Building Footprint Generation Through Convolutional Neural Networks With Attraction Field Representation

Qingyu Li, *Student Member, IEEE*, Lichao Mou, Yuansheng Hua, *Graduate Student Member, IEEE*, Yilei Shi, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Building footprint generation is a vital task in a wide range of applications, including, to name a few, land use management, urban planning and monitoring, and geographical database updating. Most existing approaches addressing this problem fall back on convolutional neural networks (CNNs) to learn semantic masks of buildings. However, one limitation of their results is blurred building boundaries. To address this, we propose to learn attraction field representation for building boundaries, which is capable of providing an enhanced representation power. Our method comprises two elemental modules: an Img2AFM module and an AFM2Mask module. More specifically, the former aims at learning an attraction field representation conditioned on an input image, which is capable of enhancing building boundaries and suppressing the background. The latter module predicts segmentation masks of buildings using the learned attraction field map. The proposed method is evaluated on three datasets with different spatial resolutions: the ISPRS dataset, the INRIA dataset, and the Planet dataset. From experimental results, we find that the proposed framework can well preserve geometric shapes and sharp boundaries of buildings, which brings significant improvements over other competitors. The trained model and code are available at https://github.com/lqycrystal/AFM_building.**

***Index Terms*— Attraction field map (AFM), building footprint, convolutional neural network (CNN), semantic segmentation.**

Manuscript received February 16, 2021; revised June 21, 2021; accepted August 19, 2021. Date of publication September 15, 2021; date of current version January 31, 2022. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement ERC-2016-StG-714087 (*So2Sat*), in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01 [Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)], in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the project "Investigation of Building Cases Using AI" funded by the Bavarian State Ministry of Finance and Regional Identity (StMFH) and the Bavarian Agency for Digitization, High-Speed Internet and Surveying. *(Corresponding author: Xiao Xiang Zhu.)*

Qingyu Li, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu are with Data Science in Earth Observation, Technische Universität München (TUM), 80333 Munich, Germany, and also with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: qingyu.li@tum.de; lichao.mou@dlr.de; yuansheng.hua@dlr.de; xiaoxiang.zhu@dlr.de).

Yilei Shi is with the Chair of Remote Sensing Technology, Technische Universität München (TUM), 80333 Munich, Germany (e-mail: yilei.shi@tum.de).

Digital Object Identifier 10.1109/TGRS.2021.3109844

I. Introduction

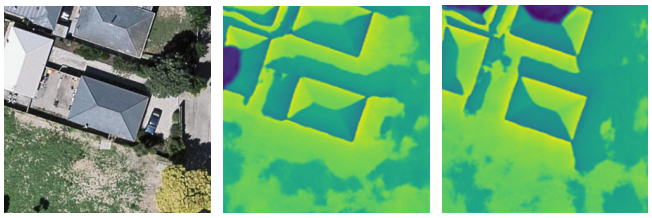
AUTOMATIC building footprint generation from remote sensing data has been of great interest in the community for a range of applications, such as land use management, urban planning and monitoring, and disaster management. However, accurate and reliable building footprint generation remains particularly challenging due to two reasons. On the one hand, different materials and structures lead to large variations of buildings in terms of color, shape, size, and texture. On the other hand, buildings and other man-made objects (e.g., roads and sidewalks) share similar spectral signatures, which can result in a low between-class variability.

Early efforts have been gone into seeking out hand-crafted features of being to effectively exploit spectral, structural, and context information. For example, Huang *et al.* [1] propose a framework for automatic building extraction, which utilizes spectral, geometrical, and contextual features extracted from imagery. Nonetheless, these methods still fail to satisfy accuracy requirements because they rely on a heuristic feature design procedure and usually have poor generalization capabilities.

More recently, convolutional neural networks (CNNs) have surpassed traditional methods in many remote sensing tasks [2]–[10]. CNNs can directly learn feature representations from the raw data; thus, they provide an end-to-end solution to generate building footprints from remote sensing data. Most of the studies in this field assign a label "building" or "non-building" to every pixel in the image, thus yielding semantic masks of buildings. The existing CNNs seem to be able to deliver very promising segmentation results for the purpose of building footprint generation at a large scale (cf. Fig. 1). However, when we zoom in on some segmentation masks (see results from U-Net [11] in Fig. 1), it can be clearly seen that such results are not that perfect, and the boundaries of some buildings are blurred.

We have observed that buildings usually have clear patterns (e.g., corners and straight lines). Therefore, geometric primitives of buildings can be exploited as the most distinguishable features for extraction purposes. There have been several works based on this idea [12]–[15]. In this work, we want to exploit building boundaries as a primary visual cue to achieve our task.

Fig. 1. Building footprints generated by U-Net [11] and our proposed method (U-Net with attraction field representation) at large scale and two zoomed in areas.



(a) (b) (c)

Fig. 2. (a) Satellite imagery, and the AFMs in both (b) x- and (c) y-directions estimated by our method.

Recently, attraction field representation is used for the task of line segment detection in computer vision [16], which seeks the most attractive line segment for each pixel. Our observation is that, when building boundaries in remote sensing images are represented by the attraction field, they can be greatly enhanced, while background clutters (e.g., car, courtyard, and road) are suppressed. Fig. 2 shows an example. Motivated by this observation, in this work, we want to make use of the attraction field to represent buildings and propose an end-to-end trainable network for automatic building footprint generation. This network consists of two modules: Img2AFM and AFM2Mask. The former takes as input an image and is responsible for learning a corresponding attraction field map (AFM) using a CNN. By doing so, fine-grained building boundaries can be preserved, and the impact of background clutters can be alleviated. The latter module learns another subnetwork to obtain semantic masks of buildings from augmented building edges in the learned AFM. Note that both these two modules are jointly optimized. In addition, the AFM2Mask module is flexible enough to use different semantic segmentation network architectures.

This work's contributions are threefold.

1) We propose to use the boundary-aware attraction field to represent building footprints in remote sensing images. This helps to enhance building boundaries while suppressing the impact of background clutters. To the best of our knowledge, it is the first work that utilizes the attraction field for the task of building footprint generation.
2) We propose a novel network that first learns an AFM by a subnetwork, termed Img2AFM, and then uses another subnetwork called AFM2Mask to reconstruct segmentation masks of buildings. These two modules are trained in an end-to-end fashion.
3) The proposed framework obtains satisfactory performance on three datasets with different spatial resolutions, including ISPRS, INRIA, and Planet datasets. Compared with naive semantic segmentation networks and networks with other visual cues (e.g., building boundary maps), our method can significantly improve accuracies in terms of both semantic mask and boundary.

The remainder of this article is organized as follows. Related work is reviewed in Section II. Section III details the proposed framework for building footprint generation. The experiments are described in Section IV. Results and discussion are provided in Section V. Eventually, Section VI summarizes this work.

II. Related Work

There are a significant number of studies working on building footprint generation from remote sensing imagery. According to used visual cues, they can be categorized into three classes: semantic mask, corner, and boundary of the building.

A. Building Footprint Generation Based on the Semantic Mask

Most methods for building footprint generation involve learning semantic masks of buildings from remote sensing imagery. Early efforts include segmentation-, classification-, and index-based methods. The segmentation-based methods extract buildings using image segmentation algorithms. For example, based on a two-level graph theory, Ok [17] proposes a segmentation approach to identify building regions. For classification-based methods, building masks are extracted by machine-learning classifiers which take spectral information and/or spatial features as input to make a prediction for each pixel. For instance, Turker and Koc-San [18] utilize a support vector machine (SVM) to identify building regions based on spectral bands and the normalized difference vegetation index (NDVI). The objective of index-based approaches is to design a feature index that can be directly applied to obtain building regions without any classification or segmentation process. Morphological building index (MBI) [19] is a widely used one, and this index integrates multiscale and multidirectional morphological operators. However, a general limitation of these early works is the use of handcrafted features and complex feature engineering, which leads to a poor generalization.

Instead of the heuristic design of features, CNNs can offer a better generalization capability. Driven by recent advances in semantic segmentation networks, results of building footprint generation have been significantly improved. These networks are usually fully convolutional network (FCN) [20] and encoder–decoder architecture, such as U-Net [11], SegNet [21], and FC-DenseNet [22]. In [23], FCN has been demonstrated to be effective in processing large amounts of remote sensing data and providing reliable building segmentation results. SegNet is used in [24] to

generate the first seamless building footprint map for the United States. In order to improve the accuracy of segmenting large buildings, a U-Net-based architecture is proposed in [25], where original images and their downsampled counterparts are taken as inputs of two branches sharing the same weights. In [26], an adversarial training strategy is proposed for building extraction from remote sensing imagery, and FC-DenseNet is exploited as a base semantic segmentation network to generate accurate building footprints.

However, many experiments show that predicted semantic masks of buildings from CNNs are still not that satisfactory, where building boundaries are blurred. In this regard, signed-distance transform (SDT) [27] is proposed to represent building footprints. The signed-distance function value is derived as the distance from a pixel to its closest point on a building boundary; positive values indicate the interior of a building and negative values otherwise. Then, the learning problem of the SDT representation can be regarded as a multiclass classification problem, which categorizes signed-distance values into a certain number of classes [24]. Compared to the widely used binary building mask, SDT can encode more fine-grained information for network learning.

B. Building Footprint Generation Based on the Corner

Some algorithms generate building footprints based on geometrical primitives, such as building corners. In these methods, geometric primitives are first detected and then grouped together to reconstruct individual building hypotheses. A building corner refers to a point with its local neighborhoods in two varying line segment directions and is invariant to translation, rotation, and illumination [28]. Early studies extract building corners with the help of some point feature operators, such as Harris corner detector [29] and scale-invariant feature transform (SIFT) operator [30]. Cote and Saeedi [12] and Zangrandi *et al.* [31] employ a Harris corner detector to extract corner points of buildings. Afterward, these detected corner points are connected in the order of their polar angles with respect to building central markers. By doing so, polygonal representations of buildings can be constructed. In [32], SIFT is exploited to extract corners that are regarded as seed points to estimate rectangle shapes of buildings with a region growing method.

