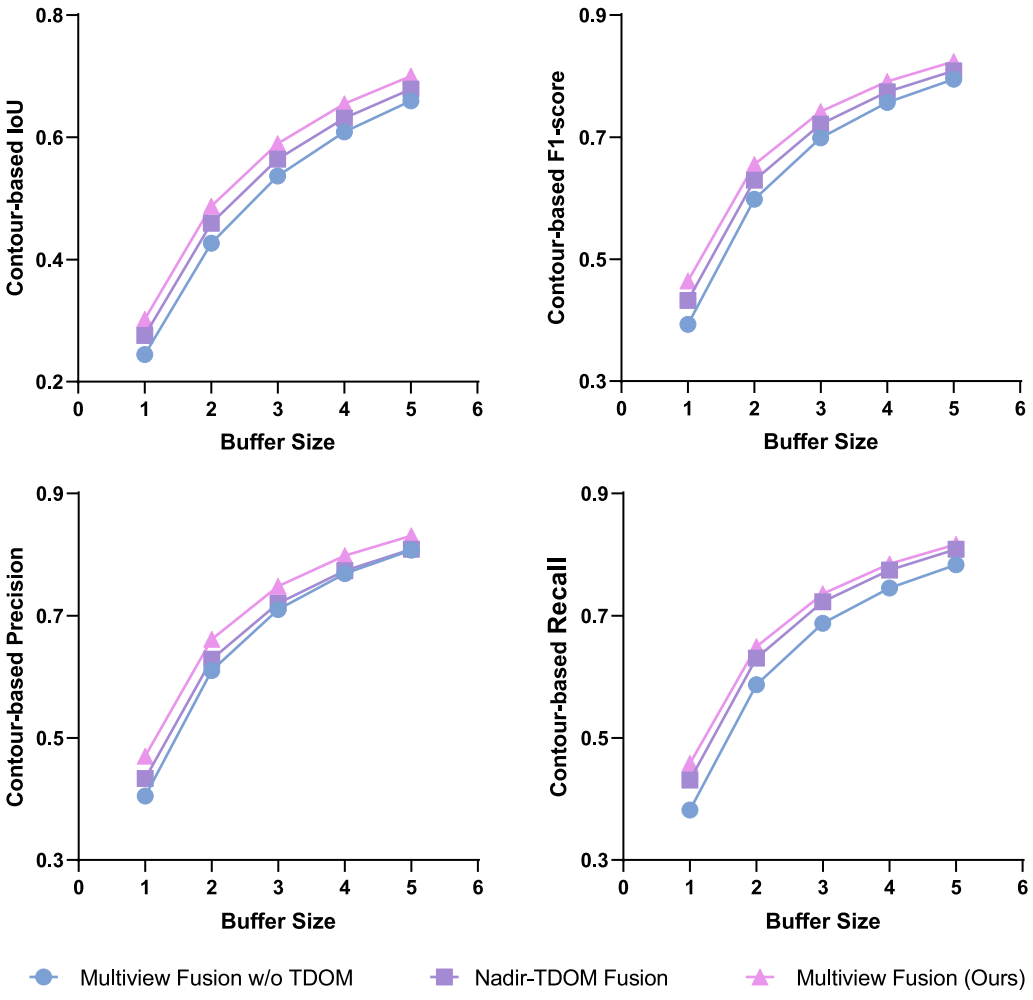


Fig. 10. Contour-based evaluation results under varying buffer sizes using different feature fusion strategies.

Table 4
The evaluation results using different image-to-object feature projection strategies.

Method	Segmentation metrics				Contour-based metrics			
	IoU	F1-score	Precision	Recall	IoU	F1-score	Precision	Recall
Collinearity-based Projection	0.785	0.879	0.896	0.863	0.469	0.639	0.645	0.633
Global Affine Projection	0.843	0.915	0.927	0.903	0.564	0.721	0.733	0.710
Piecewise Affine Projection (ours)	**0.861**	**0.925**	**0.936**	**0.915**	**0.600**	**0.750**	**0.757**	**0.743**

5.6. Transferability validation

To validate MVMapper's applicability to other datasets, given that the Xuzhou dataset is privately owned and not publicly shareable, we performed additional tests using an openly available oblique aerial imagery dataset in Dortmund city center [15]. To maintain consistency with the Xuzhou experiment, we only utilized the nadir camera data from the five cameras available in this dataset. We selected a 2.5 square kilometer area for our experiment. Due to the absence of building labels, skilled personnel were employed to create 3D roof annotations through stereo mapping, resulting in a total of 4568 polygonal vectors. As shown in Fig. 15, we divided the entire area evenly along the X-axis into five subregions and randomly selected one subregion for model tuning. The remaining four subregions were used for accuracy validation and comparison.

Table 6 shows the quantitative evaluation results of the methods on Dortmund dataset. MVMapper achieved the highest Segmentation IoU of 0.830, representing an 6.3% improvement over Object-space Fusion. Additionally, MVMapper demonstrated significant improvements in contour-based metrics, particularly with an improvement of 11.0% in Contour-based IoU. This indicates its advantage in generating more accurate building contours compared to alternative methods. It is worth noting the more significant improvement observed with MVMapper compared to Xuzhou experiment. This may be attributed to the comparatively smaller proportion of data used for training across all models (about 20% of the data utilized for model tuning), which demonstrates MVMapper's capability to effectively extract and leverage feature information from the multiview data in image space.

Fig. 16 illustrates the detailed contour evaluation across various buffer sizes in the Dortmund experiment. It is evident that MVMapper achieved significant improvements across all buffer sizes. Part of this improvement may be attributed to the noticeable decline in contour accuracy of all three methods compared to the Xuzhou experiment. This decline could be due to the higher spatial resolution of the Dortmund dataset, which is 10 centimeters, compared to the 30-centimeter resolution of the Xuzhou data. The higher resolution allows for more detailed depiction and annotation of building edges, posing a greater challenge for all three methods in capturing these intricate details.

