Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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

Delineation of agricultural fields using multi-task BsiNet from high-resolution satellite images

Jiang Long [a], Mengmeng Li [a,*], Xiaoqin Wang [a], Alfred Stein [b]

[a] Key Lab of Spatial Data Mining & Information Sharing of Ministry of Education, Academy of Digital China (Fujian), Fuzhou University, Fuzhou, China
[b] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands

ARTICLE INFO

Keywords:
Agricultural fields
High-resolution satellite images
Multi-task network
BsiNet
Spatial group-wise enhancement

ABSTRACT

This paper presents a new multi-task neural network, called BsiNet, to delineate agricultural fields from high-resolution satellite images. BsiNet is modified from a Psi-Net by structuring three parallel decoders into a single encoder to improve computational efficiency. BsiNet learns three tasks: a core task for agricultural field identification and two auxiliary tasks for field boundary prediction and distance estimation, corresponding to mask, boundary, and distance tasks, respectively. A spatial group-wise enhancement module is incorporated to improve the identification of small fields. We conducted experiments on a GaoFen1 and three GaoFen2 satellite images collected in Xinjiang, Fujian, Shandong, and Sichuan provinces in China, and compared BsiNet with 13 different neural networks. Our results show that the agricultural fields extracted by BsiNet have the lowest global over-classification (GOC) of 0.062, global under-classification (GUC) of 0.042, and global total errors (GTC) of 0.062 for the Xinjiang dataset. For the Fujian dataset with irregular and complex fields, BsiNet outperformed the second-best method from the Xinjiang dataset analysis, yielding the lowest GTC of 0.291. It also produced satisfactory results on the Shandong and Sichuan datasets. Moreover, BsiNet has fewer parameters and faster computation than existing multi-task models (i.e., Psi-Net and ResUNet-a D7). We conclude that BsiNet can be used successfully in extracting agricultural fields from high-resolution satellite images and can be applied to different field settings.

1. Introduction

Detailed spatial information on agricultural fields is essential to many applications in agriculture, such as crop mapping, crop yield estimation, and sustainable agriculture planning (McCarty et al., 2017; Belgiu and Csillik, 2018). Traditional methods of delineating the boundaries of agricultural fields are usually done by field surveys or the visual interpretation of remote sensing images. Such practices are labour-intensive and time-consuming. Recently, considerable efforts have been devoted to automatically delineating field boundaries using satellite images (Graesser and Ramankutty, 2017; Waldner et al., 2021), especially those with high spatial resolutions, providing opportunities to identify fields with small sizes and irregular shapes (Persello et al., 2019).

A widespread practice for field delineation is using image segmentation techniques based upon edge detection, region segmentation, and machine learning (Hossain and Chen, 2019). Commonly used edge detection methods only consider the changes between neighbouring pixels, limiting the use of high-level image features. A major problem in edge detection is to obtain closed boundaries. To address this problem, increasing attention has been given to region-based methods, e.g., using multi-resolution segmentation algorithms (Wassie et al., 2018). Region-based methods provide solutions to form closed boundaries when delineating agricultural fields. These methods, however, depend on well-segmented image objects, for which poor segmentation results may lead to low-quality field delineation (Hossain and Chen, 2019).

Recently, more focus has been given to deep learning techniques in extracting agricultural fields from satellite images (Volpi and Tuia, 2018; Masoud et al., 2020; Wagner and Oppelt, 2020; Zhang et al., 2021). Convolutional neural networks (CNNs) are popular for image semantic segmentation, providing an alternative solution for extracting agricultural fields. For example, Persello et al. (2019) used an encoder-decoder fully convolutional network, i.e., SegNet, and Garcia-Pedrero et al. (2019) used a UNet to extract field boundaries automatically. Although these methods can obtain acceptable results, they ignore the role of related tasks that share an optimal hypothesis class (Ruder,

* Corresponding author.
E-mail addresses: 205527028@fzu.edu.cn (J. Long), mli@fzu.edu.cn (M. Li), wangxq@fzu.edu.cn (X. Wang), a.stein@utwente.nl (A. Stein).

https://doi.org/10.1016/j.jag.2022.102871
Received 9 February 2022; Received in revised form 12 June 2022; Accepted 13 June 2022

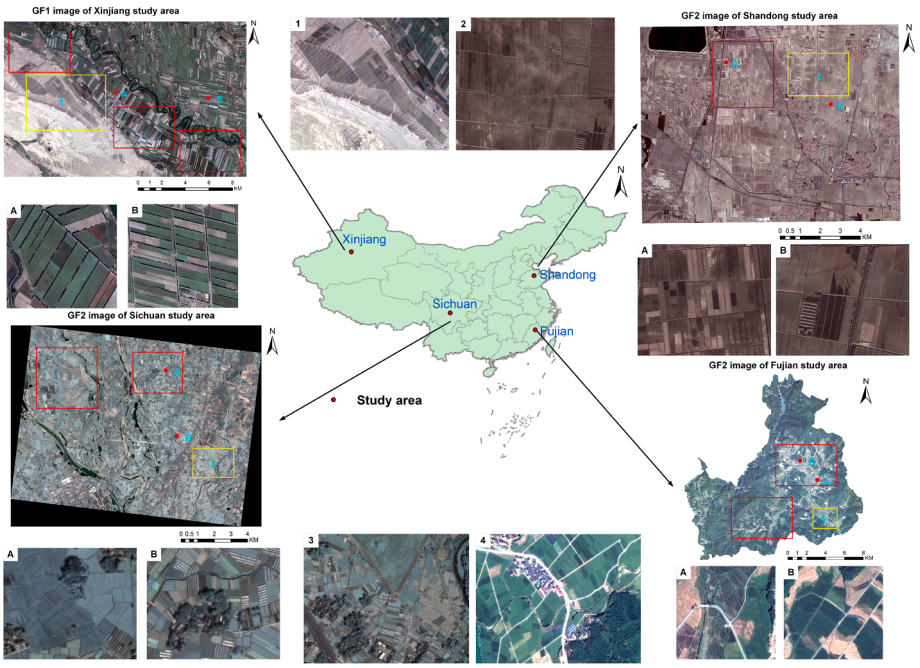

Fig. 1. Overview of the Xinjiang, Fujian, Shandong and Sichuan study areas and satellite images. Red and yellow boxes refer to subsets of training and testing areas, A and B are enlarged views of agricultural fields, and 1–4 are the test areas of the four study areas. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Details of the study areas and images. XJ, FJ, SD, and SC refer to Xinjiang, Fujian, Sichuan, and Shandong, respectively. M and P refer to multi-spectral and panchromatic bands. SR, TR, and Dates refer to spatial resolution, temporal resolution, and acquisition dates.

Areas	Satellites	SR	TR	Dates	Size (pixels)	Area (km^2)
XJ	GF1	M = 8 m, P = 2 m	4 days	20180901	10272 × 6556	270.76
FJ	GF2	M = 4 m, P = 1 m	5 days	20190919	19149 × 21018	207.90
SD	GF2	M = 4 m, P = 1 m	5 days	20211220	12965 × 9905	125.97
SC	GF2	M = 4 m, P = 1 m	5 days	20210321	15094 × 12013	147.46

2017), which can help to reduce network overfitting and improve generalisation.

An interesting method is based upon multi-task learning that learns various tasks through a jointed representation to improve model effectiveness (Ruder, 2017). A recent multi-task learning model of interest is Psi-Net, which consists of three learning modules corresponding to boundary, mask, and distance tasks (Murugesan et al., 2019). The addition of relevant tasks is conducive to improving the shape and boundary of the primary task (i.e., mask), resulting in better generalisation than conventional CNNs. Waldner and Diakogiannis (2020) used a multi-task neural network ResUNet-a to identify field boundaries, extent, and distance from Sentinel-2 images, followed by a watershed segmentation to obtain closed boundaries for individual fields. Their methods obtained satisfactory results for medium-resolution satellite images, but the potential for field extraction on high spatial resolution images remains to be investigated.