With the development of keypoint detection networks, several novel studies propose to delineate building footprints by detecting corner points using CNNs. PolyMapper [33] extracts corner points with a CNN in the first stage and then connects them by a recurrent neural network (RNN) to realize closed polygon representations of individual buildings. The other research [34] utilizes the same pipeline as PolyMapper [33], and various blocks are integrated to enhance the feature extraction and object detection modules. Another method [13] also exploits a CNN to detect corners but adopts a fully geometric-based grouping strategy without any deep feature learning. Recently, Girard's method [35] proposes to learn a frame field output instead of building corners. The frame field is regarded as a geometric feature that can help to improve the segmentation of buildings.

C. Building Footprint Generation Based on the Boundary

Building boundary is another commonly used geometric primitive and can be taken as a primary visual cue to generate building footprints. Early works extract building boundaries from remote sensing data in two steps. Given that lines are strongly relevant to building boundaries, the first step is to detect line segments. Afterward, the extracted lines are grouped to form closed boundaries for individual buildings. A commonly used line detection algorithm is the Hough transformation [36] that utilizes a voting procedure to find straight lines in parameter space. Compared with the Hough transformation, the Burns algorithm [37] only uses gradient orientations and, therefore, requires a relatively lower computation cost. In [14] and [38], line segment sets are extracted with the Hough transformation and the Burns algorithm. Then, intersection nodes of the two line segment sets are employed to build a structural graph. Finally, building boundaries are identified with a graph search algorithm. However, both Hough transformation and Burns algorithm highly depend on parameter settings and have a very high false alarm rate. In this regard, EDLines [39] are proposed to avoid parameter tuning. Moreover, it has a faster computation speed and a lower false alarm rate. In [40] and [41], EDLines are, therefore, adopted for the automatic extraction of line segments, but they make use of different strategies to group these line segments.

These early works still encounter issues when dealing with more complex building shapes and large-scale applications. Considering that, nowadays, CNNs are the de facto leading approach for building footprint generation tasks, two novel works, [15] and [42], propose to learn building boundaries in their end-to-end CNNs. Marcos *et al.* [15] present a method termed deep structured active contours (DSACs), which learns active contour model (ACM) [43] parameterizations per instance using a CNN. Although DSAC improves geometric correctness, results are still not that satisfactory, e.g., there exist blob-like shapes and some self-intersections of building. Besides, the representation of boundary points in DSAC adopts Euclidean coordinates, which leads to extra computational overheads during energy minimization. On this point, another research [42] proposes to use polar coordinates, as this can not only simplify the energy function but also prevent self-intersection. However, these two methods still have two limitations. On the one hand, the initialization of them relies on external methods that are not included in an end-to-end learning process. On the other hand, their results are promising only in very high-resolution remote sensing images where strong geometric priors exist.

III. METHODOLOGY

In this work, we explicitly take building boundaries as a primary visual cue. By doing so, building footprint generation tasks can be benefited from the precise delineation of building boundaries. In this section, an overview of the proposed approach is first presented. Then, two key modules, Img2AFM and AFM2Mask, are introduced in detail, respectively. Finally, the method of integrating and jointly optimizing the two modules in an end-to-end architecture is described.

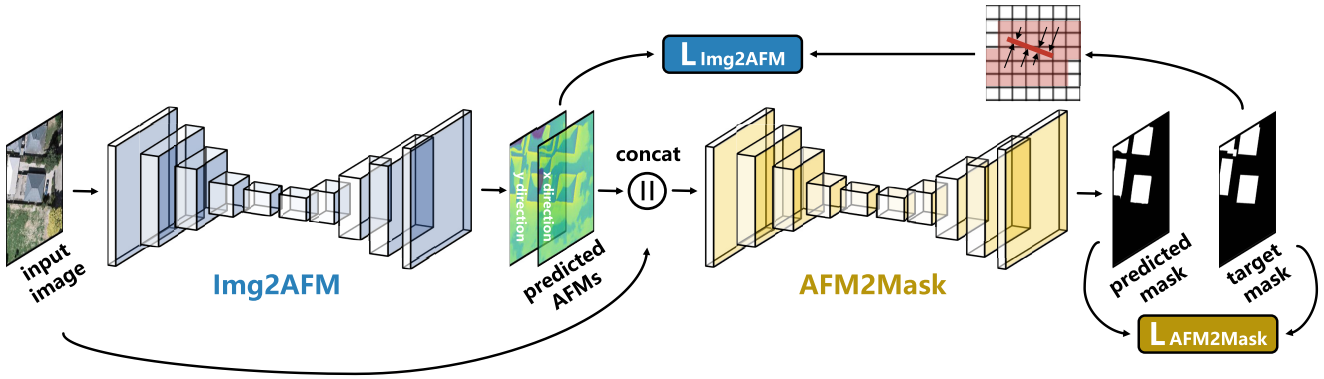


Fig. 3. Overview of the proposed framework. The Img2AFM module takes an image as input and outputs two AFMs in x- and y-directions. Afterward, the output is then fed into the AFM2Mask module along with the input image to generate a building mask. Notable that these two modules are trained in an end-to-end fashion.

A. Overview

As shown in Fig. 3, the proposed method consists of two modules. The Img2AFM module exploits a U-Net architecture to learn the attraction field representation, which can enhance building boundaries and suppress background clutters. It takes an image as input and outputs two AFMs in x- and y-directions. Afterward, the output is then fed into the AFM2Mask module along with the input image to generate a building mask. Moreover, the AFM2Mask module is very flexible to utilize different semantic segmentation networks. Note that these two modules can be integrated into an end-to-end framework and optimized jointly. In this way, the optimal output can be obtained by the coadaptation of these two modules.

B. Img2AFM Module

1) Definition of Attraction Field Map: An image I can be regarded a lattice. Let $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ be the set of building line segments in the image lattice with n being the number of building line segments. A building line segment $\mathbf{e}_i$ is represented by two end points $\mathbf{p}_i^a$ and $\mathbf{p}_i^b$. For the sake of simplicity, the set E is named boundary map in our case. The boundary map characterizing all building boundaries in the ground reference is shown in Fig. 4(c).

For each pixel, we try to find its most "attractive" building line segment that is the closest to it. Following this criterion, a region partition map R is formed by partitioning all pixels into n regions and assigning each pixel $\mathbf{x} \in I$ to its closest building line segment. R_i denotes a region for the building line segment $\mathbf{e}_i$ in E. Specifically, in order to derive the distance between a pixel $\mathbf{x}$ and a building line segment $\mathbf{e}_i$, the pixel $\mathbf{x}$ is first projected to the straight line passing through $\mathbf{e}_i$. If the projection point is not on $\mathbf{e}_i$, the nearest endpoint is utilized as the projection point. The definition of the projection point $\mathbf{p}'$ is

$$\mathbf{p}' = \mathbf{p}_i^a + c_x \cdot \left(\mathbf{p}_i^b - \mathbf{p}_i^a\right). \tag{1}$$

When $c_x \in (0, 1)$, $\mathbf{p}'$ belongs to the original point-to-line projection, and if $c_x = 0$ or 1, $\mathbf{p}'$ is its nearest endpoint of $\mathbf{e}_i$.

Then, the distance $d(\mathbf{x}, \mathbf{e}_i)$ between $\mathbf{x}$ and $\mathbf{e}_i$ can be defined as the Euclidean distance between the pixel and the projection point. Then, R_i in the image lattice for $\mathbf{e}_i$ can be defined as

$$R_i = \left\{\mathbf{x} \mid \mathbf{x} \in I; d(\mathbf{x}, \mathbf{e}_i) < d\left(\mathbf{x}, \mathbf{e}_j\right) \forall j \neq i, \mathbf{e}_j \in E\right\}. \tag{2}$$

It should be noted that $R_i \cap R_j = \emptyset$ and $\cup_{i=1}^n R_i = R$. Fig. 4 shows an example that, for the green building line segment, its corresponding region partition map is highlighted in green.

Afterward, the geometric property of a building line segment can be characterized by a 2-D attraction of all individual pixels in R_i. For instance, the attraction function of the pixel $\mathbf{x}$ in R_i is defined as

$$\mathbf{a}_i(\mathbf{x}) = \mathbf{p}' - \mathbf{x}. \tag{3}$$

When $c_x \in (0, 1)$, the attraction vector is perpendicular to the line segment. Fig. 4(d) shows the attraction vectors of the green line segment.

Finally, by enumerating (3) over all pixels in I, the AFM A with respect to E can be obtained as follows:

$$A = \{\mathbf{a}(\mathbf{x}) \mid \mathbf{x} \in I\}. \tag{4}$$

The superiority of AFM lies in two aspects compared with the boundary map used in previous studies (see [15] and [42]). One is that the geometry of boundaries can be depicted more precisely by the AFM, while the boundary map is only characterized by few pixels. Thus, directly learning boundary maps can lead to a zig-zag effect that results from the extreme imbalance between the number of boundary pixels and that of nonboundary pixels. The other benefit is that the AFM associates each line segment with a support region, which avoids the blurring effect.