Fig. 17 presents the segmentation results evaluation, revealing that MVMapper generated more false detections in this experimental group

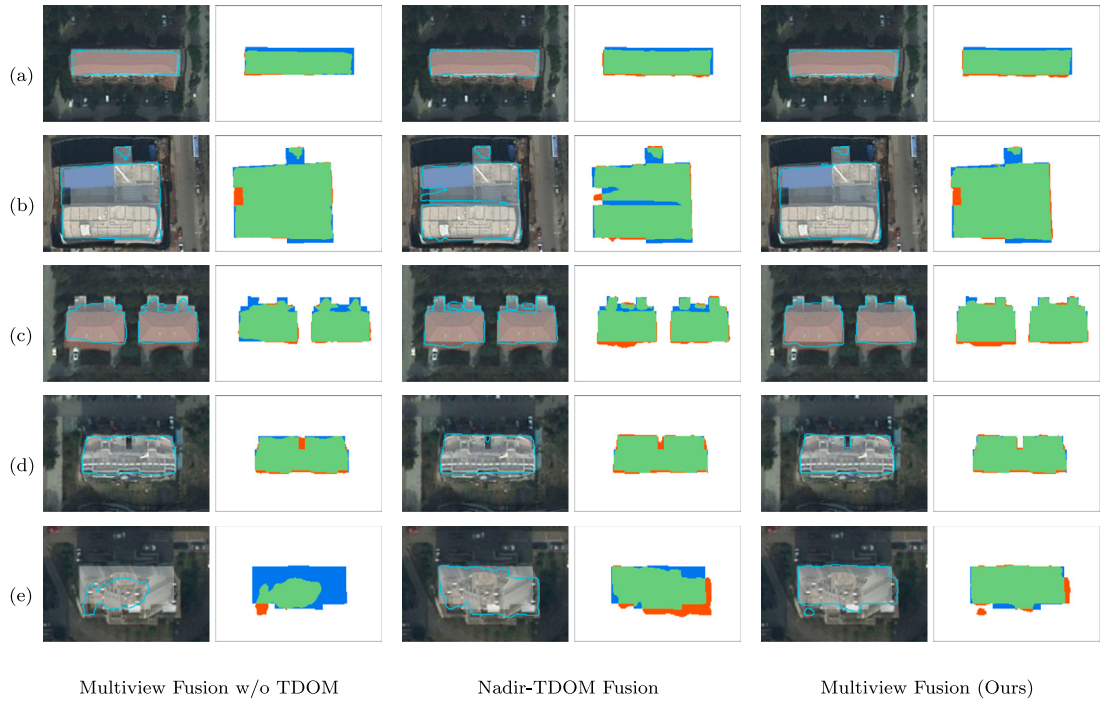


Fig. 11. Segmentation results of typical building instances using different feature fusion strategies. The cyan polygons represent the building contours generated by the models, while the ground truths are overlaid on the TDOM as white transparent masks. Evaluation maps are color-coded, with true positives in green, false positives in red, and false negatives in blue.

Table 5
5-fold cross validation results of the comparative methods.

Method	Segmentation metrics				Contour-based metrics			
	IoU	F1-score	Precision	Recall	IoU	F1-score	Precision	Recall
Fold 1								
Object-space Fusion	0.838	0.912	0.923	0.901	0.552	0.712	0.728	0.696
Multiview Voting	0.832	0.909	0.926	0.892	0.519	0.683	0.706	0.662
MVMapper (Ours)	**0.861**	**0.925**	**0.936**	**0.915**	**0.600**	**0.750**	**0.757**	**0.743**
Fold 2								
Object-space Fusion	0.796	0.886	0.909	0.865	0.499	0.666	0.689	0.644
Multiview Voting	0.782	0.877	**0.911**	0.846	0.465	0.635	0.652	0.619
MVMapper (Ours)	**0.815**	**0.898**	0.901	**0.895**	**0.526**	**0.689**	**0.699**	**0.680**
Fold 3								
Object-space Fusion	0.816	0.898	0.922	0.876	0.507	0.672	0.699	0.648
Multiview Voting	0.816	0.899	**0.925**	0.874	0.483	0.652	0.678	0.627
MVMapper (Ours)	**0.850**	**0.919**	0.922	**0.916**	**0.559**	**0.717**	**0.734**	**0.702**
Fold 4								
Object-space Fusion	0.850	0.919	**0.938**	0.901	0.550	0.710	0.747	0.676
Multiview Voting	0.839	0.912	0.934	0.892	0.530	0.693	0.725	0.663
MVMapper (Ours)	**0.866**	**0.928**	0.922	**0.934**	**0.573**	**0.729**	**0.771**	**0.691**
Fold 5								
Object-space Fusion	0.819	0.901	0.914	**0.888**	0.482	0.651	0.660	0.641
Multiview Voting	0.779	0.876	**0.941**	0.820	0.441	0.612	0.638	0.588
MVMapper (Ours)	**0.823**	**0.903**	0.921	0.885	**0.516**	**0.681**	**0.677**	**0.684**

Table 6
Quantitative evaluation results of the comparative methods applied to Dortmund dataset.

Method	Segmentation metrics				Contour-based metrics			
	IoU	F1-score	Precision	Recall	IoU	F1-score	Precision	Recall
Object-space Fusion	0.767	0.868	0.937	0.809	0.216	0.355	0.357	0.352
Multiview Voting	0.755	0.860	**0.978**	0.768	0.180	0.306	0.312	0.299
MVMapper (Ours)	**0.830**	**0.907**	0.898	**0.916**	**0.326**	**0.492**	**0.465**	**0.522**