This paper aims to extract agricultural fields from high spatial resolution satellite images using multi-task learning methods. We propose a single encoder-decoder multi-task network unifying boundary and shape perception, called BsiNet, to improve the geometric accuracy of extracted field boundaries. The novelty and contributions of the paper are as follows:

- A new multi-task model, called BsiNet, is proposed to delineate agricultural fields from high-resolution satellite images (accessed at: https://github.com/long123524/BsiNet-torch);
- The model has a flexible backbone structure that can be replaced by various state-of-the-art networks;
- An extensive experimental analysis is conducted to evaluate the effectiveness and efficiency of BsiNet using high-resolution images collected in four different areas. Moreover, BsiNet is compared to 13 different methods with varying distance features, network backbones, and attention mechanisms.

The remainder of this paper is organised as follows: Section 2 illustrates the study areas and datasets, Section 3 describes the proposed method in detail. Section 4 presents the experimental results and related analysis, followed by a discussion in Section 5 and the conclusions of this research in Section 6.

2. Study areas and datasets

Experiments were conducted in four study areas located in Xinjiang, Fujian, Shandong, and Sichuan, in China, respectively (Fig. 1). Agricultural fields in Xinjiang and Shandong areas have regular shapes, big sizes, and clear boundaries, while fields in Fujian and Sichuan areas are small, irregularly shaped, and have ragged boundaries. It is therefore challenging to delineate field boundaries in the Fujian and Sichuan areas. For the Xinjiang area, a GaoFen1 (GF1) satellite image was acquired, while GaoFen2 (GF2) images were obtained for the other three areas. Detailed information regarding the study areas and images is given in Table 1. We used the Gram-Schmidt pan-sharpening method to fuse multi-spectral GF1 and GF2 images with the corresponding panchromatic bands (Laben and Brower, 2000). For experimental convenience, we further resampled the GF1 and GF2 pan-sharpened images to 2 m and 1 m, respectively.

Ground-truth agricultural fields were obtained by manually delineating the high-resolution images of the study areas at the local scale.

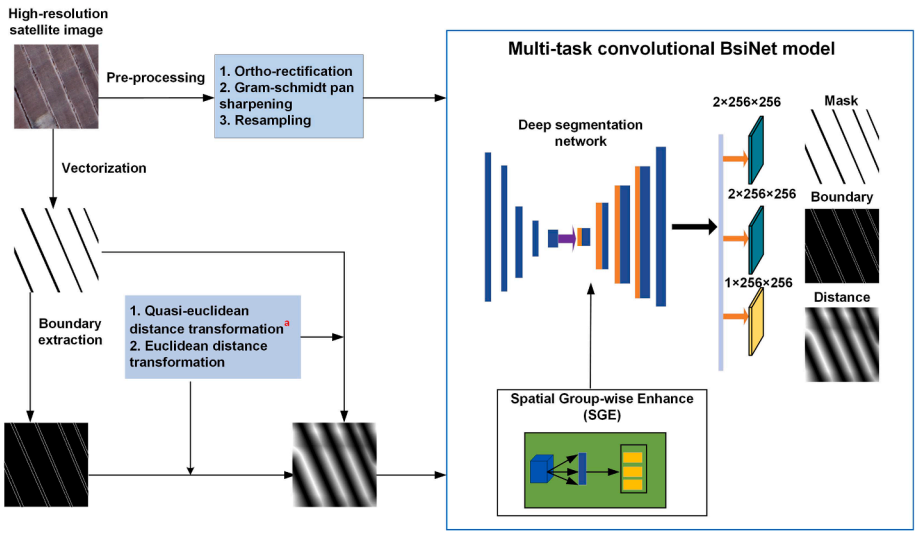

Fig. 2. Workflow of the proposed field delineation using BsiNet from high-resolution satellite images.

We selected three training and one testing areas for the Xinjiang dataset, two training and one testing areas for the Fujian dataset, one training area and one testing area for the Shandong dataset, and two training areas and one testing area for the Sichuan dataset, corresponding to the red and yellow boxes in Fig. 1. The field samples within each training area were further randomly partitioned into training and validation samples, while field samples within testing areas were used as testing samples.

3. Methods

We considered the field delineation task a multi-task semantic segmentation problem, and used BsiNet to do the semantic segmentation. Fig. 2 shows the workflow of the proposed BsiNet method. BsiNet was constructed based upon a single encoder-decoder segmentation network Psi-Net, consisting of a mask, boundary, and distance components. More specifically, a boundary map was derived from a mask map, while a distance map was obtained by performing a quasi-Euclidean distance transformation on the boundary. The distance map alleviates the errors of isolated segmentation in the results. Moreover, we incorporated a spatial group-wise enhancement (SGE) module into the decoder to improve the identification of small fields.

3.1. Multi-task network model Psi-Net

For semantic segmentation, Psi-Net is a popular multi-task model with a UNet-like encoder-decoder architecture, consisting of one encoder and three decoders (Murugesan et al., 2019). The encoder comprises repeated downsampling operations after a 3×3 convolution with a stride of 1 (Fig. 3A). The last downsampling refers to a 4×4 max pooling and is used as the input of decoders. The three decoders are parallel and have similar structures, corresponding to a mask, boundary, and distance decoders. Each decoder concatenates image features from the corresponding encoder layer, forming multi-scale features. The features of the three tasks are then extracted by the last layer of the decoders through a 1×1 convolution.

Here the mask is seen as the core task, while boundary and distance estimation are auxiliary tasks. The two auxiliary tasks are used to obtain the target's shape and boundary information and refine the mask prediction of agricultural fields. Moreover, the mask and boundary predictions are classification tasks, and the distance map estimation is a regression task.

3.2. BsiNet

3.2.1. SGE attention module

Inspired by Psi-Net, we propose a new multi-task network based upon a single encoder-decoder Psi-Net and a SGE attention module, called BsiNet (Fig. 3B). Unlike Psi-Net, our network uses one single decoder to produce three types of information mask, boundary, and distance. We argue that such a decoder is sufficient to replace the three decoders provided by a Psi-Net and thus reduce computation time. Another advantage is that BsiNet learns information on distance and contours together. To improve the identification of small fields, we added the SGE attention module at the end of BsiNet (Fig. 3B). Generally, high-level image features generated through a CNN contain apparent noise and have ill-distributed feature responses (Li et al., 2019). To deal with these problems, we incorporate a SGE attention module into BsiNet to extract enhanced features, utilising the global information of each channel of the image features and scaling the importance of different channels.

Let X be the matrix of convolutional features with c channels (Fig. 3B). Each channel has a length h and a width w. The SGE module first divides the feature matrix X into k groups according to channel dimension, where $X = \{x_1, \cdots, x_m\}, m = h \times w$, and $x_i \in R^{\frac{c}{k}}$ indicates the feature vector of a specific point in the $h \times w$ space (Li et al., 2019). To retain key feature responses while alleviating feature noise, SGE reassigns the importance of each point x_i by calculating the similarity between its feature vector with a global feature vector g that is obtained by spatially averaging all possible points. Suppose f indicates a spatial averaging function, then the global feature vector g is defined as:

$$g = f(x) = \frac{1}{m}\sum_{i=1}^{m} x_i. \tag{1}$$

We obtain an importance coefficient c_i for each x_i, as,

$$c_i = g \cdot x_i = \|g\| \cdot cos\|x_i\| \cdot cos(\theta_i) \tag{2}$$

where θ_i indicates the angle between x_i and g in feature space.