2) Learning Attraction Field Map: Each pixel in the attraction field representation has two components (x- and y-directions) that are represented by attraction vectors from it to its projection point. In this respect, an attraction field representation can be regarded as a 2-D feature map, which is feasible to be learned by a network. Hence, in this article, we view the learning of the AFM as a dense prediction problem and solve it using a semantic segmentation network architecture. Among all semantic segmentation networks, U-Net

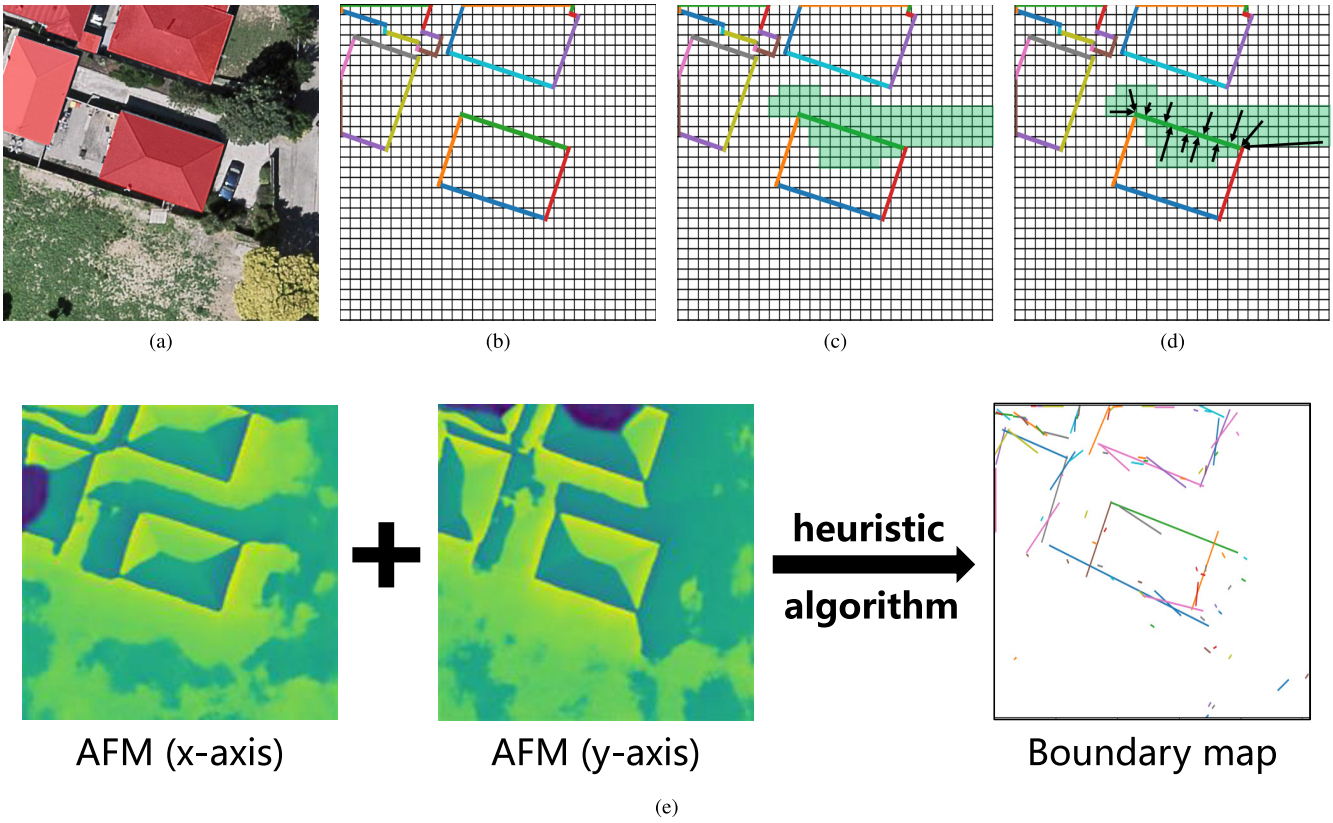


Fig. 4. (a) and (b) Semantic masks and boundaries of buildings in an image. (c) and (d) Region partition map and attraction vectors of the green building line segmentaccording to the method in [16]. (e) Recovered boundary map obtained by the heuristic algorithm in [16].

is more favorable than others for this task. Because learning the attraction field representation relies heavily on low-level visual cues (e.g., object edges) that exist in lower layers, and multiscale skip connections of U-Net are able to effectively use such information. In fact, in our experiments, we found that taking other network architectures as the Img2AFM module fails.

C. AFM2Mask Module

By learning the AFM, a representation encoding building boundaries can be obtained. Then, we need to remap the learned AFM into building masks. In [16], a heuristic algorithm has been proposed to recover line segments from the AFM. In this heuristic algorithm, attraction vectors are rearranged mathematically to generate a proposal map of line segments, and final line segments are then extracted with a greedy grouping strategy. However, we found that, in our building footprint generation task, the recovered boundary map from this algorithm is not satisfactory [cf. Fig. 4(e)] since there is a relatively high false alarm rate [see short line segments in Fig. 4(e)]. The reason is that predicted attraction vectors from CNNs are not mathematically precise enough. In this case, some potential outliers have been included in the following heuristic method, which leads to inaccurate line segment detections. Another reason is that this heuristic algorithm is not robust to imprecise estimates of the AFM. Furthermore, it requires a set of heuristics and makes the whole process inefficient. Therefore, in this work, we propose to learn this process, i.e., recovering building masks from the learned AFM, using a network. By doing so, the whole process can be trained in an end-to-end manner, which makes it more efficient and robust.

In the AFM2Mask module, the input image and learned attraction field representation from the previous module are concatenated as the input to this module. Afterward, the network can directly generate building masks without using math heuristics (that do not work well in our case). It is noteworthy that different semantic segmentation network architectures are quite flexible to be utilized in this module, which makes full use of the power of state-of-the-art networks to generate building footprint maps.

D. End-to-End Network Learning

We propose an end-to-end training pipeline for the supervised learning of our network. More specifically, the Img2AFM module is appended before the AFM2Mask module, and the two modules are jointly trained by minimizing a global loss function. The global loss function L is defined as follows:

$$L = L_{\text{Img2AFM}} + \lambda \cdot L_{\text{AFM2Mask}} \tag{5}$$

where L_{Img2AFM} and L_{AFM2Mask} are two loss functions for optimizing the Img2AFM and AFM2Mask modules, respectively. λ is a hyperparameter to introduce a weight on the second loss and can model the relative importance of two modules.

For the first term, an L_1 loss function is utilized, and it is defined as follows:

$$L_{\text{Img2AFM}} = \sum_{\mathbf{x}\in I} ||\hat{\mathbf{a}}(\mathbf{x}) - \mathbf{a}(\mathbf{x})||_1 \tag{6}$$

where $\hat{\mathbf{a}}(\mathbf{x})$ is the predicted AFM and $\mathbf{a}(\mathbf{x})$ is ground reference AFM for the input image.

For the AFM2Mask module, we make use of a cross entropy loss function to guide the learning. L_{AFM2Mask} is defined as

$$L_{\text{AFM2Mask}} = \sum_{x\in I} \begin{cases} -\log(f(x)) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{if } y = 0 \end{cases} \tag{7}$$

where y is the ground truth of pixel x, $y = 1$ denotes building, and $y = 0$ represents non-building. $f(x) \in [0, 1]$ is the output probability value of x.

In the backward propagation, L_{AFM2Mask} is first backpropagated through the AFM2Mask module and then together with $\lambda \cdot L_{\text{Img2AFM}}$ propagated backward through the Img2AFM module.

IV. Experiments

A. Dataset

We validate the proposed method on three datasets with different spatial resolutions, i.e., the ISPRS dataset, the INRIA dataset, and the Planet dataset.

1) ISPRS Dataset: The ISPRS dataset [44] is a benchmark dataset consisting of 38 tiles of aerial imagery over the city of Potsdam [cf. Fig. 5(a)]. Each aerial imagery includes 6000 × 6000 pixels at a spatial resolution of 5 cm/pixel. The provided ground reference has six land cover classes. In this work, we only use RGB bands of aerial images, and for the ground reference, the class of building is a positive class, while the other five categories are viewed as the class of non-building. Following the training/validation/test split in [45], 20 tiles (tile id: 2-10, 2-12, 3-10, 3-11, 3-12, 4-11, 4-12, 5-10, 5-11, 6-7, 6-8, 6-9, 6-10, 6-11, 6-12, 7-7, 7-9, 7-10, 7-11, and 7-12) are used for training, four tiles (tile id: 7-8, 4-10, 2-11, and 5-11) are for validation, and the remaining 14 tiles are used to test models.

2) INRIA Dataset: The INRIA dataset [46] is composed of 360 large-scale aerial images that are collected over ten different cities. The size of each imagery is 5000 × 5000, and each image consists of three bands (RGB) at a spatial resolution of 30 cm/pixel. A sample aerial image is showed in Fig. 5(b). The ground reference data of this dataset provide building masks but are only publicly available for five cities (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna). In this article, data are split into training and test sets according to the setup in [46] and [47]. For each city, images with ids 1–5 are used for validation, and the remaining 31 images are for training. The statistics are derived from the validation set.

3) Planet Dataset: In addition to the aforementioned two public datasets, we create a Planet dataset by collecting PlanetScope satellite images and their corresponding building footprints from OpenStreetMap. The PlanetScope satellite images are gathered from eight European cities (Munich, Berlin, Amsterdam, Paris, Cologne, Milan, Rome, and Zurich) with three bands (RGB) at 3-m spatial resolution. Compared to the former two datasets, the Planet dataset is more challenging due to its coarser spatial resolution. Fig. 5(c) shows an example of Munich. In our experiment, the image of Munich is used as the test set to evaluate the performance of models. The remaining seven cities are utilized as training and validation

(a)

(b)

(c)

Fig. 5. (a) Aerial imagery in the ISPRS dataset (spatial resolution: 5 cm/pixel). (b) Aerial image in the INRIA dataset (spatial resolution: 30 cm/pixel). (c) Satellite imagery in the Planet dataset (spatial resolution: 3 m/pixel).

sets. Specifically, for each city, 80% of samples are used for training, while 20% of samples are for validation purposes.

B. Experiment Setup

Our proposed model consists of two modules in an end-to-end framework, where the Img2AFM module utilizes a U-Net to learn the attraction field representation of an image with respect to building edges, and the AFM2Mask module can learn building masks from the representation using different semantic segmentation networks. To explore the flexibility of the AFM2Mask module, we select four state-of-the-art semantic segmentation networks: FCN-8s [20], SegNet [21], U-Net [11], and FC-DenseNet [22]. The attraction field representation encodes the geometric relation between pixels and building boundaries in an image, and it can be considered as a variant of distance transform, such as SDT [27] that measures the distance from the pixel to the boundary. Hence, we compare our model with existing works [24], [27] learning SDT representations of buildings. On the other hand, it is clearly seen that the learned AFMs from the Img2AFM module can well enhance building boundaries. In this aspect, the function of the attraction field representation seems to be similar to other visual cues, such as building boundaries and SDT masks. Thus, we also compare our network with two methods, SDT-recursive and boundary-recursive, where, basically, we incorporate SDT/edge learning into the proposed framework (cf. Fig. 3). Comparing the proposed approach and the two models can verify whether the attraction field representation is effective. Besides, the sensitivity of the hyper-parameter λ, being the coefficient of loss of the AFM2Mask module, is investigated.