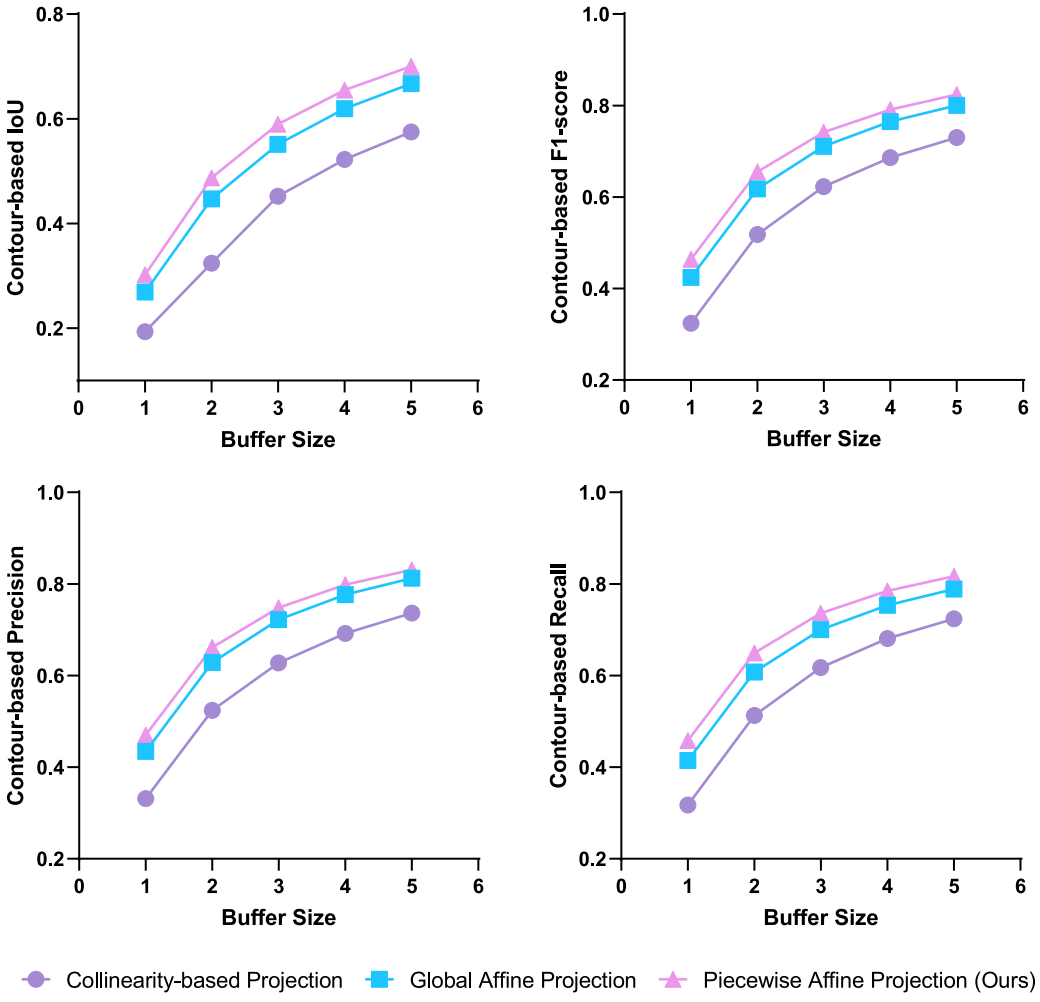


Fig. 12. Contour-based evaluation results under varying buffer sizes using different feature projection strategies.

(as demonstrated by the lower segmentation Precision value in Table 6). However, it excelled in more comprehensive detection of building targets, leading to better overall performance. The samples depicted in Fig. 18 further illustrate MVMapper's superior ability to capture complete roof structures. However, the false positives generated at the edges do impact its contour accuracy.

5.7. Limitations and other potential applications

One of the limitations of our proposed method is the requirement for 3D annotations. Although establishing a 3D-to −2D projection relationship can eliminate the need for additional 2D annotations in the image space, the cost of annotating 3D labels in object space is still relatively high. In future work, we plan to explore the development of a self-supervised learning framework in the image space, which may enable model training with only 2D labels applied in the object space. Additionally, our method relies on DSM data generated using traditional approaches. Incorporating deep-learning-based DSM generation frameworks with MVMapper's building segmentation framework is another area for potential improvement.

While our framework is tailored for photogrammetric data, its core concept lies in establishing connections between 2D and 3D spaces and conducting fusion learning to better utilize multiview features. Therefore, provided that the imaging geometry of dual spaces can be well modeled, we believe our method holds potential for applications beyond photogrammetry, such as medical image diagnosis and crop growth monitoring.

6. Conclusion

In this work, we introduce MVMapper, an end-to-end learning framework that establishes a direct connection between original multiview images and the object-space segmentation task, enabling effective multiview feature fusion and enhancing building segmentation and contour accuracy. MVMapper initiates by running an image-space building segmentation module to obtain multiview building contour predictions. It subsequently executes an object-space segmentation workflow, mirroring the design of the image-space module, and utilizes TDOM and DSM as inputs. This process facilitates the fusion learning of image-space and object-space features to predict the final segmentation results. The piecewise affine projection method applied to the multiview original features ensures robust image-to-object transformation for effective fusion. Cross-validation experiments and hypothesis testing results demonstrate that MVMapper achieves statistically significant improvements in segmentation IoU and contour-based IoU compared to traditional multiview fusion strategies for building segmentation. Transferability validation experiments on a separate open-source dataset further support the higher performance of our method.

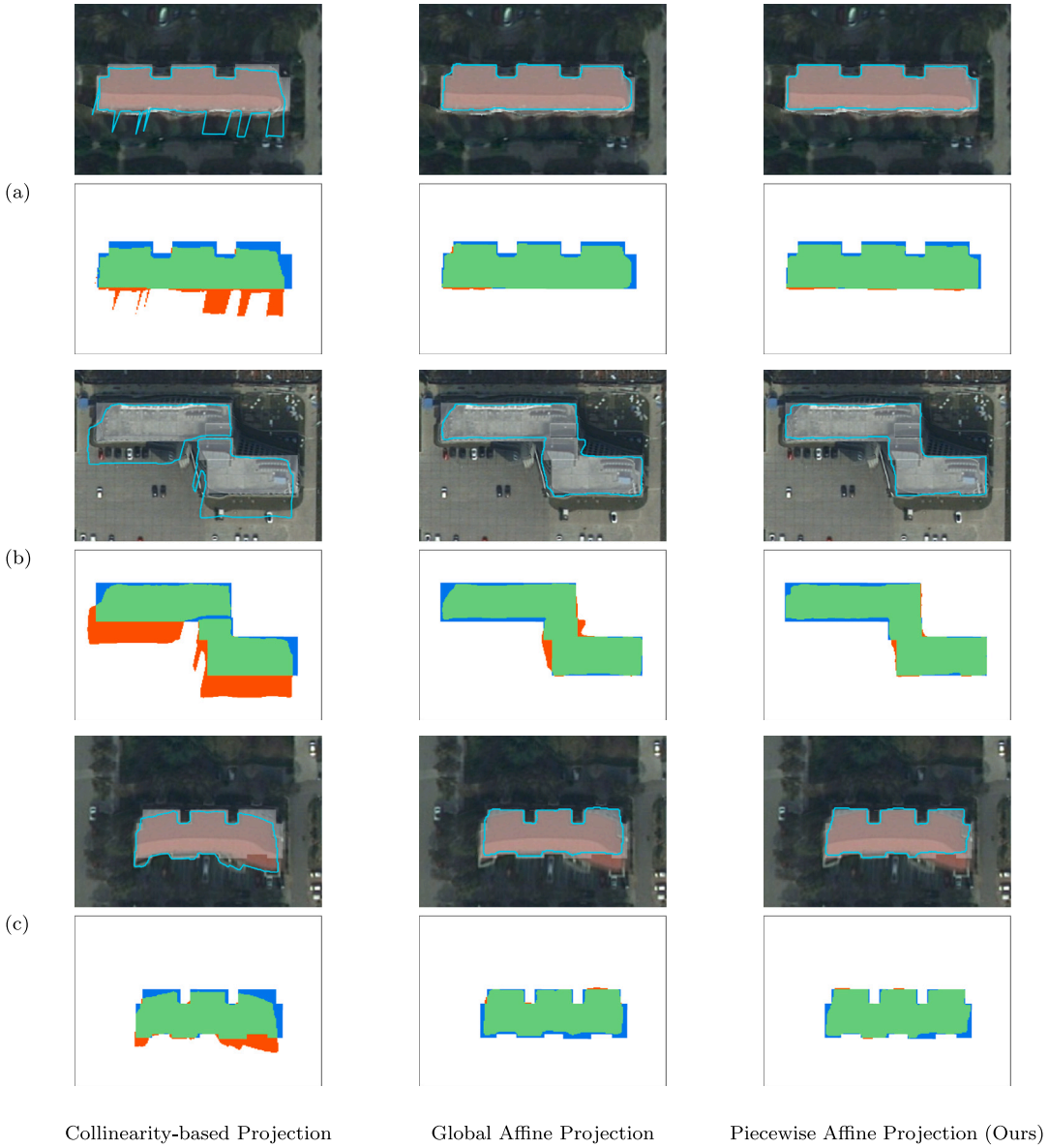


Fig. 13. Segmentation results of typical building instances using different image-to-object feature projection strategies. The cyan polygons represent the building contours generated by the models, while the ground truths are overlaid on the TDOM as white transparent masks. Evaluation maps are color-coded, with true positives in green, false positives in red, and false negatives in blue.