To avoid the influence of the biased magnitude of coefficients between various samples, the importance coefficient c_i is further normalised as:

$$\widehat{c_i} = \frac{c_i - u_d}{\delta_c + \epsilon}, u_d = \frac{1}{m}\sum_{j}^{m} c_j, \delta_c^2 = \frac{1}{m}\sum_{j}^{m}(c_j - u_c)^2 \tag{3}$$

where u_d and δ_c^2 are the mean and variance of the importance coefficient

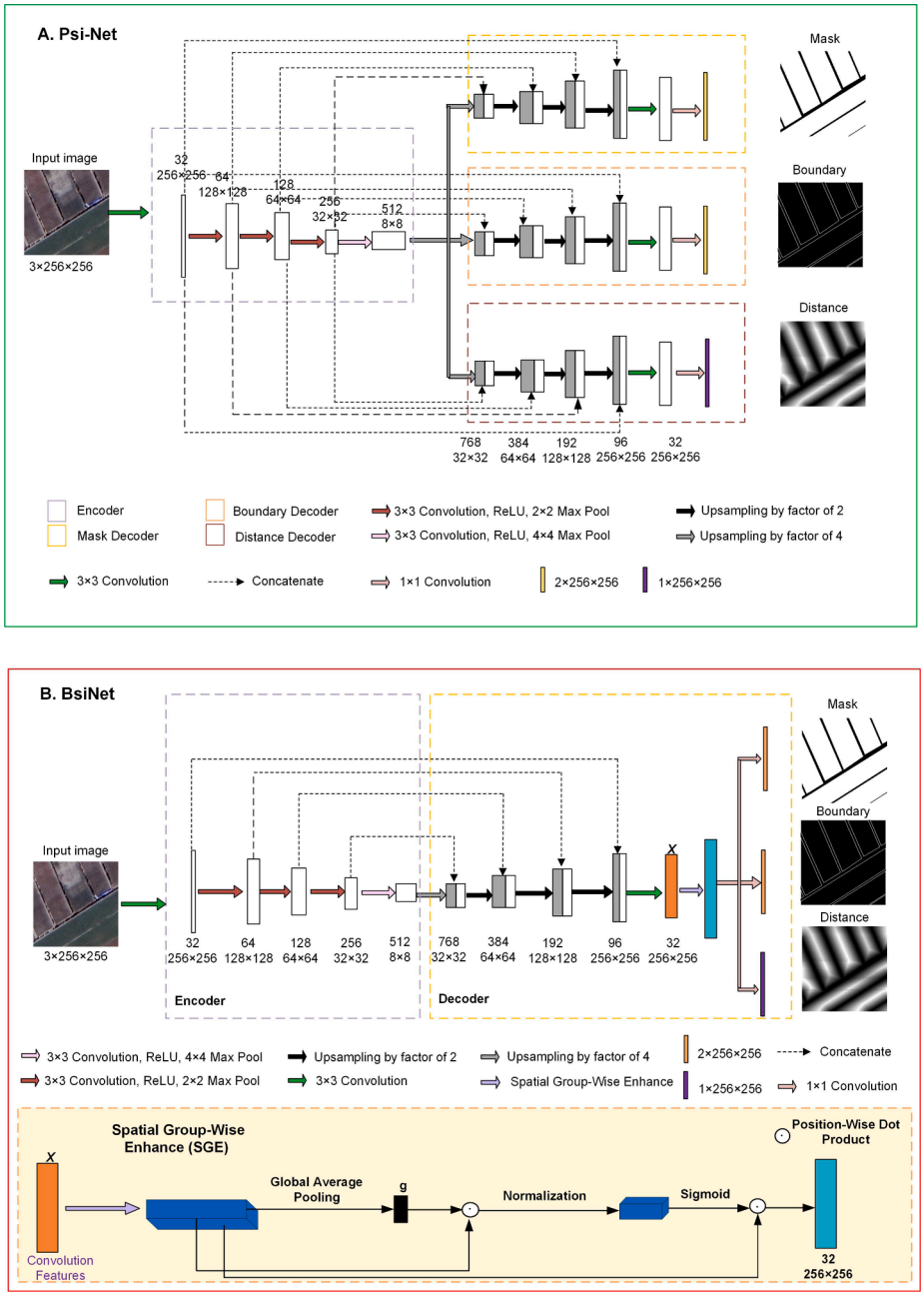

Fig. 3. Architecture of multi-task Psi-Net (A) adjusted from Murugesan et al. (2019), and the proposed BsiNet (B).

c_i and ϵ is a constant ensuring the denominator is positive. Furthermore, we adjust the normalized $\widehat{c}_i$ using a linear function to ensure that the normalization can represent the identity transformation and that the normalized operation can be restored, i.e.,

$$p_i = \alpha \widehat{c}_i + \beta \tag{4}$$

where α and β are function parameters, and p_i is the adjusted coefficient. Here we set $\alpha = \beta$ equal to k (i.e., 32), referring to Li et al. (2019).

The enhanced feature vector $\widehat{x}_i$ is obtained by scaling the original feature vector x_i with an importance coefficient that is derived by applying a sigmoid function δ over the adjusted coefficient p_i,

$$\widehat{x}_i = x_i \cdot \delta(p_i). \tag{5}$$

3.2.2. Quasi-Euclidean distance

In a multi-task network, the distance to the boundaries helps to alleviate the errors of isolated segmentation (Tan et al., 2018). Obtaining distance features is thus essential to multi-task model performance. Unlike Psi-Net, which obtains distance features based upon the corresponding mask channel, we use the quasi-Euclidean distance transformation to obtain distances from the corresponding boundary map. The quasi-Euclidean distance considers the influence of horizontal, vertical, and diagonal pixels (Paglieroni, 1992). It provides better performance in the distance metric compared to the Euclidean distance that only considers diagonal pixels (Singh et al., 2020).

3.2.3. Loss function

In multi-task learning, we often set different loss functions for other tasks. Here, we refer to Murugesan et al. (2019) for constructing loss functions. Mask and boundary predictions adopt a negative log-likelihood (NLL) loss, while the distance map estimation adopts a mean square error (MSE) loss. Let l_{cls} be the pixel-wise classification loss

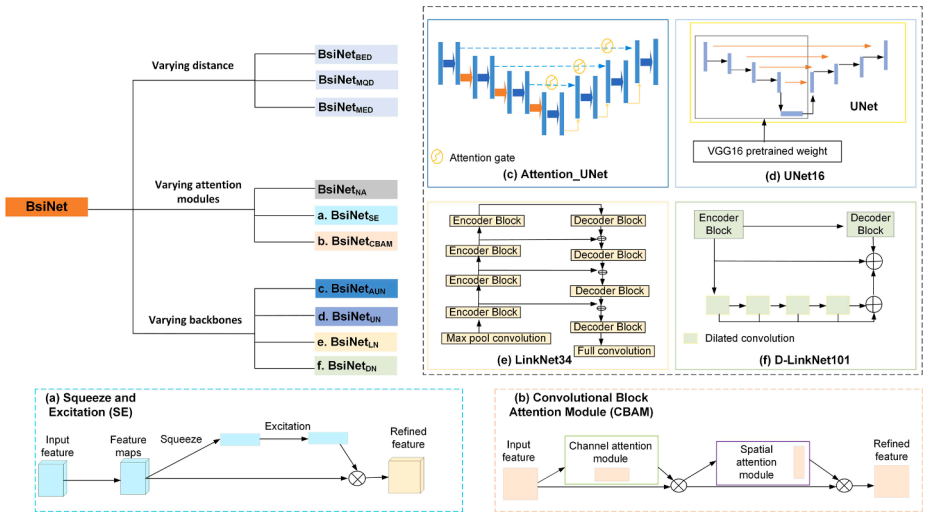

Fig. 4. BsiNet variants. It consists of 10 multi-task networks via varying distance, attention modules and backbones. Backbones adjusted from AttentionUNet, UNet16, LinkNet34, and D-Linknet101 (Oktay et al., 2018; Iglovikov and Shvets, 2018; Chaurasia and Culurciello, 2017; Zhou et al., 2018).