C. Training Details

Our experiments are conducted within a Pytorch framework on an NVIDIA Tesla P100 GPU with 16 GB of memory. For the model training, remote sensing images and their corresponding ground reference building masks are cropped into small patches with a size of 256×256 pixels. Afterward, the boundaries, SDT, and AFMs are generated from the ground-truth building masks for further experiments as a ground reference in the training set. All models are trained for 100 epochs, and the optimizer is stochastic gradient descent (SGD) with a learning rate of 0.00001. The training batch size of all models is set as 4. The cross-entropy function is used as the loss function for other competitors.

The configurations of competitors included in experiments are listed as follows.

1) FCN-8s adopts a VGG16 architecture [48] as the backbone.
2) The encoder in SegNet is based on VGG16, and the decoder utilizes a reversed VGG16 architecture.
3) U-Net is composed of five blocks in both the encoder and the decoder. Each block in the encoder has two convolution layers, and in the decoder, it has one transposed convolution layer.
4) Both the encoder and the decoder in FC-DenseNet consist of five dense blocks, and each dense block has five convolutional layers.
5) For the SDT-based network that directly learns the SDT representations of buildings, they utilize the aforementioned four semantic segmentation networks and, finally, convert the learned SDT representations of buildings to semantic masks by definition [24], [27].
6) The SDT-recursive model or boundary-recursive model first utilizes a U-Net to learn the SDT representation or boundaries of buildings. Afterward, they also utilize the aforementioned four semantic segmentation networks to reconstruct semantic masks of building. It should be noted that the whole method is trained in an end-to-end fashion.

D. Evaluation Metrics

The performance of models is evaluated from two aspects. Mask metrics are focused on building masks, while boundary metrics are exploited to measure the quality of boundaries of the predicted building masks.

1) Mask Metrics: In our experiments, F1 score and intersection over union (IoU) are selected as two mask metrics. They can be computed as follows:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

where TP indicates the number of true positives, FN is the number of false negatives, and FP is the number of false positives. Notable that these metrics are calculated based on building pixels rather than building objects. F1 score realizes a harmonic between precision and recall.

2) Boundary Metrics: In order to assess building boundaries, structural similarity index (SSIM) [49] and F-measure [50] are exploited as two evaluation criteria. SSIM is a measure to calculate the similarity between two images, which can be used for the quality assessment of boundaries [51]. Before the calculation of F-measure, building boundaries are extracted first from predicted semantic masks by the Sobel edge operator [52]. F-measure is used to score the boundary and is defined as the geometric mean of the precision and recall

$$\text{precision}' = \frac{\text{TP}'}{\text{TP}' + \text{FP}'} \tag{12}$$

$$\text{recall}' = \frac{\text{TP}'}{\text{TP}' + \text{FN}'} \tag{13}$$

$$\text{F-measure} = \frac{2 \times \text{precision}' \times \text{recall}'}{\text{precision}' + \text{recall}'} \tag{14}$$

where TP' is the number of correctly identified boundary pixels, FN' is the number of boundary pixels in the ground

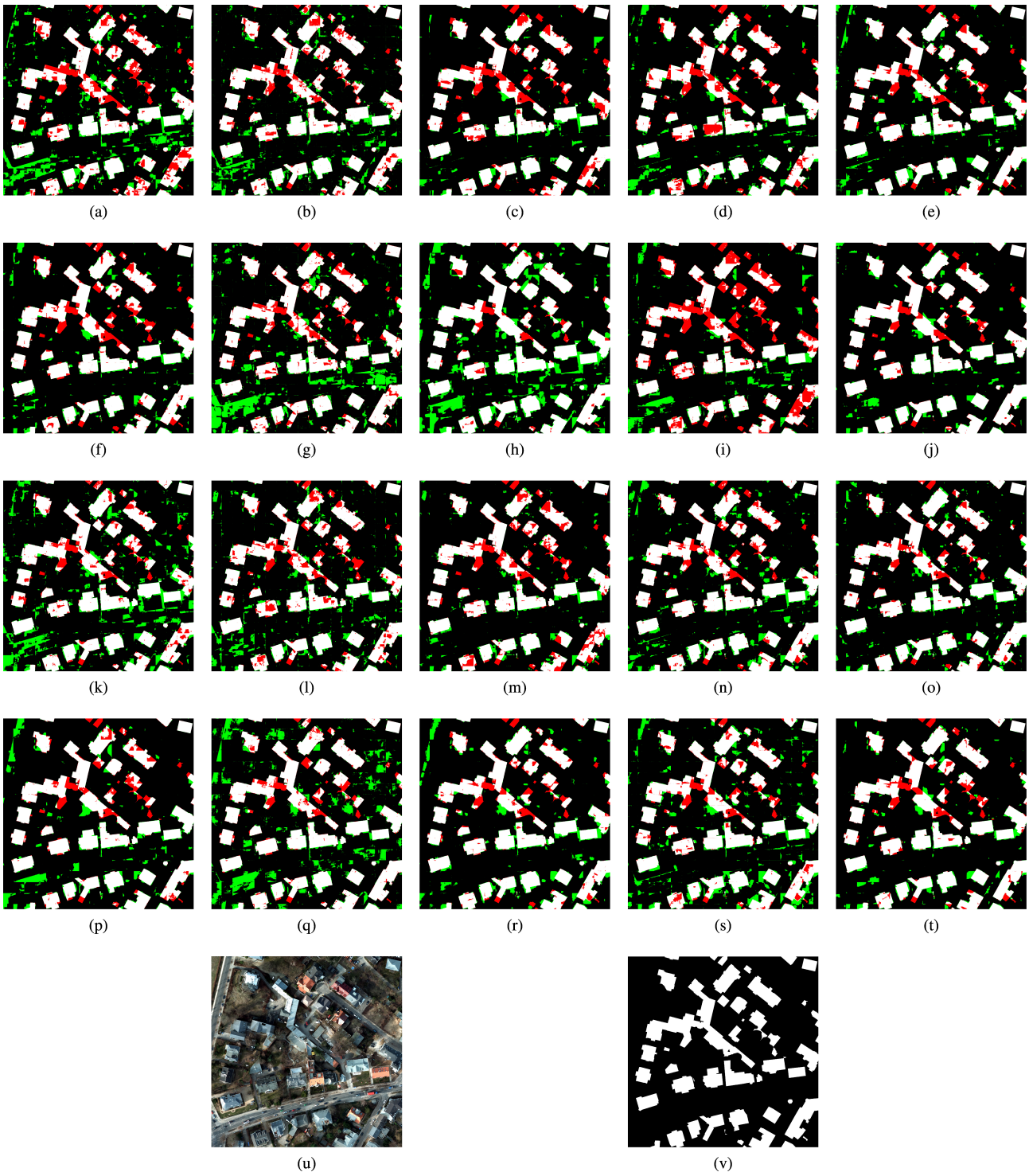


Fig. 6. Predicted results obtained from (a) FCN-8s, (b) FCN-8s-SDT, (c) FCN-8s-SDT-recursive, (d) FCN-8s-boundary-recursive, (e) proposed FCN-8s-AFM, (f) SegNet, (g) SegNet-SDT, (h) SegNet-SDT-recursive, (i) SegNet-boundary-recursive, (j) proposed SegNet-AFM, (k) U-Net, (l) U-Net-SDT, (m) U-Net-SDT-recursive, (n) U-Net-boundary-recursive, (o) proposed U-Net-AFM, (p) FC-DenseNet, (q) FC-DenseNet-SDT, (r) FC-DenseNet-SDT-recursive, (s) FC-DenseNet-boundary-recursive, and (t) proposed FC-DenseNet-AFM. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (u) and (v) Aerial imagery and ground reference from the ISPRS dataset (spatial resolution: 5 cm/pixel).

reference but being failed to be detected, and FP' is the number of nonboundary pixels mislabeled as "boundary."

V. Results and Discussion

A. Comparison With Other Competitors

The comparisons among the proposed method, naive semantic segmentation networks, SDT-based networks, SDT-learning methods, and boundary-learning methods are presented in this section. Their respective performance is evaluated according to both quantitative (cf. Tables I–III) and qualitative results (see Figs. 6–8) on three datasets.