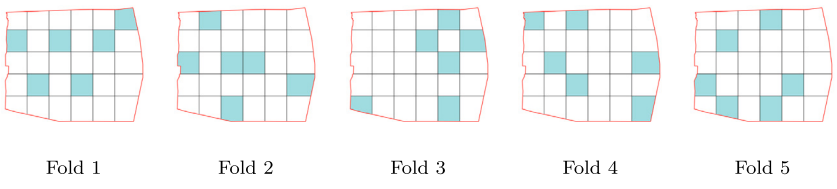
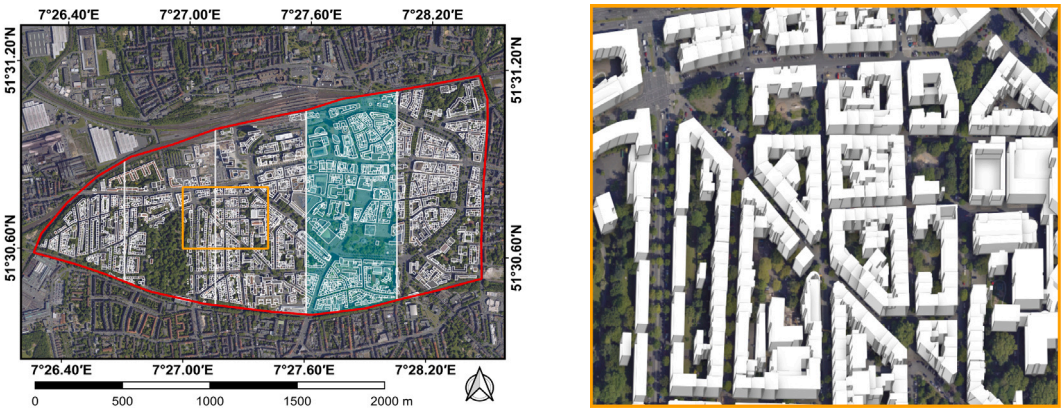


Fig. 14. Data splits for the 5-fold cross-validation experiment, where Fold 1 refers to the initial trial corresponding to Fig. 1a.



(a) (b)

Fig. 15. Overview of the Dortmund dataset utilized for transferability validation. (a) delineates the selected experiment area with a red polygon, showcasing annotated buildings represented by white vectors. The light green shaded area is designated for model tuning, while the remaining regions are reserved for accuracy validation. (b) presents 3D perspectives of building annotations within the orange box outlined in (a).

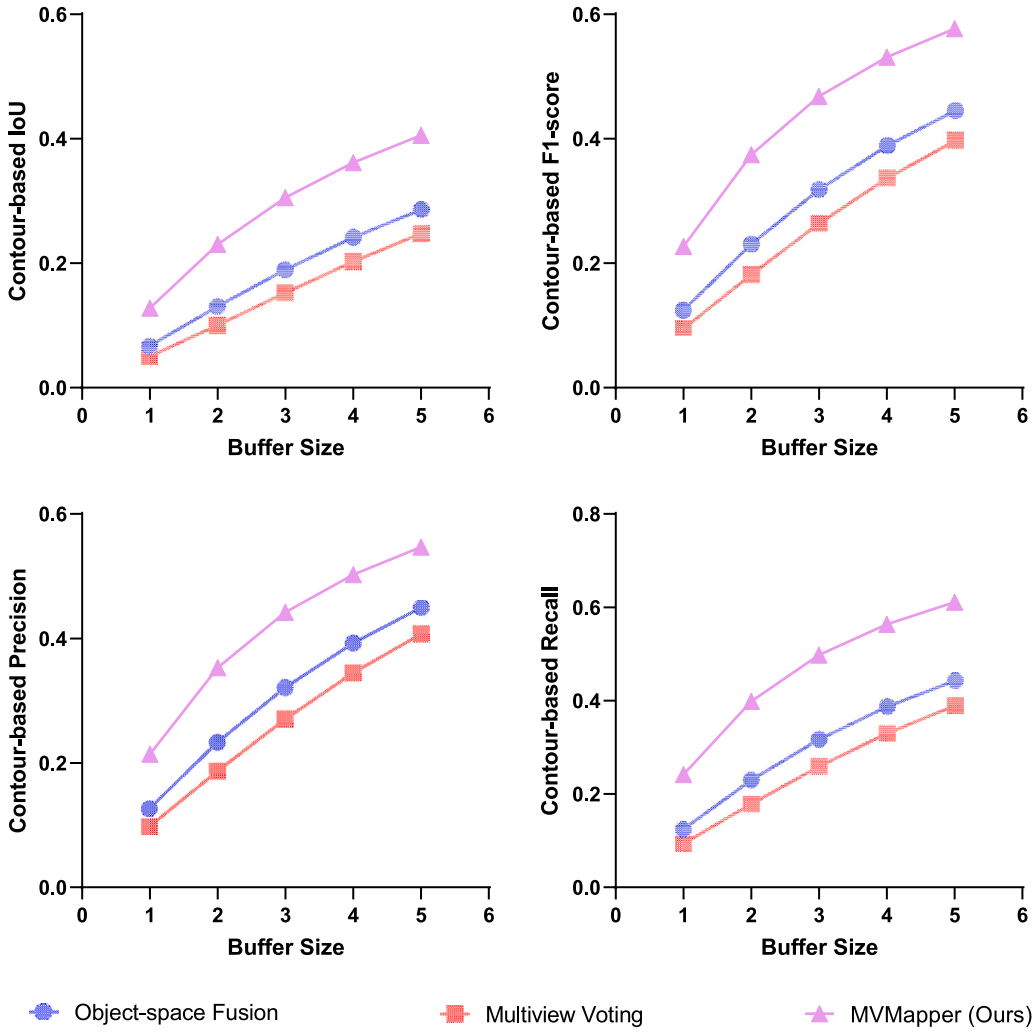


Fig. 16. Contour-based evaluation results under varying buffer sizes in Dortmund experiment.