(i.e., NLL), x be the pixel position in image space R. The loss function of mask and boundary prediction tasks with N input instances are defined as:

$$l_{cls} = -\frac{1}{N}\sum_{i=1}^{N} log p_{mc}(x; l_{mc}(x)) \tag{6}$$

where $p_{mc}(x; l_{mc}(x))$ indicates the predicted probability, by softmax activation, of the true label l_{mc}. To obtain the MSE loss (i.e., l_{reg}) of the regression task, namely the distance map estimation, we calculate the bias from predicted distance $\widehat{D}(x)$ and ground truth distance $D(x)$. More specifically, l_{reg} is obtained as:

$$l_{reg} = \frac{1}{N}\sum_{i=1}^{N} (\widehat{D}(x) - D(x))^2. \tag{7}$$

For multiple tasks, we aggregate the loss values of different subtasks to formulate the total loss:

$$l_{total} = w_1 \cdot l_{cls1} + w_2 \cdot l_{cls2} + w_3 \cdot l_{reg} \tag{8}$$

where l_{cls1}, l_{cls2}, and l_{reg} refer to the loss values of the three subtasks, corresponding to mask, boundary, and distance prediction, and w_1, w_2, and w_3 are the corresponding weight coefficients. For comparison, we use the above loss functions for all multi-task networks, and set equal weights, i.e., $w_1{=}w_2{=}w_3$ for all separate sub-tasks, referring to Murugesan et al. (2019).

3.3. Variations on BsiNet and comparison with existing models

3.3.1. Variations on BsiNet

(A) Varying distance features generation methods

In a multi-task network, the subtask for distance map estimation regularises the boundaries and shape of the objects of interest segmented by the subtask for core task predication. Murugesan et al. (2019) found that for multi-task semantic segmentation, the distance map derived from the Euclidean distance transform of the boundary map is better than that of the mask map of the ground-truth fields. Here, we compare BsiNet with three variants (Fig. 4): $BsiNet_{MED}$ and $BsiNet_{BED}$ derive the distance maps using the conventional Euclidean distance transformation of the mask image and the boundary image, respectively, and $BsiNet_{MQD}$ derives the distance map using the quasi-Euclidean distance transformation of the mask image.

(B) Varying attention modules

To investigate the effect of attention modules on BsiNet, we varied the models using different attention modules, i.e., squeeze and excitation (SE) (Hu et al., 2018) and convolutional block attention module (CBAM) (Woo et al., 2018), leading to two multi-task models, i.e., $BsiNet_{SE}$ and $BsiNet_{CBAM}$, respectively. Moreover, we removed the attention module from BsiNet, labeled as $BsiNet_{NA}$. The choice of these two attention modules was motivated by the following:

SE, improves the representation power of a network by adaptively recalibrating channel-wise feature responses based upon the interdependencies of each channel (Hu et al., 2018). It is a general-purpose attention module for semantic segmentation and image classification (Fig. 4a).

CBAM, derives a channel and a spatial attention map from a convolutional feature map by two sequential sub-modules, referring to channel and spatial modules (Woo et al., 2018). The global responses of convolutional features are combined from the channel and spatial perspectives (Fig. 4b). In contrast to the SE module, CBAM improves the characterization of small objects, and the corresponding agreement with their positions and geometric accuracy.

(C) Varying backbones

To vary the backbones of BsiNet, we replaced its encoder-decoder components with AttentionUNet (Oktay et al., 2018), UNet16 (Iglovikov and Shvets, 2018), LinkNet34 (Chaurasia and Culurciello, 2017), and D-linkNet101 (Zhou et al., 2018), leading to $BsiNet_{AUN}$, $BsiNet_{UN}$, $BsiNet_{LN}$, $BsiNet_{DN}$, respectively. The motivations for selecting these methods are as follows:

AttentionUNet, is a variant of UNet, which adds an attention gate module at the end of each jump junction (Fig. 4c). The attention gate derives attention responses on the targets with various shapes and sizes and improves model sensitivity for dense label predictions.

UNet16, uses a pre-trained VGG16 as its encoder for initialising model weights, rather than training the model from scratch with random initialisation weights like a classical UNet (Fig. 4d). Such fine-tuning practice improves model convergence and generalizability.

LinkNet34, connects the encoder of a network with its decoder in a different way than those encoder-decoder networks of the UNet family (Chaurasia and Culurciello, 2017). Usually, some spatial information may be lost by performing multiple downsampling (e.g., pooling) operations in the encoder, which is hard to recover using the decoder. The LinkNet34 retains the spatial information by removing the downsampling operation, commonly done in a conventional encoder-decoder network, from each encoder block (Fig. 4e).

D-LinkNet101, uses a LinkNet with a pre-trained encoder as its

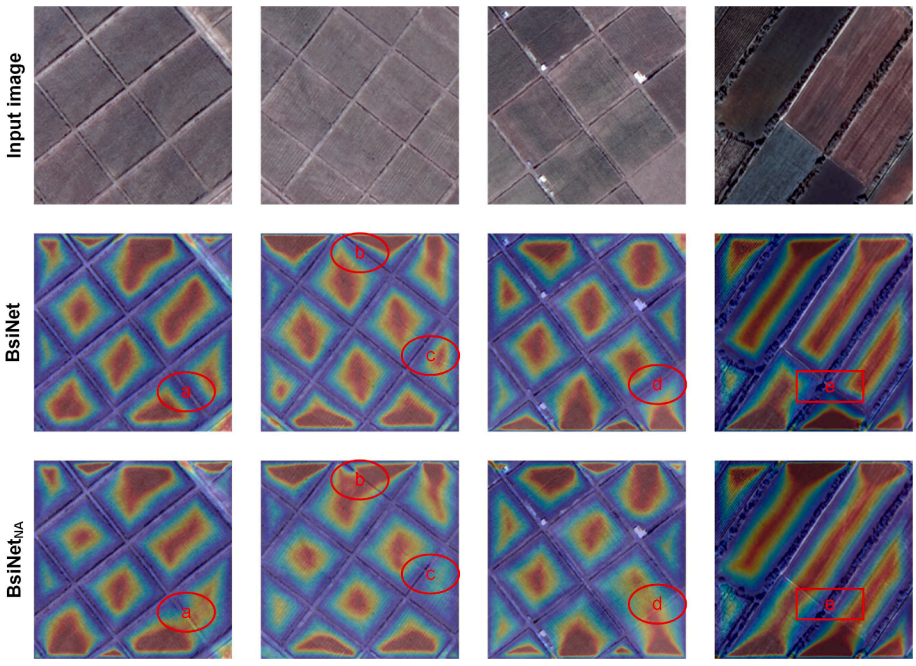

Fig. 5. Attention features extracted by BsiNet and *BsiNet*$_{NA}$.

backbone, and adds a set of dilated convolutions between the decoder and encoder (Zhou et al., 2018) (Fig. 4f). Adding these dilation convolutions expands the feature's receptive field while retaining feature maps' resolution. D-LinkNet101 is thus capable of using global information in a broad context for feature extraction.

3.3.2. Comparison with existing models and accuracy assessment

We compared BsiNet with three existing methods, two multi-task models, i.e., ResUNet-a D7 and Psi-Net, and a single-task UNet. Here, Psi-Net serves as a baseline for field delineation by multi-task networks, while ResUNet-a D7 was recently proposed to delineate fields from Sentinel-2 images (Waldner and Diakogiannis, 2020). UNet is a popular single-task encoder-decoder network, and was investigated by Xia et al. (2018) for field delineation.