Naive semantic segmentation networks that are regarded as baseline methods are first compared with the proposed framework. Specifically, we implement four baseline models,

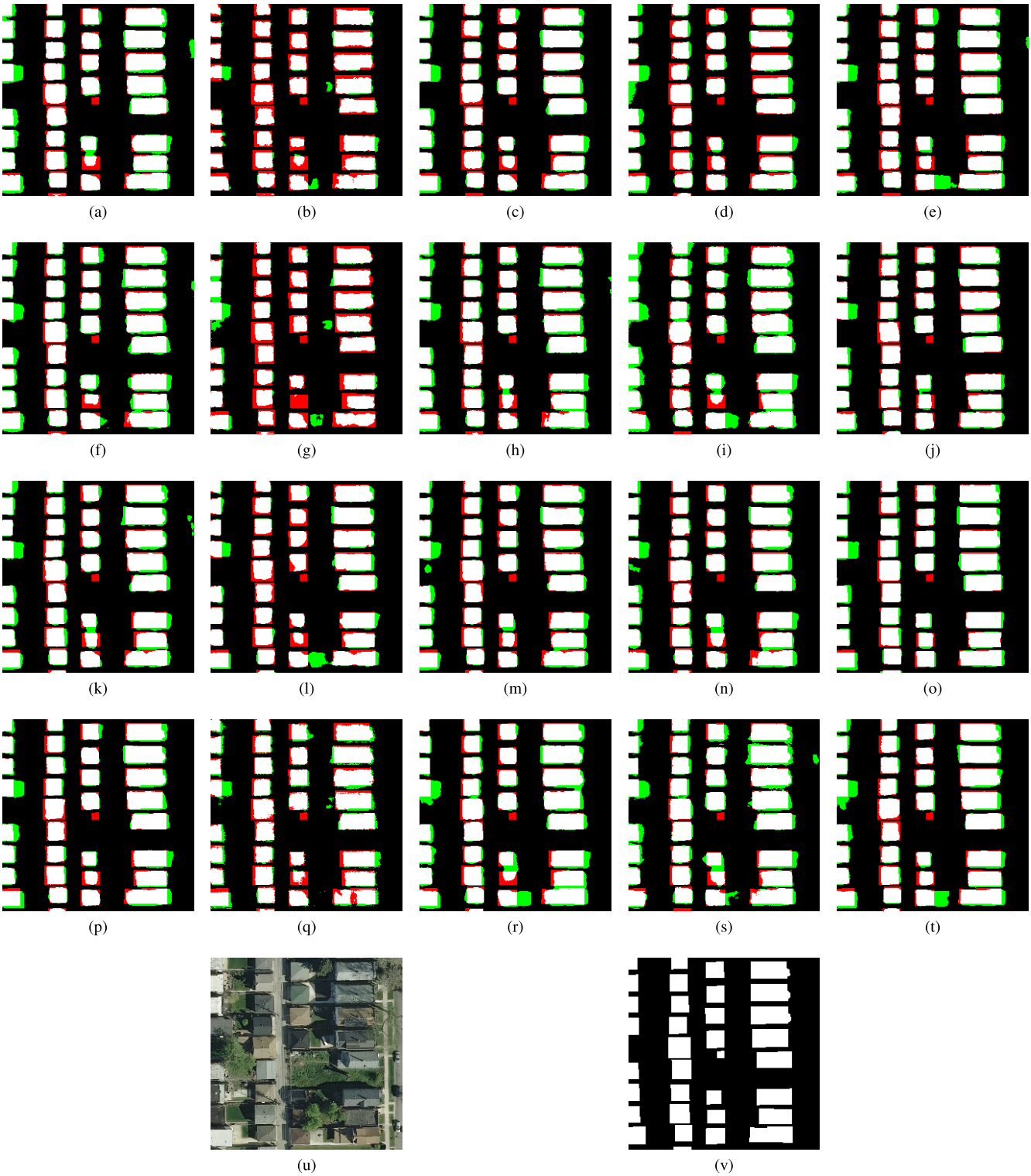


Fig. 7. Predicted results obtained from (a) FCN-8s, (b) FCN-8s-SDT, (c) FCN-8s-SDT-recursive, (d) FCN-8s-boundary-recursive, (e) proposed FCN-8s-AFM, (f) SegNet, (g) SegNet-SDT, (h) SegNet-SDT-recursive, (i) SegNet-boundary-recursive, (j) proposed SegNet-AFM, (k) U-Net, (l) U-Net-SDT, (m) U-Net-SDT-recursive, (n) U-Net-boundary-recursive, (o) proposed U-Net-AFM, (p) FC-DenseNet, (q) FC-DenseNet-SDT, (r) FC-DenseNet-SDT-recursive, (s) FC-DenseNet-boundary-recursive, and (t) proposed FC-DenseNet-AFM. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (u) and (v) Aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

i.e., FCN-8s, SegNet, U-Net, and FC-DenseNet. For a fair comparison, the AFM2Mask module is instantiated with these four networks separately. It can be seen from the statistics of three datasets that the proposed approach significantly boosts performance in both mask and boundary metrics compared with baseline networks. This indicates that the integration of learning attraction field representation is effective, and our framework can offer more robust results for the

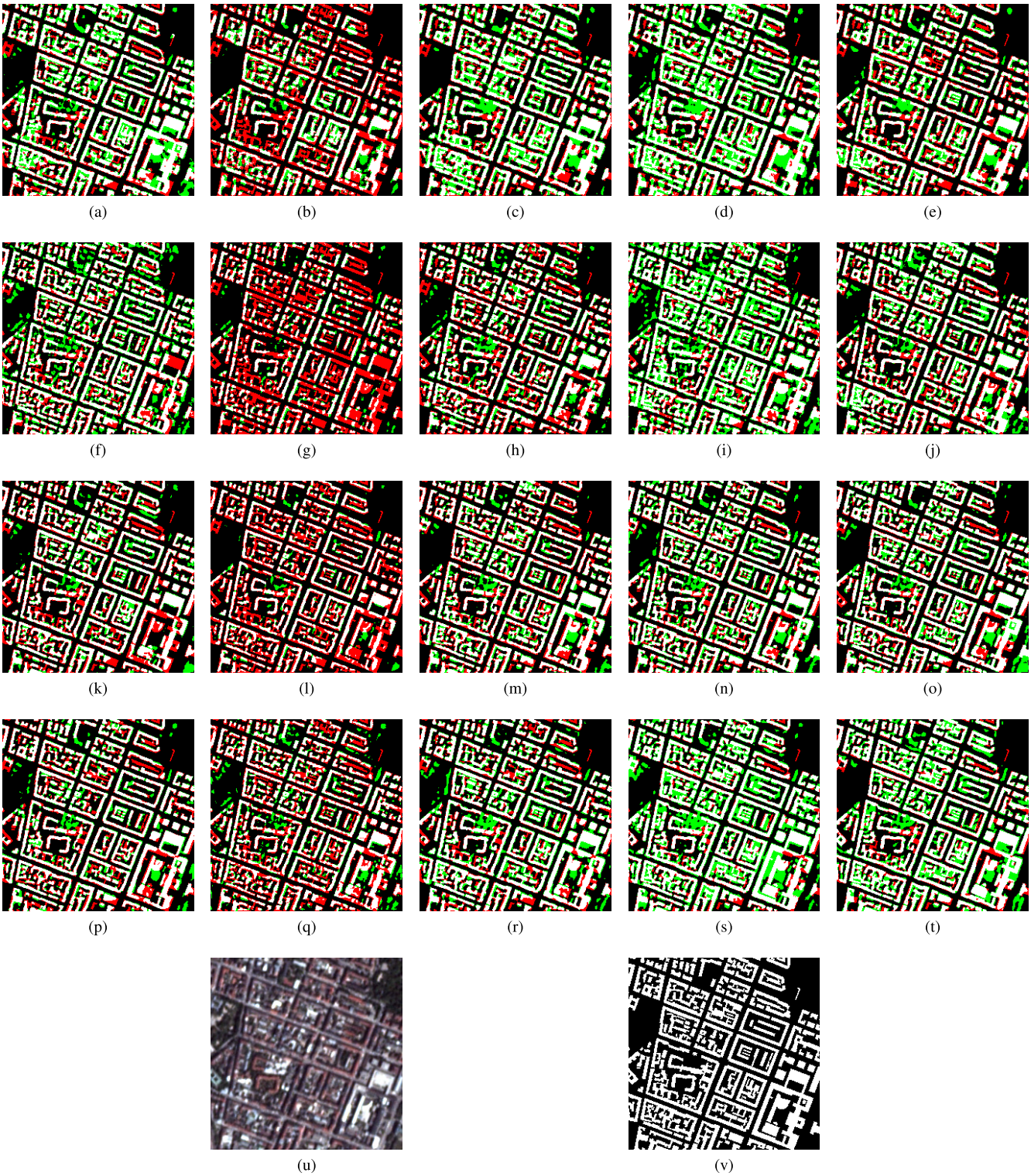


Fig. 8. Predicted results obtained from (a) FCN-8s, (b) FCN-8s-SDT, (c) FCN-8s-SDT-recursive, (d) FCN-8s-boundary-recursive, (e) proposed FCN-8s-AFM, (f) SegNet, (g) SegNet-SDT, (h) SegNet-SDT-recursive, (i) SegNet-boundary-recursive, (j) proposed SegNet-AFM, (k) U-Net, (l) U-Net-SDT, (m) U-Net-SDT-recursive, (n) U-Net-boundary-recursive, (o) proposed U-Net-AFM, (p) FC-DenseNet, (q) FC-DenseNet-SDT, (r) FC-DenseNet-SDT-recursive, (s) FC-DenseNet-boundary-recursive, and (t) proposed FC-DenseNet-AFM. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (u) and (v) Satellite imagery and ground reference from the Planet dataset (spatial resolution: 3 m/pixel).

task of building footprint generation. For the ISPRS dataset (cf. Table I), our proposed FCN-8s-AFM obtains increments of 6.65% and 10.1% in F1 score and IoU, respectively. Moreover, the proposed U-Net-AFM reaches improvements of 4.65% and 4.18% in SSIM and F-measure, respectively. The increases in boundary metrics suggest that our method can better preserve geometric details. The spatial resolution and image quality of the Planet dataset are much lower

TABLE I

ACCURACIES (%) OF DIFFERENT NETWORKS FOR BUILDING FOOTPRINT GENERATION IN THE ISPRS DATASET (SPATIAL RESOLUTION: 5 cm/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
FCN-8s	81.82	69.23	79.00	18.71
FCN-8s-SDT	84.38	72.97	79.52	22.94
FCN-8s-SDT-recursive	86.53	76.26	84.67	20.28
FCN-8s-boundary-recursive	84.83	73.66	82.67	15.97
proposed FCN-8s-AFM	**88.47**	**79.33**	**86.15**	**19.39**
SegNet	87.81	78.28	85.92	17.11
SegNet-SDT	83.88	72.24	80.86	20.61
SegNet-SDT-recursive	87.09	77.14	84.12	15.01
SegNet-boundary-recursive	81.79	69.20	80.36	14.38
proposed SegNet-AFM	**90.56**	**82.75**	**88.76**	**20.34**
U-Net	85.37	74.48	82.59	19.32
U-Net-SDT	84.77	73.57	82.65	19.98
U-Net-SDT-recursive	86.65	76.44	84.65	18.11
UNet-boundary-recursive	86.33	75.94	83.63	19.18
proposed U-Net-AFM	**89.30**	**80.67**	**87.24**	**23.50**
FC-DenseNet	88.34	79.11	86.24	20.76
FC-DenseNet-SDT	88.03	78.61	85.16	23.95
FC-DenseNet-SDT-recursive	87.98	78.53	85.62	18.96
FC-DenseNet-boundary-recursive	85.63	74.88	82.75	19.20
proposed FC-DenseNet-AFM	**89.38**	**80.81**	**87.70**	**21.17**