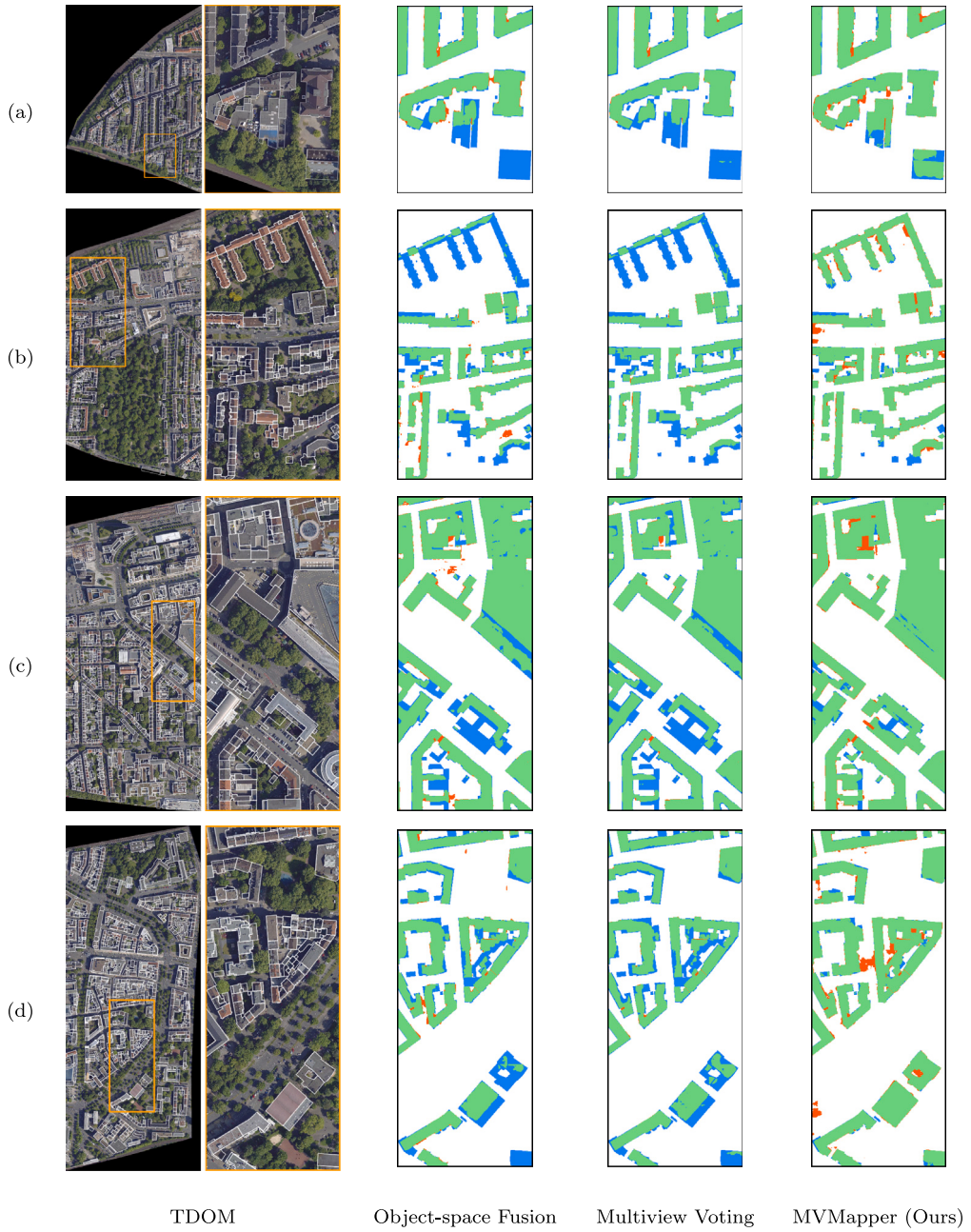


Fig. 17. Building segmentation results of typical scenes in Dortmund experiment. Ground truths are depicted by white polygons on the TDOM. Evaluation maps are color-coded, with true positives in green, false positives in red, and false negatives in blue.

CRediT authorship contribution statement

Qi Chen: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Wenxiang Gan:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Pengjie Tao:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition. **Penglei Zhang:** Visualization, Validation. **Rongyong Huang:** Writing – review & editing, Validation. **Lei Wang:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 42371475 and 41801390. It is also supported by the High Performance Computing Center at Eastern Institute of Technology, Ningbo, and High Performance Computing Center at Ningbo Institute of Digital Twin. The authors would like to acknowledge the provision of the datasets by ISPRS and EuroSDR, released in conjunction with the ISPRS scientific initiatives 2014 and 2015, led by ISPRS ICWG I/Vb. We also thank the support of Wuhan University-Huawei Geoinformatics Innovation Laboratory.

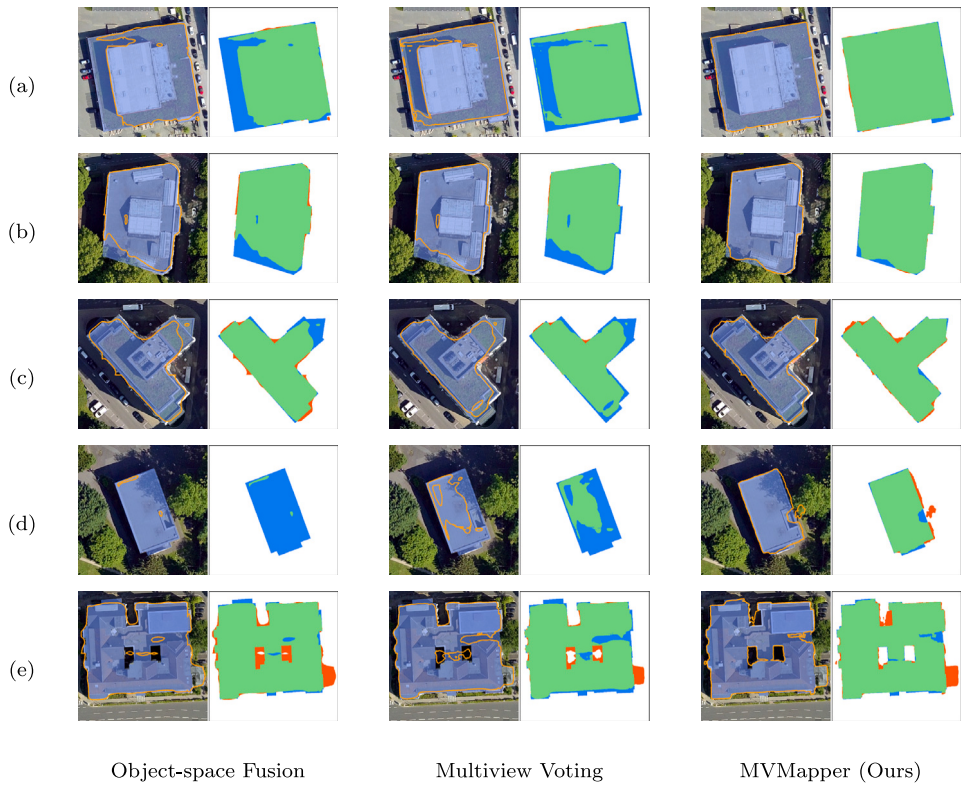


Fig. 18. Results of typical building instances in Dortmund experiment. The orange polygons represent the building contours generated by the models, while the ground truths are overlaid on the TDOM as blue transparent masks. Evaluation maps are color-coded, with true positives in green, false positives in red, and false negatives in blue.