3.4. Accuracy assessment and performance evaluation

We used a set of object-based error indices to evaluate both the attribute and geometrical accuracy of extracted fields (Persello and Bruzzone, 2009; Li et al., 2015). Let M_i be a classified object, $i = 1,\cdots,m$, and O_i be the reference object that has the highest overlapping area with M_i. Let $area(M_i)$ and $area(O_i)$ be the areas of M_i and O_i, respectively, and $area(M_i \cap O_i)$ be their overlapping area. The over-classification error $OC(M_i)$ and under-classification error $UC(M_i)$ are then defined as:

$$OC(M_i) = 1 - \frac{area(M_i \cap O_i)}{area(O_i)}, \tag{9}$$

$$UC(M_i) = 1 - \frac{area(M_i \cap O_i)}{area(M_i)}. \tag{10}$$

We then use $OC(M_i)$ and $UC(M_i)$ to obtain the total error $TC(M_i)$,

$$TC(M_i) = \sqrt{\frac{OC(M_i)^2 + UC(M_i)^2}{2}}. \tag{11}$$

Based upon the obtained $OC(M_i)$, $UC(M_i)$, and $TC(M_i)$ indices, three global error indices are derived, referring to as GOC, GUC, and GTC:

$$GOC = \sum_{i=1}^{m} \left(OC(M_i) \times \frac{area(M_i)}{\sum_{i=1}^{m} area(M_i)} \right), \tag{12}$$

$$GUC = \sum_{i=1}^{m} \left(UC(M_i) \times \frac{area(M_i)}{\sum_{i=1}^{m} area(M_i)} \right), \tag{13}$$

$$GTC = \sum_{i=1}^{m} \left(TC(M_i) \times \frac{area(M_i)}{\sum_{i=1}^{m} area(M_i)} \right). \tag{14}$$

Moreover, we used Giga floating-point operations per second (i.e., GFLOPs) and the number of parameters (Hu et al., 2018) to evaluate the efficiency of BsiNet.

3.5. Implementation details

To facilitate model training, we cropped high-resolution satellite images into smaller images with a size of 256 × 256 pixels. In this work, we conducted data augmentation to enlarge the training dataset. More specifically, a sample image was flipped horizontally and vertically, followed by a clockwise image rotation of 90 and 180 degrees. To reduce the influence of invalid samples, we removed images with field pixel values of less than 40%. In total, we generated (2528, 150, 336), (5000, 150,160), (5676, 150, 180), and (4128, 150, 180) samples for the training, validation, and testing of the Xinjiang, Fujian, Shandong, and Sichuan datasets. To train the BsiNet, we used the Adam optimiser with a batch size of 4 and an initial learning rate of 0.0001, and trained 150 epochs using an NVIDIA GeForce RTX 2080Ti GPU. The same parameters were used to train BsiNet variants and the existing Psi-Net, ResUNet-a D7, and UNet models.

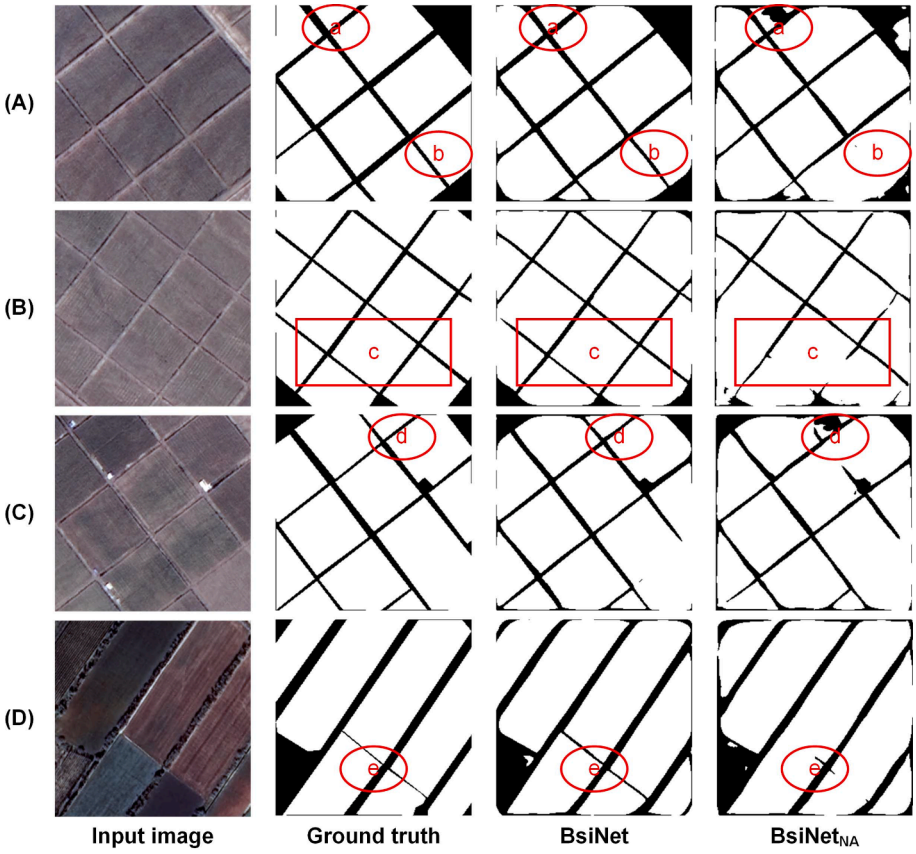

Fig. 6. Extraction results obtained by BsiNet and *BsiNet*$_{NA}$.

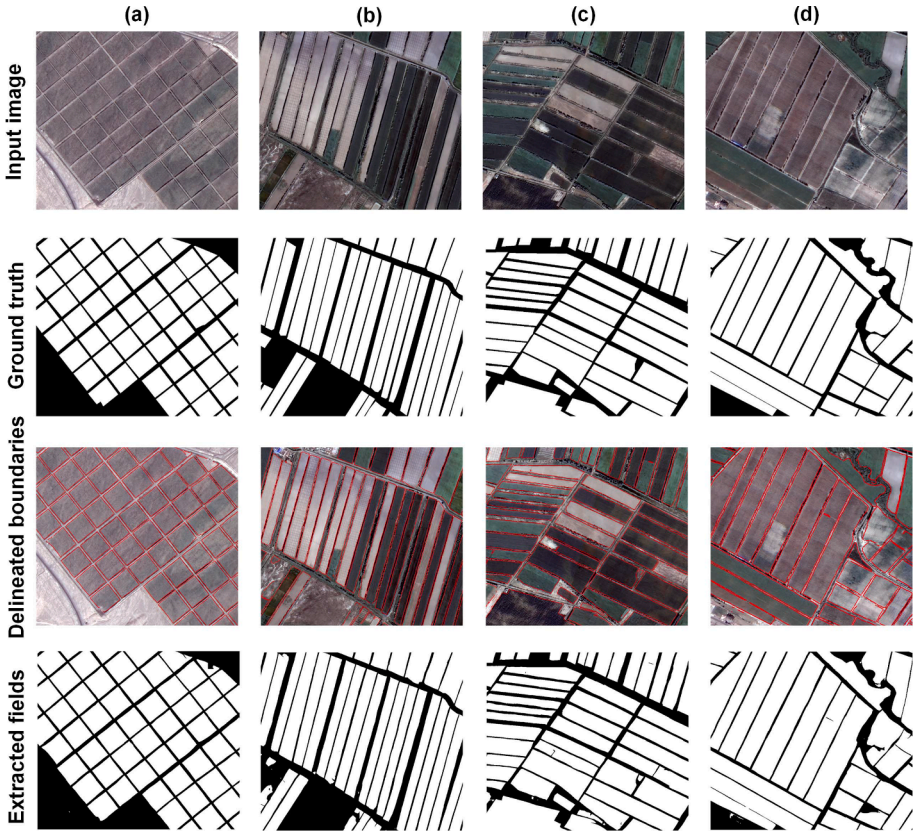

Fig. 7. Examples of agricultural fields and field boundaries extracted by BsiNet on the Xinjiang dataset.