TABLE II

ACCURACIES (%) OF DIFFERENT NETWORKS FOR BUILDING FOOTPRINT GENERATION IN THE INRIA DATASET (SPATIAL RESOLUTION: 30 cm/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
FCN-8s	84.79	73.60	85.65	27.01
FCN-8s-SDT	78.07	64.03	82.12	20.58
FCN-8s-SDT-recursive	85.11	74.08	86.06	28.16
FCN-8s-boundary-recursive	84.84	73.68	85.73	26.86
proposed FCN-8s-AFM	**85.92**	**75.31**	**86.47**	**29.92**
SegNet	84.43	73.05	85.47	28.16
SegNet-SDT	76.72	62.23	82.07	20.34
SegNet-SDT-recursive	84.78	73.58	85.79	27.39
SegNet-boundary-recursive	83.08	71.06	84.63	23.40
proposed SegNet-AFM	**86.18**	**75.72**	**86.88**	**31.38**
U-Net	84.83	73.66	86.76	28.98
U-Net-SDT	83.27	71.33	85.12	28.28
U-Net-SDT-recursive	85.41	74.54	86.48	28.06
U-Net-boundary-recursive	85.06	74.01	86.18	28.79
proposed U-Net-AFM	**86.68**	**76.49**	**87.07**	**33.77**
FC-DenseNet	84.66	73.41	85.92	28.96
FC-DenseNet-SDT	77.90	63.80	81.38	27.68
FC-DenseNet-SDT-recursive	84.86	73.70	85.81	27.67
FC-DenseNet-boundary-recursive	83.69	71.95	84.61	27.52
proposed FC-DenseNet-AFM	**85.41**	**74.53**	**86.20**	**29.72**

than the other two datasets, which may lead to a negative effect on accurately extracting buildings. In this case, although improvements in both mask and boundary metrics on the Planet dataset are less significant than those on the other two datasets, the nearly 1% gain is still not trivial.

From qualitative results, we can observe that building boundaries obtained from naive semantic segmentation networks are blurred, which is also pointed out in [53]–[55]. The visual comparisons (cf. Figs. 6–8) demonstrate the effectiveness of the proposed method. As illustrated in Fig. 7, semantic masks provided by naive networks have blob-like shapes. Even with skip connections that help compensate spatial details in networks, U-Net and FC-DenseNet fail to achieve accurate building boundaries. Moreover, this scene is a residential area, and some consecutive buildings are identified as a large building by most of the baseline models. Note that building boundaries produced by our algorithm are more rectilinear and precise. Even for buildings with complex structures (cf. Fig. 6 and 8), building footprints generated from our framework are more adherent to the ground reference. These observations suggest that our model really benefits from learning attraction field representation, enabling us to gain more geometric details of buildings.

The attraction field representation can be considered as a type of distance transform, which represents the relationship between the pixel and the boundary. Therefore, we also

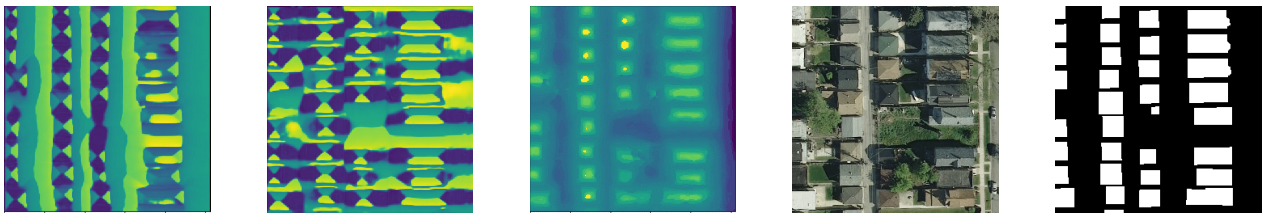
TABLE III
ACCURACIES (%) OF DIFFERENT NETWORKS FOR BUILDING FOOTPRINT GENERATION IN THE PLANET DATASET (SPATIAL RESOLUTION: 3 m/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
FCN-8s	60.45	43.32	66.86	48.84
FCN-8s-SDT	47.87	31.47	64.43	33.68
FCN-8s-SDT-recursive	60.96	43.85	67.43	46.37
FCN-8s-boundary-recursive	59.80	42.66	66.47	45.26
proposed FCN-8s-AFM	**61.32**	**44.22**	**68.42**	**49.67**
SegNet	56.30	39.18	63.11	52.48
SegNet-SDT	46.11	29.96	61.47	44.46
SegNet-SDT-recursive	57.01	39.87	63.66	49.52
SegNet-boundary-recursive	56.52	39.40	61.49	50.33
proposed SegNet-AFM	**60.38**	**43.25**	**66.21**	**53.31**
U-Net	63.74	46.78	70.63	54.03
U-Net-SDT	50.99	34.22	66.32	39.39
U-Net-SDT-recursive	63.21	46.21	69.40	53.84
U-Net-boundary-recursive	64.14	47.21	68.74	54.83
proposed U-Net-AFM	**65.03**	**48.18**	**69.71**	**56.62**
FC-DenseNet	64.54	47.65	70.80	54.35
FC-DenseNet-SDT	55.62	38.52	64.63	49.90
FC-DenseNet-SDT-recursive	61.60	44.51	67.68	52.52
FC-DenseNet-boundary-recursive	64.68	47.80	68.82	55.10
proposed FC-DenseNet-AFM	**65.68**	**48.90**	**70.56**	**56.67**

take another type of distance transform: SDT as competitors. One competitor is an SDT-based network that utilizes variant backbones to learn the SDT representation of buildings and then convert this representation to semantic masks by definition [24], [27]. Compared to baseline networks, the SDT-based network can contribute to the F-measure only on the ISPRS dataset. However, there are even decreases in mask metrics. This suggests that directly learning SDT labels as final output have the potential for the improvement of geometric details only in remote sensing data with very high resolution (e.g., 5 cm/pixel). The other competitor is the SDT-recursive model, which first learns the SDT representation of buildings with a U-Net and then reconstructs semantic masks by different backbones. Notable that the whole method is trained in an end-to-end fashion. The SDT-recursive model that feeds the learned SDT representations into semantic segmentation networks is much superior to the SDT-based network, as we can see gains in both mask and boundary metrics. This may be because the SDT representation learned from the remote sensing imagery carries useful information to capture the global semantic context in semantic segmentation networks, which indicates the potential of SDT in a recursive learning way for building footprint generation. It is worthy to note that the performances of both SDT-based network and SDT-recursive model are more sensitive to the backbone semantic segmentation networks. For the ISPRS dataset (see Table III), when the backbone is FCN-8s, both SDT-based network and SDT-recursive model can boost the performance. However, the performance of SegNet-SDT and SegNet-SDT-recursive is worse than that of SegNet.

The geometric property of building boundaries can be significantly enhanced by AFMs (see Fig. 2). From Tables I–III, it can be observed that our framework can improve results in terms of both mask and boundary metrics, which confirms that explicitly encoding geometric information is essential to building footprint generation tasks. In this regard, we investigate another competitor, the boundary-recursive model, to further validate the effectiveness of the attraction field representation. This method first learns building boundaries from remote sensing images with a U-Net and then uses them as auxiliary information to extract building masks by variant semantic segmentation networks. Notable that these two subnetworks are jointly optimized. Experimental results show that this model does not bring this task any benefits in terms of building boundary quality, and we can see decreases in boundary metrics and more blurred boundaries compared to the naive semantic segmentation network. This may be because building boundaries are characterized with very few pixels, and this class imbalance leads to ambiguity in the network learning.

By contrast, our method can always provide significant gains, regardless of which semantic segmentation network architecture is chosen as the AFM2Mask module, and the proposed approach outperforms other competitors in most of the statistical metrics for three datasets. This is due to two facts. One is that the attraction field representation can encode geometric properties in 2-D (x- and y-directions), but SDT only relies on the Euclidean distance and, thus, characterizes the information in 1-D. This indicates that the use of the information in different dimensions is more reliable and accurate. Fig. 9(a) and (b) presents the AFM learned by the proposed U-Net-AFM, and Fig. 9(c) shows the SDT representation learned by U-Net-SDT-recursive. It can be observed that attraction field representation can better delineate sharp building boundaries. The other factor is that the attraction field representation takes the nonboundary pixels into account, which have addressed the challenges of class imbalance in boundary-learning methods.



(a) (b) (c) (d) (e)

Fig. 9. (a) AFM (x-axis) and (b) AFM (y-axis) are learned by the proposed method (U-Net-AFM). (c) SDT representation learned by the U-Net-SDT-recursive. (d) and (e) Aerial imagery and ground reference from the INRIA dataset (spatial resolution: 0.3 m/pixel).