References

[1] W. Zhang, G. Yang, N. Zhang, L. Xu, X. Wang, Y. Zhang, H. Zhang, J. Del Ser, V.H.C. De Albuquerque, Multi-task learning with multi-view weighted fusion attention for artery-specific calcification analysis, Inf. Fusion (ISSN: 15662535) 71 (2021) 64–76, http://dx.doi.org/10.1016/j.inffus.2021.01.009, URL https://linkinghub.elsevier.com/retrieve/pii/S1566253521000154.

[2] S. Jiang, W. Min, L. Liu, Z. Luo, Multi-scale multi-view deep feature aggregation for food recognition, IEEE Trans. Image Process. 29 (2020) 265–276, http://dx.doi.org/10.1109/TIP.2019.2929447 (ISSN 1057-7149, 1941-0042) URL https://ieeexplore.ieee.org/document/8779586.

[3] X. Ning, Z. Yu, L. Li, W. Li, P. Tiwari, DILF: Differentiable rendering-based multi-view image–language fusion for zero-shot 3D shape understanding, Inf. Fusion (ISSN: 15662535) 102 (2024) 102033, http://dx.doi.org/10.1016/j.inffus.2023.102033, URL https://linkinghub.elsevier.com/retrieve/pii/S1566253523003494.

[4] J.L. Awange, J.B. Kyalo Kiema, Fundamentals of photogrammetry, in: J.L. Awange, J.B. Kyalo Kiema (Eds.), Environmental Geoinformatics: Monitoring and Management, in: Environmental Science and Engineering, Springer, Berlin, Heidelberg, ISBN: 978-3-642-34085-7, 2013, pp. 157–174, http://dx.doi.org/10.1007/978-3-642-34085-7_11.

[5] J. Zhang, S. Xu, Y. Zhao, J. Sun, S. Xu, X. Zhang, Aerial orthoimage generation for UAV remote sensing: Review, Inf. Fusion (ISSN: 1566-2535) 89 (2023) 91–120, http://dx.doi.org/10.1016/j.inffus.2022.08.007, URL https://www.sciencedirect.com/science/article/pii/S1566253522000999.

[6] J. Li, X. Huang, L. Tu, T. Zhang, L. Wang, A review of building detection from very high resolution optical remote sensing images, GISci. Remote Sens. 59 (1) (2022) 1199–1225, http://dx.doi.org/10.1080/15481603.2022.2101727 (ISSN 1548-1603, 1943-7226), URL https://www.tandfonline.com/doi/full/10.1080/15481603.2022.2101727.

[7] Q. Chen, Y. Zhang, X. Li, P. Tao, Extracting rectified building footprints from traditional orthophotos: A new workflow, Sensors (ISSN: 1424-8220) 22 (1) (2021) 207, http://dx.doi.org/10.3390/s22010207, URL https://www.mdpi.com/1424-8220/22/1/207.

[8] S. Gui, R. Qin, Automated LoD-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto, ISPRS J. Photogramm. Remote Sens. (ISSN: 09242716) 181 (2021) 1–19, http://dx.doi.org/10.1016/j.isprsjprs.2021.08.025, URL https://linkinghub.elsevier.com/retrieve/pii/S0924271621002318.

[9] J. Zhou, Y. Liu, G. Nie, H. Cheng, X. Yang, X. Chen, L. Gross, Building extraction and floor area estimation at the village level in rural China via a comprehensive method integrating UAV photogrammetry and the novel EDSANet, Remote Sens. (ISSN: 2072-4292) 14 (20) (2022) 5175, http://dx.doi.org/10.3390/rs14205175, URL https://www.mdpi.com/2072-4292/14/20/5175, Number: 20 Publisher: Multidisciplinary Digital Publishing Institute.

[10] D. Yu, S. Ji, J. Liu, S. Wei, Automatic 3D building reconstruction from multi-view aerial images with deep learning, ISPRS J. Photogramm. Remote Sens. (ISSN: 09242716) 171 (2021) 155–170, http://dx.doi.org/10.1016/j.isprsjprs.2020.11.011, URL https://linkinghub.elsevier.com/retrieve/pii/S092427162030318X.

[11] C. Peng, H. Li, C. Tao, Y. Li, J. Ma, MSINet: Mining scale information from digital surface models for semantic segmentation of aerial images, Pattern Recognit. (ISSN: 0031-3203) 143 (2023) 109785, http://dx.doi.org/10.1016/j.patcog.2023.109785, URL https://www.sciencedirect.com/science/article/pii/S0031320323004831.

[12] W. Liu, M. Yang, M. Xie, Z. Guo, E. Li, L. Zhang, T. Pei, D. Wang, Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network, Remote Sens. (ISSN: 2072-4292) 11 (24) (2019) 2912, http://dx.doi.org/10.3390/rs11242912, URL https://www.mdpi.com/2072-4292/11/24/2912, Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

[13] Z. Rao, M. He, Z. Zhu, Y. Dai, R. He, Bidirectional guided attention network for 3-D semantic detection of remote sensing images, IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 59 (7) (2021) 6138–6153, http://dx.doi.org/10.1109/TGRS.2020.3029527, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[14] X. Zhang, Y. Chen, H. Zhang, S. Wang, J. Lu, J. Yang, When visual disparity generation meets semantic segmentation: A mutual encouragement approach, IEEE Trans. Intell. Transp. Syst. (ISSN: 1558-0016) 22 (3) (2021) 1853–1867, http://dx.doi.org/10.1109/TITS.2020.3027556, Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[15] F. Nex, M. Gerke, F. Remondino, H.-J. Przybilla, M. Bäumker, A. Zurhorst, ISPRS benchmark for multi-platform photogrammetry, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. (ISSN: 2194-9050) II-3/W4 (2015) 135–142, http://dx.doi.org/10.5194/isprsannals-II-3-W4-135-2015, URL https://isprs-annals.copernicus.org/articles/II-3-W4/135/2015/.