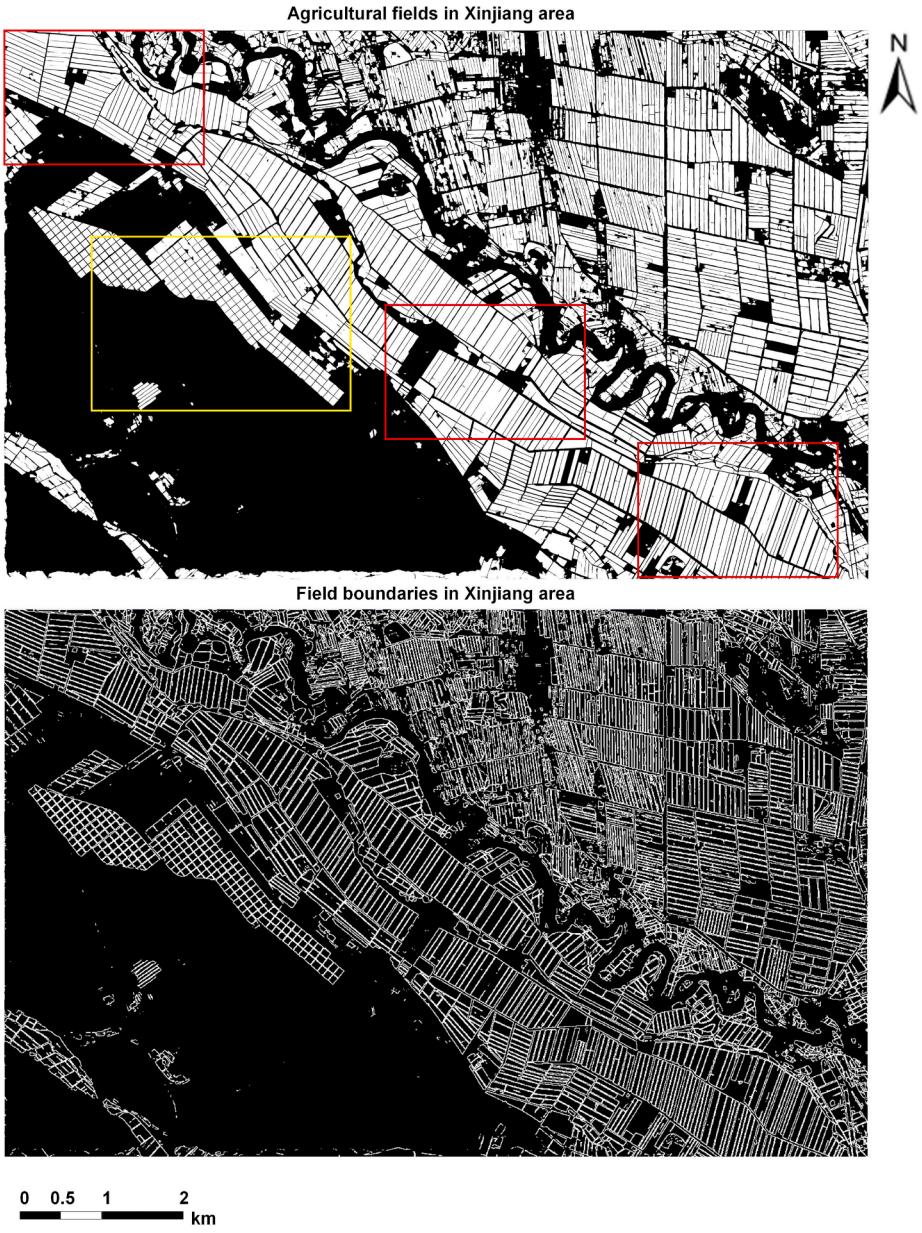

Fig. 8. Results of agricultural fields and field boundaries extracted by BsiNet on the whole Xinjiang dataset. Red and yellow boxes represent the training and testing areas. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparing geometric errors of the extraction results between 14 different methods on the Xinjiang dataset.

Groups	Methods	GOC	GUC	GTC
Proposed	**BsiNet**	**0.062**	**0.042**	**0.062**
Distance	$BsiNet_{MED}$	0.123	0.052	0.111
	$BsiNet_{BED}$	0.115	0.058	0.105
	$\boldsymbol{BsiNet_{MQD}}$	**0.112**	**0.055**	**0.101**
Attention mechanism	$BsiNet_{NA}$	0.129	0.058	0.115
	$BsiNet_{SE}$	0.120	0.062	0.110
	$\boldsymbol{BsiNet_{CBAM}}$	**0.085**	**0.082**	**0.100**
Backbones	$BsiNet_{ATN}$	0.272	0.057	0.216
	$\boldsymbol{BsiNet_{UN}}$	**0.208**	**0.053**	**0.172**
	$BsiNet_{DN}$	0.229	0.046	0.183
	$BsiNet_{LN}$	0.234	0.046	0.185
Existing models	*Psi-Net*	0.170	0.049	0.142
	ResUNet-a D7	0.205	0.122	0.204
	UNet	0.282	0.084	0.224

4. Results

4.1. SGE attention

We compared the feature maps of agricultural fields derived from BsiNet to those without an SGE attention module, i.e., $BsiNet_{NA}$. For the SGE module in BsiNet, we set the number of groups to 32, as recommended by Li et al. (2019). The feature map generated from an attention module highlights the areas of images to be identified as agricultural fields (Fig. 5). The figure shows that BsiNet more accurately emphasised every field (red box c in Fig. 5) compared to $BsiNet_{NA}$. Moreover, $BsiNet_{NA}$ failed to accurately locate fields, leading to adjacent fields (red circles a, b, d, and e in Fig. 5), while BsiNet was able to identify these fields more accurately. The agricultural fields obtained from BsiNet and $BsiNet_{NA}$ for the four examples are shown in Fig. 6. In general, the fields obtained from BsiNet are closer to the ground truth fields than $BsiNet_{NA}$ in shape completeness and boundary separation. Notably, when the SGE attention module was added, the boundary delineation of small fields