TABLE IV
ACCURACIES (%) OF DIFFERENT COEFFICIENTS OF AFM2MASK LOSS (λ) FOR BUILDING FOOTPRINT GENERATION IN THE INRIA DATASET (SPATIAL RESOLUTION: 30 cm/pixel)

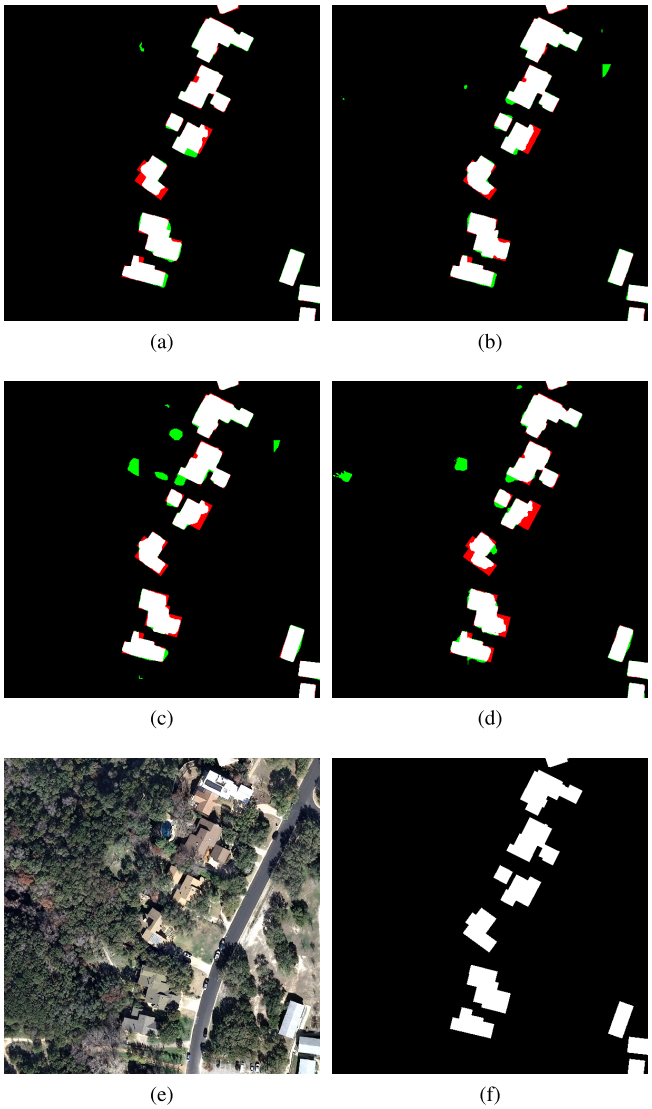
	Mask		Boundary	
λ	F1 score	IoU	SSIM	F-measure
10	86.04	75.50	86.77	31.80
1	86.37	76.01	87.06	33.22
0.1	**86.68**	**76.49**	**87.07**	**33.77**

B. Analysis of Hyperparameter Tuning

As shown in the results on three datasets, taking U-Net as the AFM2Mask module can deliver relatively satisfactory results on all three datasets. Therefore, in this section, we use U-Net-AFM for further studies. Moreover, the INRIA dataset is selected as an example dataset to carry out the following experiments.

In the proposed framework, the global loss function is utilized to guide the end-to-end learning of building masks from remote sensing data. This function is a sum of losses from two separate modules, where the hyperparameter λ is the coefficient of the AFM2Mask module. Here, λ is set as three different numbers, i.e., 0.1, 1, and 10, to investigate its impact on final results.

The statistical results with different values of λ are shown in Table IV. We can see that our model is insensitive to this parameter, and networks with all different λ values outperform the naive U-Net. Furthermore, increasing the value of λ will lead to a slight reduction in both mask and boundary metrics. The best result is obtained when $\lambda = 0.1$. A small value of λ indicates more significance of the Img2AFM module than the AFM2Mask module, which places an emphasis on the attraction field representation learning in the whole framework. It can be clearly seen from the Fig. 10 that gradually lowering λ can reduce false detections. This is mainly because the attraction field representation can alleviate the impact of background clutters.



(a) (b) (c) (d) (e) (f)

Fig. 10. Results obtained by the proposed method (U-Net-AFM) with coefficient λ = (a) 0.1, (b) 1, and (c) 10. (d) Result obtained by the naive U-Net. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (e) and (f) Corresponding aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

C. Different Methods to Incorporate Attraction Field Representation

It is worth noting that building boundaries leaned by the proposed method are significantly improved due to the exploitation of attraction field representation. In order to further explore how to well leverage attraction field representation, we investigate another three designs to incorporate this useful representation in network learning.

1) **Srivastava *et al.* [56]:** It uses a U-Net architecture followed by two separate fully connected layers to

TABLE V
ACCURACIES (%) OF DIFFERENT DESIGNS FOR THE INCORPORATION OF ATTRACTION FIELD REPRESENTATION IN THE INRIA DATASET (SPATIAL RESOLUTION: 30 cm/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
U-Net	84.83	73.66	86.76	28.98
proposed U-Net-AFM	**86.68**	**76.49**	**87.07**	**33.77**
Srivastava *et al.* [56]	85.97	75.39	86.54	29.83
Bischke *et al.* [47]	86.10	75.59	86.67	29.80
Mou & Zhu [57]	85.48	74.63	86.29	30.06

TABLE VI
ACCURACIES (%) OF DIFFERENT METHODS FOR BUILDING FOOTPRINT GENERATION IN THE ISPRS DATASET (SPATIAL RESOLUTION: 5 cm/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
proposed SegNet-AFM	**90.56**	**82.75**	**88.76**	**20.34**
Griffiths & Boehm [58]	85.00	-	-	-
Lin *et al.* [59]	88.47	79.94	85.73	18.41
Wei *et al.* [60]	88.65	80.23	86.40	19.67

learn semantic masks and attraction field representation, respectively.

2) **Bischke *et al.* [47]:** It takes a U-Net as the backbone and first adds one convolutional layer after the decoder to learn the attraction field representation. Afterward, this learned attraction field representation and feature maps produced by the decoder are concatenated and fed into another convolutional layer to learn final segmentation masks.
3) **Mou and Zhu [57]:** It utilizes an encoder and two separate decoders to jointly optimize two complementary tasks, namely, building semantic segmentation and attraction field representation learning. Note that the architecture of encoder and decoders in this design is the same as those in U-Net.

The statistical and visual results are reported in Table V and Fig. 11, respectively. From both mask and boundary metrics in Table V, all methods have shown superior results than naive U-Net, which again confirms the significance of attraction field representation in our task. Among all design options, the proposed framework has achieved the best performance. In particular, the F-measure achieved by our approach is increased by more than 3% when compared to the other methods. Besides, it can be seen that the building boundaries and corners learned by the proposed framework are more accurate than its competitors. This suggests that our approach is able to effectively leverage information of attraction field representation, which is attributed to our recursive learning strategy.

D. Comparison With State-of-the-Art Methods

To verify the superiority of our approach on datasets with different spatial resolutions, we make a comparison with other

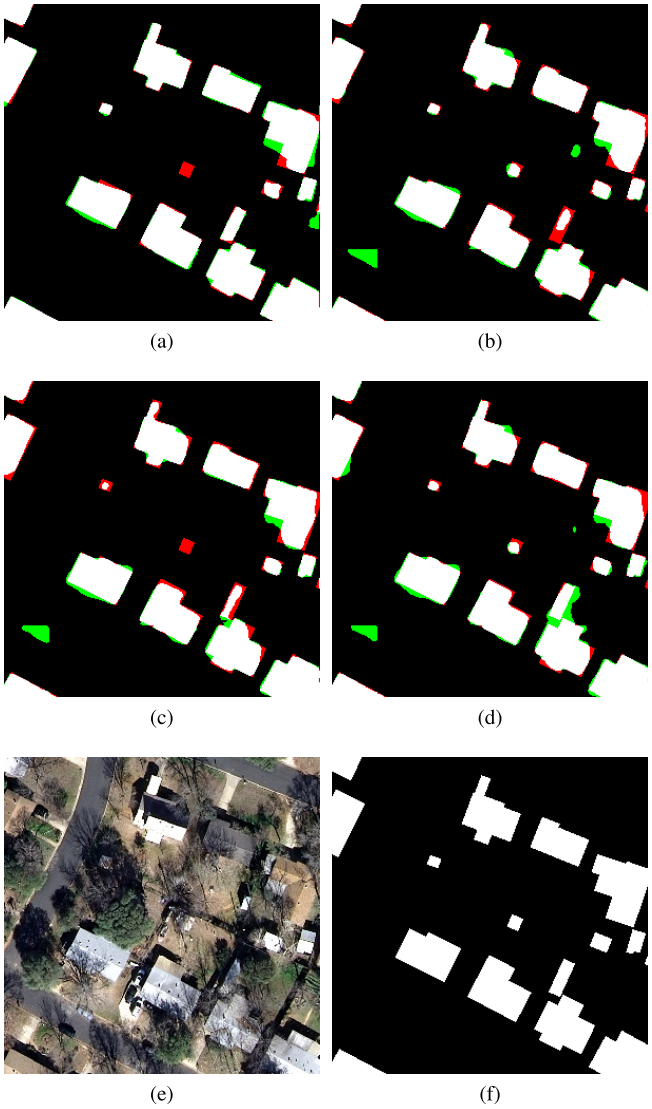


Fig. 11. Results obtained by (a) proposed U-Net-AFM, (b) Srivastava *et al.* [56], (c) Bischke *et al.* [47], and (d) Mou and Zhu [57]. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (e) and (f) Corresponding aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

state-of-the-art methods on the ISPRS, INRIA, and Planet datasets. The statistical results of different algorithms on three datasets are shown in Tables VI–VIII, respectively. On both ISPRS and Planet datasets, the proposed method surpasses all other models in both mask and boundary metrics. For the INRIA dataset, our approach achieves the highest scores in boundary metrics and comparative performance in mask prediction. Compared to our methods, Girard's method [35] gains a marginal improvement in mask metrics at the cost of additional ground-truth annotations (i.e., vector format of building footprints). For an intuitive comparison, the visual results of our method and Girard's method [35] are illustrated in Fig. 12. As we can see, Girard's method [35] fails to recover detailed structures of complicated buildings. On the contrary, our approach can accurately capture more geometric details,

TABLE VII
ACCURACIES (%) OF DIFFERENT METHODS FOR BUILDING FOOTPRINT GENERATION IN THE INRIA DATASET (SPATIAL RESOLUTION: 30 cm/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
proposed U-Net-AFM	86.68	76.49	**87.07**	**33.77**
Ji *et al.* [25]	-	71.40	-	-
Liu *et al.* [61]	-	71.76	-	-
Bischke *et al.* [47]	-	73.00	-	-
Audebert *et al.* [62]	-	74.17	-	-
Girard *et al.* [35]	**86.82**	**76.71**	86.49	32.00