[16] H. Jung, H.-S. Choi, M. Kang, Boundary enhancement semantic segmentation for building extraction from remote sensed image, IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 60 (2022) 1–12, http://dx.doi.org/10.1109/TGRS.2021.3108781, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[17] T. Liu, L. Yao, J. Qin, N. Lu, H. Jiang, F. Zhang, C. Zhou, Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction, Int. J. Appl. Earth Obs. Geoinf. (ISSN: 1569-8432) 109 (2022) 102768, http://dx.doi.org/10.1016/j.jag.2022.102768, URL https://www.sciencedirect.com/science/article/pii/S0303243422000940.

[18] L. Wang, S. Fang, X. Meng, R. Li, Building extraction with vision transformer, IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 60 (2022) 1–11, http://dx.doi.org/10.1109/TGRS.2022.3186634, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[19] P. Song, J. Li, Z. An, H. Fan, L. Fan, CTMFNet: CNN and transformer multiscale fusion network of remote sensing urban scene imagery, IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 61 (2023) 1–14, http://dx.doi.org/10.1109/TGRS.2022.3232143, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[20] S. Zorzi, S. Bazrafkan, S. Habenschuss, F. Fraundorfer, PolyWorld: Polygonal building extraction with graph neural networks in satellite images, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, New Orleans, LA, USA, ISBN: 978-1-66546-946-3, 2022, pp. 1938–1947, http://dx.doi.org/10.1109/CVPR52688.2022.00189, URL https://ieeexplore.ieee.org/document/9880425/.

[21] Y. Hu, Z. Wang, Z. Huang, Y. Liu, PolyBuilding: Polygon transformer for building extraction, ISPRS J. Photogramm. Remote Sens. (ISSN: 0924-2716) 199 (2023) 15–27, http://dx.doi.org/10.1016/j.isprsjprs.2023.03.021, URL https://www.sciencedirect.com/science/article/pii/S0924271623000813.

[22] B. Xu, J. Xu, N. Xue, G.-S. Xia, HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision, ISPRS J. Photogramm. Remote Sens. (ISSN: 0924-2716) 198 (2023) 284–296, http://dx.doi.org/10.1016/j.isprsjprs.2023.03.006, URL https://www.sciencedirect.com/science/article/pii/S0924271623000667.

[23] A.D. Schlosser, G. Szabó, L. Bertalan, Z. Varga, P. Enyedi, S. Szabó, Building extraction using orthophotos and dense point cloud derived from visual band aerial imagery based on machine learning and segmentation, Remote Sens. (ISSN: 2072-4292) 12 (15) (2020) 2397, http://dx.doi.org/10.3390/rs12152397, URL https://www.mdpi.com/2072-4292/12/15/2397, Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.

[24] H.A.H. Al-Najjar, B. Kalantar, B. Pradhan, V. Saeidi, A.A. Halin, N. Ueda, S. Mansor, Land cover classification from fused DSM and UAV images using convolutional neural networks, Remote Sens. (ISSN: 2072-4292) 11 (12) (2019) 1461, http://dx.doi.org/10.3390/rs11121461, URL https://www.mdpi.com/2072-4292/11/12/1461.

[25] Z. Rao, M. He, Z. Zhu, Y. Dai, R. He, SDBF-net: Semantic and disparity bidirectional fusion network for 3D semantic detection on incidental satellite images, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, IEEE, Lanzhou, China, ISBN: 978-1-72813-248-8, 2019, pp. 438–444, http://dx.doi.org/10.1109/APSIPAASC47483.2019.9023223, URL https://ieeexplore.ieee.org/document/9023223/.

[26] H. Chen, M. Lin, H. Zhang, G. Yang, G.-S. Xia, X. Zheng, L. Zhang, Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, (ISSN: 2153-7003) 2019, pp. 4967–4970, http://dx.doi.org/10.1109/IGARSS.2019.8899306.

[27] N. Lu, Y. Wu, H. Zheng, X. Yao, Y. Zhu, W. Cao, T. Cheng, An assessment of multi-view spectral information from UAV-based color-infrared images for improved estimation of nitrogen nutrition status in winter wheat, Precis. Agric. (ISSN: 1573-1618) 23 (5) (2022) 1653–1674, http://dx.doi.org/10.1007/s11119-022-09901-7.

[28] X. Huang, S. Li, J. Li, X. Jia, J. Li, X.X. Zhu, J.A. Benediktsson, A multispectral and multiangle 3-D convolutional neural network for the classification of ZY-3 satellite images over urban areas, IEEE Trans. Geosci. Remote Sens. 59 (12) (2021) 10266–10285, http://dx.doi.org/10.1109/TGRS.2020.3037211 (ISSN 0196-2892, 1558-0644) URL https://ieeexplore.ieee.org/document/9266127.

[29] T. Liu, A. Abd-Elrahman, Multi-view object-based classification of wetland land covers using unmanned aircraft system images, Remote Sens. Environ. (ISSN: 00344257) 216 (2018) 122–138, http://dx.doi.org/10.1016/j.rse.2018.06.043, URL https://linkinghub.elsevier.com/retrieve/pii/S0034425718303237.

[30] Q. Hu, W. Woldt, C. Neale, Y. Zhou, J. Drahota, D. Varner, A. Bishop, T. LaGrange, L. Zhang, Z. Tang, Utilizing unsupervised learning, multi-view imaging, and CNN-based attention facilitates cost-effective wetland mapping, Remote Sens. Environ. (ISSN: 00344257) 267 (2021) 112757, http://dx.doi.org/10.1016/j.rse.2021.112757, URL https://linkinghub.elsevier.com/retrieve/pii/S0034425721004776.

[31] F. Kurz, S. Azimi, C.-Y. Sheu, P. d'Angelo, Deep learning segmentation and 3D reconstruction of road markings using multiview aerial imagery, ISPRS Int. J. Geo-Inf. (ISSN: 2220-9964) 8 (1) (2019) 47, http://dx.doi.org/10.3390/ijgi8010047, URL http://www.mdpi.com/2220-9964/8/1/47.

[32] Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning, IEEE Trans. Knowl. Data Eng. 31 (10) (2019) 1863–1883, http://dx.doi.org/10.1109/TKDE.2018.2872063 (ISSN 1041-4347, 1558-2191, 2326-3865), URL https://ieeexplore.ieee.org/document/8471216.

[33] G. Chao, S. Sun, J. Bi, A survey on multiview clustering, IEEE Trans. Artif. Intell. (ISSN: 2691-4581) 2 (2) (2021) 146–168, http://dx.doi.org/10.1109/TAI.2021.3065894, URL https://ieeexplore.ieee.org/document/9395530/.

[34] X. Zhang, H. Gao, G. Li, J. Zhao, J. Huo, J. Yin, Y. Liu, L. Zheng, Multi-view clustering based on graph-regularized nonnegative matrix factorization for object recognition, Inform. Sci. (ISSN: 00200255) 432 (2018) 463–478, http://dx.doi.org/10.1016/j.ins.2017.11.038, URL https://linkinghub.elsevier.com/retrieve/pii/S0020025517311015.