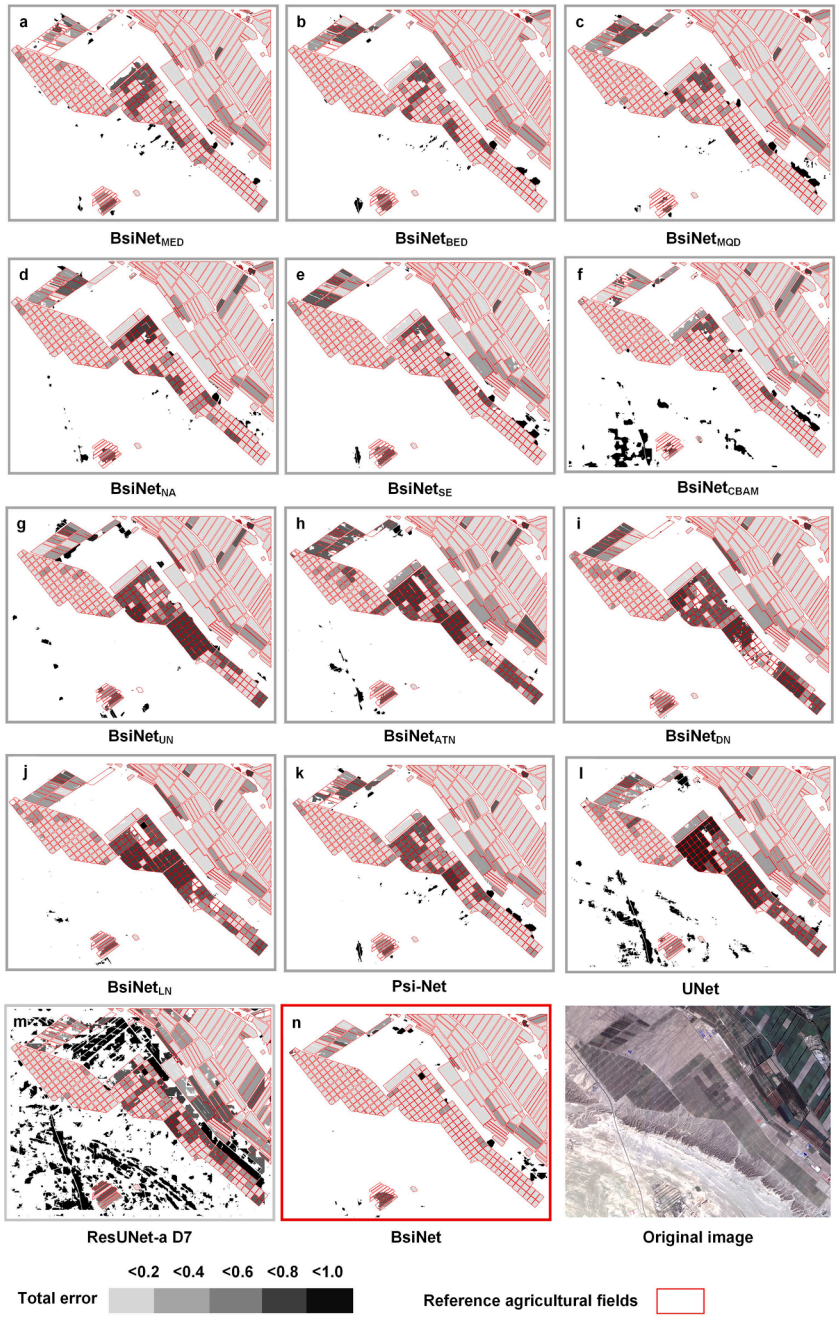

Fig. 9. Distribution of the GTC error of the agricultural fields extracted by BsiNet and other methods on the Xinjiang dataset.

Table 3
Comparing computation load and model parameters between different models. GFLOPs represent Giga floating-point operations per second and 1 M equals 10^6.

Models	Image size	GFLOPs	Parameters
BsiNet	***256 × 256***	**13.30**	**7.84 M**
UNet	256 × 256	46.13	31.04 M
Psi-Net	256 × 256	32.61	14.11 M
ResUNet-a D7	256 × 256	70.90	131.47 M

(red circles a and d in Figs. 6A and 6C) and adjacent fields (red circles b, c, and e in Figs. 6A, 6B, and 6D) was improved.

Table 4
Geometric errors of extracted fields by BsiNet, *BsiNet$_{CBAM}$*, and ResUNet-a D7 on the Fujian dataset.

Types	GOC	GUC	GTC
BsiNet	**0.322**	**0.165**	**0.291**
BsiNet$_{CBAM}$	0.359	0.238	0.353
ResUNet-a D7	0.817	0.073	0.583

4.2. Agricultural fields obtained by BsiNet on the Xinjiang GF1 image

We applied BsiNet to delineate agricultural fields from the Xinjiang GF1 image. It was trained from scratch using the training dataset collected from three subsets of the Xinjiang area. The parameter setting for model training was detailed in Section 3.5. Fig. 7 displays extracted

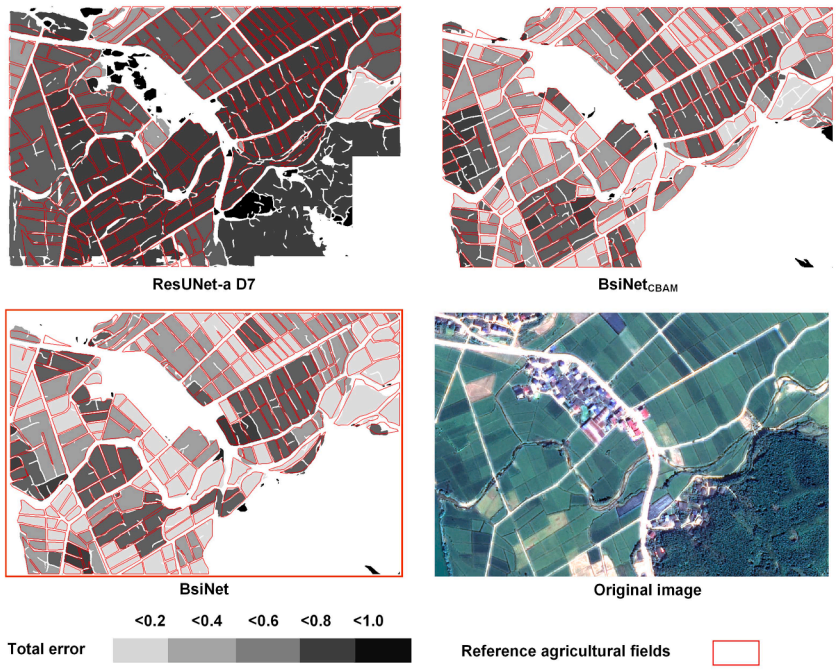

Fig. 10. Distribution of the GTC error of the agricultural fields extracted by BsiNet, *BsiNet*$_{CBAM}$, and ResUNet-a D7 on the Fujian dataset.

results by BsiNet for the different examples. BsiNet produced well-delineated boundaries for regions characterised by regular fields (Figs. 7a-c) and irregular fields (Fig. 7d), and was able to deal with the spectral heterogeneity between different fields (Figs. 7b and 7c). Moreover, the results of fields delineation for the entire GF1 image using BsiNet are shown in Fig. 8. This figure shows that BsiNet produced clear field boundaries for the Xinjiang dataset.

4.3. Accuracy assessment on the Xinjiang dataset

We compared field delineation using BsiNet with ten BsiNet variants, two existing multi-task Psi-Net and ResUNet-a D7, and an existing single task UNet. The ten BsiNet variants were formulated by varying distance features (i.e., *BsiNet*$_{MED}$, *BsiNet*$_{BED}$, and *BsiNet*$_{MQD}$), attention modules (i.e., *BsiNet*$_{NA}$, *BsiNet*$_{SE}$, and *BsiNet*$_{CBAM}$), and backbones (i.e., *BsiNet*$_{UN}$, *BsiNet*$_{DN}$, *BsiNet*$_{LN}$, and *BsiNet*$_{ATN}$). All methods were tested using the same testing dataset. Table 2 evaluates the geometric accuracy of extracted results using GUC, GOU, and GTC error indices. This table shows that BsiNet yielded the lowest GOC of 0.062, GUC of 0.042, and GTC of 0.062 on the Xinjiang dataset as compared to the other 13 methods. *BsiNet*$_{MQD}$ produced the best results among the BsiNet variants within the distance group, leading to the smallest GOC and GTC errors. This implies that the use of quasi-Euclidean transformation improves delineation results. Among the BsiNet variants within the attention group, *BsiNet*$_{CBAM}$ generated better results compared with *BsiNet*$_{NA}$ and *BsiNet*$_{SE}$, resulting in lower GOC and GTC errors. Results imply that incorporating spatial and channel attention can further improve the geometric accuracy of field extraction. Furthermore, using BsiNet produced lower GOC, GUC, and GTC values than *BsiNet*$_{CBAM}$. This indicates that SGE attention shows a better performance than CBAM attention. Table 2 also shows that BsiNet outperforms *BsiNet*$_{UN}$, *BsiNet*$_{DN}$, *BsiNet*$_{LN}$, and *BsiNet*$_{ATN}$. The multi-task networks produced lower GOC, GUC, and GTC errors than the single-task UNet model.