TABLE VIII
ACCURACIES (%) OF DIFFERENT METHODS FOR BUILDING FOOTPRINT GENERATION IN THE PLANET DATASET (SPATIAL RESOLUTION: 3 m/pixel)

	Mask		Boundary	
Method	F1 score	IoU	SSIM	F-measure
proposed FC-DenseNet-AFM	**65.68**	**48.90**	**70.56**	**56.67**
Lin *et al.* [59]	59.54	42.39	64.53	53.03
Wei *et al.* [60]	64.85	47.98	69.06	54.85

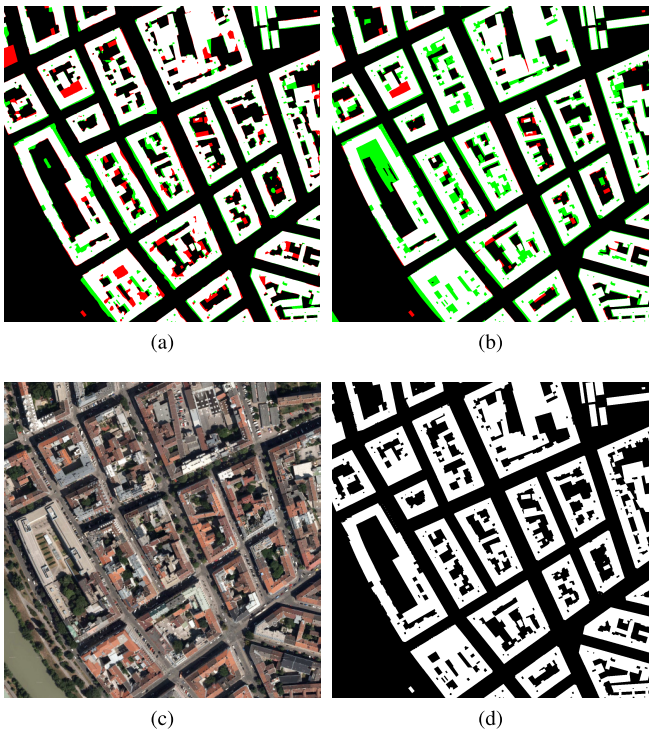


Fig. 12. Results obtained by (a) proposed U-Net-AFM and (b) Girard *et al.* [35]. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (c) and (d) Corresponding aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

which again demonstrates the strength of the AFM for the task of building footprint generation.

VI. CONCLUSION

Considering that building boundaries are easily blurred when using semantic segmentation networks to directly learn building footprints, a new end-to-end building footprint generation method through learning the attraction field representation is proposed in this article. The proposed model comprises two modules: an Img2AFM module and an AFM2Mask module. More specifically, the former is designed to learn the attraction field representation, which enables not only the enhancement of building boundaries but also the suppression of background clutters. Afterward, the latter exploits the input remote sensing image and learned AFM to reconstruct building masks. The performance of the proposed end-to-end network is assessed on three datasets with different spatial resolutions: the ISPRS dataset (5 cm/pixel), the INRIA dataset (30 cm/pixel), and the Planet dataset (3 m/pixel). Experimental results suggest that the incorporation of the attraction field representation in our framework can offer more satisfactory building footprint maps. On the one hand, sharp boundaries and geometric details of buildings can be better preserved. On the other hand, non-building objects that are wrongly detected as buildings can be avoided to a large extent. Thus, we believe that our method has the potential to be a robust solution for building footprint generation at a large scale. Looking into the future, we intend to investigate the potential of the attraction field representation in other tasks, e.g., road extraction and vehicle detection.

REFERENCES

[1] X. Huang, W. Yuan, J. Li, and L. Zhang, "A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017.

[2] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[3] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.

[4] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271619301261

[5] C. Qiu, M. Schmitt, C. Geiß, T.-H.-K. Chen, and X. X. Zhu, "A framework for large-scale mapping of human settlement extent from sentinel-2 images via fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 152–170, May 2020.

[6] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[7] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.

[8] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, early access, Dec. 16, 2020, doi: 10.1109/TGRS.2020.3042607.

[9] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, Dec. 2020.

[10] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, "SRAF-net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6154–6168, Jul. 2021.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[12] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.

[13] Q. Li *et al.*, "Instance segmentation of buildings using keypoints," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep./Oct. 2020, pp. 1452–1455.

[14] S. Cui, Q. Yan, and P. Reinartz, "Complex building description and extraction based on Hough transformation and cycle detection," *Remote Sens. Lett.*, vol. 3, no. 2, pp. 151–159, Mar. 2012.

[15] L. Zhang *et al.*, "Learning deep structured active contours end-to-end," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8877–8885.

[16] N. Xue, S. Bai, F. Wang, G.-S. Xia, T. Wu, and L. Zhang, "Learning attraction field representation for robust line segment detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1595–1603.

[17] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, Dec. 2013.

[18] M. Turker and D. Koc-San, "Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 58–69, Feb. 2015.

[19] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral geoeye-1 imagery," *Photogramm. Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, 2011.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[22] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

[23] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[24] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[25] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[26] X. Li, X. Yao, and Y. Fang, "Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3680–3687, Oct. 2018.

[27] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[28] K. G. Derpanis, *The Harris Corner Detector*. Toronto, ON, Canada: York Univ., 2004, pp. 1–2.

[29] C. Harris and S. Mike, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, Manchester, U.K., vol. 15, no. 50, pp. 10–5244, Sep. 1988.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[31] M. Zangrandi, E. Baccaglini, and L. Boulard, "An enhanced corner-based automatic rooftop extraction algorithm leveraging drlse segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1024–1027.

[32] M. Wang, S. Yuan, and J. Pan, "Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed Hough transform," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. - IGARSS*, Jul. 2013, pp. 508–511.

[33] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1715–1724.

[34] W. Zhao, C. Persello, and A. Stein, "Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 119–131, May 2021.

[35] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building segmentation by frame field learning," 2020, *arXiv:2004.14875*. [Online]. Available: http://arxiv.org/abs/2004.14875

[36] R. O. Duda and R. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972.

[37] J. B. Burns, A. R. Hanson, and E. M. Riseman, "Extracting straight lines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 4, pp. 425–455, Jul. 1986.

[38] M. Izadi and P. Saeedi, "Three-dimensional polygonal building model estimation from single satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2254–2272, Jun. 2012.

[39] C. Akinlar and C. Topal, "EDLines: A real-time line segment detector with a false detection control," *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1633–1642, Oct. 2011.

[40] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 487–491, Mar. 2015.

[41] X. Qin, S. He, X. Yang, M. Dehghan, Q. Qin, and J. Martin, "Accurate outline extraction of individual building from very high-resolution optical images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1775–1779, Nov. 2018.

[42] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "DARNet: Deep active ray network for building segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7431–7439.

[43] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[44] *ISPRS*. Accessed: Dec. 15, 2018. [Online]. Available: http:http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[45] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[46] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sensors Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[47] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1480–1484.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[50] I. Kokkinos, "Boundary detection using F-measure-, filter- and feature-(F^3) boost," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 650–663.

[51] D. Sadykova and A. P. James, "Quality assessment metrics for edge detection and edge-aware filtering: A tutorial review," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 2366–2369.

[52] I. Sobel, *An Isotropic 3×3 Gradient Operator, Machine Vision for Three–Dimensional Scenes*. New York, NY, USA: Academic, 1990, pp. 376–379.

[53] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020.

[54] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.

[55] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[56] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNS," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5173–5176.

[57] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[58] D. Griffiths and J. Boehm, "Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne lidar and image data using active contours," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 70–83, Aug. 2019.

[59] J. Lin, W. Jing, H. Song, and G. Chen, "ESFNet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019.

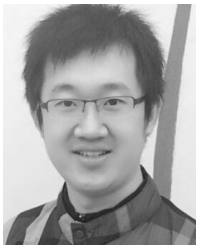
[60] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[61] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, 2019.

[62] N. Audebert, A. Boulch, B. Le Saux, and S. Lefèvre, "Distance transform regression for spatially-aware deep semantic segmentation," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102809.

Qingyu Li (Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2015, and the master's degree in earth oriented space science and technology (ESPACE) from the Technische Universität München (TUM), Munich, Germany, in 2018. She is currently pursuing the Ph.D. degree with TUM and the German Aerospace Center (DLR), Weßling, Germany.

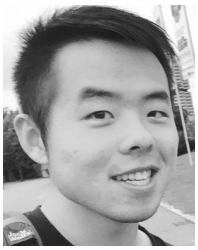
Her research interests include deep learning, remote sensing mapping, and remote sensing applications.



Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. He is currently a Guest Professor with the Munich AI Future Lab AI4EO, TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been a Research Scientist with DLR-IMF and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU).

Dr. Mou was a recipient of the First Place in the 2016 IEEE GRSS Data Fusion Contest and the finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.



Yuansheng Hua (Graduate Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2014, and the double master's degrees in earth oriented space science and technology (ESPACE) and photogrammetry and remote sensing from the Technical University of Munich (TUM), Munich, Germany, and Wuhan University in 2018 and 2019, respectively. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

In 2019, he was a Visiting Researcher with Wageningen University & Research, Wageningen, The Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.

Yilei Shi (Member, IEEE) received the Dipl.Ing degree in mechanical engineering and the Dr.Ing degree in signal processing from the Technische Universität München (TUM), Munich, Germany, in 2010 and 2019, respectively.

In April and May 2019, he was a Guest Scientist with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, U.K. He is currently a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and parallel computing for large-scale problems, high-performance computing and computational intelligence, advanced methods on synthetic-aperture radar (SAR) and interferometric SAR (InSAR) processing, machine learning and deep learning for a variety of data sources, such as SAR, optical images, and medical images, and partial differential equation (PDE)-related numerical modeling and computing.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School, TUM. Since 2019, she has been the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She is currently a Professor of data science in earth observation (former: signal processing in earth observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is currently also an AI Professor with ESA's Phi-Lab, Frascati, Italy. Her main research interests are remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is also a member of the young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as an Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.