[35] Y. Qin, G. Feng, Y. Ren, X. Zhang, Consistency-induced multiview subspace clustering, IEEE Trans. Cybern. 53 (2) (2023) 832–844, http://dx.doi.org/10.1109/TCYB.2022.3165550 (ISSN 2168-2267, 2168-2275), URL https://ieeexplore.ieee.org/document/9764657.

[36] Y. Qin, C. Qin, X. Zhang, D. Qi, G. Feng, NIM-nets: Noise-aware incomplete multiview learning networks, IEEE Trans. Image Process. 32 (2023) 175–189, http://dx.doi.org/10.1109/TIP.2022.3226408 (ISSN 1057-7149, 1941-0042), URL https://ieeexplore.ieee.org/document/9975265.

[37] D. Robert, B. Vallet, L. Landrieu, Learning multi-view aggregation in the wild for large-scale 3D semantic segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, New Orleans, LA, USA, ISBN: 978-1-66546-946-3, 2022, pp. 5565–5574, http://dx.doi.org/10.1109/CVPR52688.2022.00549, URL https://ieeexplore.ieee.org/document/9879920/.

[38] Z. Qi, H. Chen, C. Liu, Z. Shi, Z. Zou, Implicit ray transformers for multiview remote sensing image segmentation, IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 61 (2023) 1–15, http://dx.doi.org/10.1109/TGRS.2023.3285659, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[39] Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan, MVSNet: Depth inference for unstructured multi-view stereo, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, in: Lecture Notes in Computer Science, vol. 11212, Springer International Publishing, Cham, 2018, pp. 785–801, http://dx.doi.org/10.1007/978-3-030-01237-3_47, (ISBN 978-3-030-01236-6 978-3-030-01237-3), URL https://link.springer.com/10.1007/978-3-030-01237-3_47.

[40] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, P. Tan, Cascade cost volume for high-resolution multi-view stereo and stereo matching, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Seattle, WA, USA, ISBN: 978-1-72817-168-5, 2020, pp. 2492–2501, http://dx.doi.org/10.1109/CVPR42600.2020.00257, URL https://ieeexplore.ieee.org/document/9157551/.

[41] T. Toutin, Review article: Geometric processing of remote sensing images: models, algorithms and methods, Int. J. Remote Sens. (ISSN: 0143-1161, 1366-5901) 25 (10) (2004) 1893–1924, http://dx.doi.org/10.1080/0143116031000101611, URL https://www.tandfonline.com/doi/full/10.1080/0143116031000101611.

[42] J. Gao, J. Liu, S. Ji, Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, Montreal, QC, Canada, ISBN: 978-1-66542-812-5, 2021, pp. 6128–6137, http://dx.doi.org/10.1109/ICCV48922.2021.00609, URL https://ieeexplore.ieee.org/document/9711432/.

[43] V. Arevalo, J. Gonzalez, Improving piecewise linear registration of high-resolution satellite images through mesh optimization, IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 46 (11) (2008) 3792–3803, http://dx.doi.org/10.1109/TGRS.2008.924003, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[44] Q. Chen, S. Wang, B. Wang, M. Sun, Automatic registration method for fusion of ZY-1-02C satellite images, Remote Sens. (ISSN: 2072-4292) 6 (1) (2013) 157–179, http://dx.doi.org/10.3390/rs6010157, URL http://www.mdpi.com/2072-4292/6/1/157.

[45] Han, Kim, Yeom, Improved piecewise linear transformation for precise warping of very-high-resolution remote sensing images, Remote Sens. (ISSN: 2072-4292) 11 (19) (2019) 2235, http://dx.doi.org/10.3390/rs11192235, URL https://www.mdpi.com/2072-4292/11/19/2235.

[46] H. Guo, H. Xu, Y. Wei, Y. Shen, X. Rui, Outlier removal and feature point pairs optimization for piecewise linear transformation in the co-registration of very high-resolution optical remote sensing imagery, ISPRS J. Photogramm. Remote Sens. (ISSN: 09242716) 193 (2022) 299–313, http://dx.doi.org/10.1016/j.isprsjprs.2022.09.008, URL https://linkinghub.elsevier.com/retrieve/pii/S0924271622002520.

[47] S. Suzuki, K. be, Topological structural analysis of digitized binary images by border following, Comput. Vis. Graph. Image Process. (ISSN: 0734-189X) 30 (1) (1985) 32–46, http://dx.doi.org/10.1016/0734-189X(85)90016-7, URL https://www.sciencedirect.com/science/article/pii/0734189X85900167.

[48] Y. Zhang, S. Zou, X. Liu, X. Huang, Y. Wan, Y. Yao, LiDAR-guided stereo matching with a spatial consistency constraint, ISPRS J. Photogramm. Remote Sens. (ISSN: 09242716) 183 (2022) 164–177, http://dx.doi.org/10.1016/j.isprsjprs.2021.11.003, URL https://linkinghub.elsevier.com/retrieve/pii/S0924271621002951.

[49] S. Zou, X. Liu, X. Huang, Y. Zhang, S. Wang, S. Wu, Z. Zheng, B. Liu, Edge-preserving stereo matching using LiDAR points and image line features, IEEE Geosci. Remote Sens. Lett. 20 (2023) 1–5, http://dx.doi.org/10.1109/LGRS.2023.3239030 (ISSN 1545-598X, 1558-0571), URL https://ieeexplore.ieee.org/document/10024831.

[50] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3349–3364, http://dx.doi.org/10.1109/TPAMI.2020.2983686 (ISSN 0162-8828, 2160-9292, 1939-3539), URL https://ieeexplore.ieee.org/document/9052469.

[51] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019, arXiv:1711.05101 [cs, math].

[52] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, IEEE, Stanford, CA, USA, ISBN: 978-1-5090-5407-7, 2016, pp. 565–571, http://dx.doi.org/10.1109/3DV.2016.79, URL http://ieeexplore.ieee.org/document/7785132/.

[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection.

[54] P. Jaccard, The distribution of the flora in the alpine zone.1, New Phytol. (ISSN: 1469-8137) 11 (2) (1912) 37–50, http://dx.doi.org/10.1111/j.1469-8137.1912.tb05611.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x.

[55] Y. Sasaki, The truth of the F-measure, Teach Tutor. Mater. 1 (5) (2007) 1–5.

[56] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Las Vegas, NV, ISBN: 978-1-4673-8851-1, 2016, pp. 724–732, http://dx.doi.org/10.1109/CVPR.2016.85, URL https://ieeexplore.ieee.org/document/7780454/.

[57] F. Wilcoxon, Your use of the JSTOR archive indicates your acceptance of JSTOR's terms and conditions of use, available at., Biometrics 1 (6) (1945) 80–83.