Fig. 9 visualizes the GTC error of extracted fields by BsiNet and the other 13 methods. Among the BsiNet variants within the distance feature groups, *BsiNet*$_{MED}$ (Fig. 9a) and *BsiNet*$_{BED}$ (Fig. 9b) have higher segmentation errors than *BsiNet*$_{MQD}$ (Fig. 9c). Among the BsiNet variants within the attention group, Figs. 9d and 9f exhibit a higher number of segmentation errors in *BsiNet*$_{NA}$ and *BsiNet*$_{CBAM}$ as compared with *BsiNet*$_{SE}$ (Fig. 9e). Among the BsiNet variants within the backbone groups, *BsiNet*$_{UN}$, *BsiNet*$_{ATN}$, and *BsiNet*$_{DN}$ (Figs. 9g, 9h, and 9i) have higher segmentation errors as compared with *BsiNet*$_{LN}$ (Fig. 9j). This figure shows that BsiNet yields the lowest segmentation errors (Fig. 9n). Moreover, Psi-Net (Fig. 9k) produces lower segmentation errors than existing models UNet (Fig. 9l) and ResUNet-a D7 (Fig. 9m).

We also compared the computation load of the model parameters of the BsiNet with existing models (Table 3). This table shows that when the input image size is the same, BsiNet has the smallest GFLOPs (13.30) and fewest model parameters (7.84 M), whereas ResUNet-a D7 has the most model parameters (131.47 M), leading to the highest computation load (70.90 GFLOPs).

4.4. Agricultural fields obtained by BsiNet on the Fujian GF2 image

We further evaluated BsiNet on the Fujian GF2 image. We trained BsiNet from scratch using the training dataset collected from two subsets of the Fujian area, and tested on one subset (Fig. 1). We compared the results of field delineation between BsiNet and *BsiNet*$_{CBAM}$, which achieved the second-best accuracy on the Xinjian dataset (see Table 2). Furthermore, we compared BsiNet with ResUNet-a D7, which was recently used for field delineation from Sentinel-2 images. Table 4 lists the geometrical errors of extracted fields by the three methods. This table shows that BsiNet produced the lowest extraction errors than *BsiNet*$_{CBAM}$ and ResUNet-a D7 for GOC and GTC, while the highest GOC error was produced by ResUNet-a D7. It implies that the ResUNet-a D7 developed based upon medium-resolution images may be not suitable for field delineation from high-resolution images. Fig. 10 visualises the GTC error of the fields extracted by BsiNet, *BsiNet*$_{CBAM}$ and ResUNet-a D7. Clearly, *BsiNet*$_{CBAM}$ and ResUNet-a D7 produced higher GTC errors than BsiNet. Fig. 11 shows the extracted agricultural fields for the whole Fujian dataset obtained with BsiNet. BsiNet generated smooth fields in the Fujian area, where agricultural fields are commonly found with irregular shapes. Moreover, the results also show that BsiNet achieved excellent extraction results in small field areas (Fig. 11e).

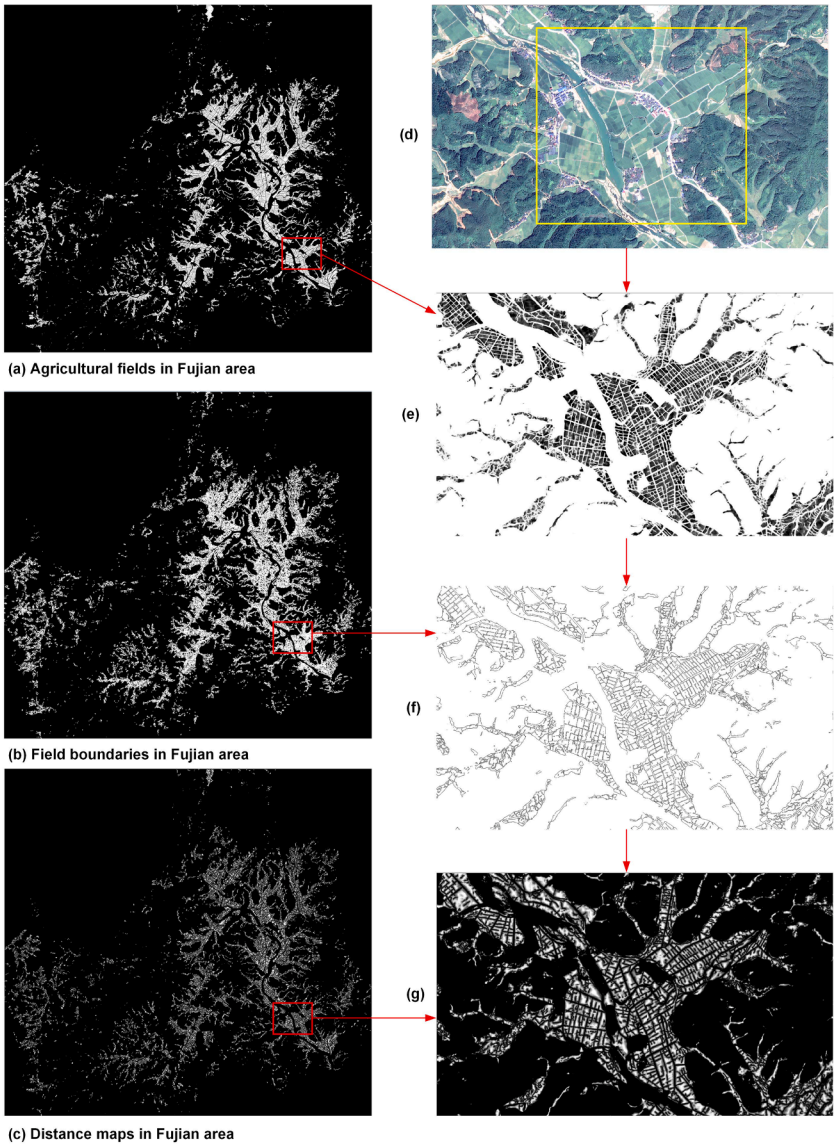

Fig. 11. Results of field, boundary, and distance maps extracted by BsiNet on the Fujian dataset. The yellow box represents the testing area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5. Agricultural fields obtained by BsiNet on the Shandong and Sichuan GF2 images

Next, we applied BsiNet to delineate fields on the Shandong and Sichuan GF2 images. The (GOC, GUC, GTC) errors for the Shandong and Sichuan datasets are (0.024, 0.148, 0.115) and (0.029, 0.335, 0.241). Fig. 12 shows the results of field extraction on the two datasets. Although BsiNet produced satisfactory results for all four datasets, we can see that regular fields in Xinjiang and Shandong areas were extracted with lower errors than irregular fields in Fujian and Sichuan areas.

5. Discussion

This study investigated to use multi-task neural networks to delineate agricultural fields from high-resolution satellite images. Inspired by Psi-Net network, we presented a new multi-task network, referred to as BsiNet, to learn high-level boundary and shape information of fields. We considered field delineation a multi-task semantic segmentation problem, and used BsiNet to extract field mask, boundaries and distance. Our first finding is that BsiNet evidently improves the extraction accuracy of agricultural fields, and can be used to extract fields over different areas. BsiNet is also more efficient than existing models, i.e., Psi-Net (Mugesan et al., 2019) and ResUNet-a D7 (Waldner and Diakogiannis, 2020). The second finding is that the addition of auxiliary tasks retains the shape and boundary information of fields, yielding smooth fields. This further strengthens the applicability of the multi-task network developed in Murugesan et al. (2019). We observed that the shape and boundary information was also highlighted in a conventional image segmentation problem (Li et al., 2015). This shows a way that it is of high potential to design more advanced multi-task models of semantic segmentation by referencing the designing logic of conventional image segmentation methods.

Recent studies have demonstrated the effectiveness of distance maps in reducing errors in isolated segmentation (Tan et al., 2018). In a multi-task network, we consider predicting distance map as an auxiliary task to generate smooth fields. The distance maps provide rich information (e.g., the shape and boundary of the agricultural fields) of the extracted fields. Therefore, training the multi-task network while predicting the distance maps corresponds to enforcing boundary and shape constraints for the delineation task. Moreover, studies have shown that adding boundary direction and intensity information can further constrain the

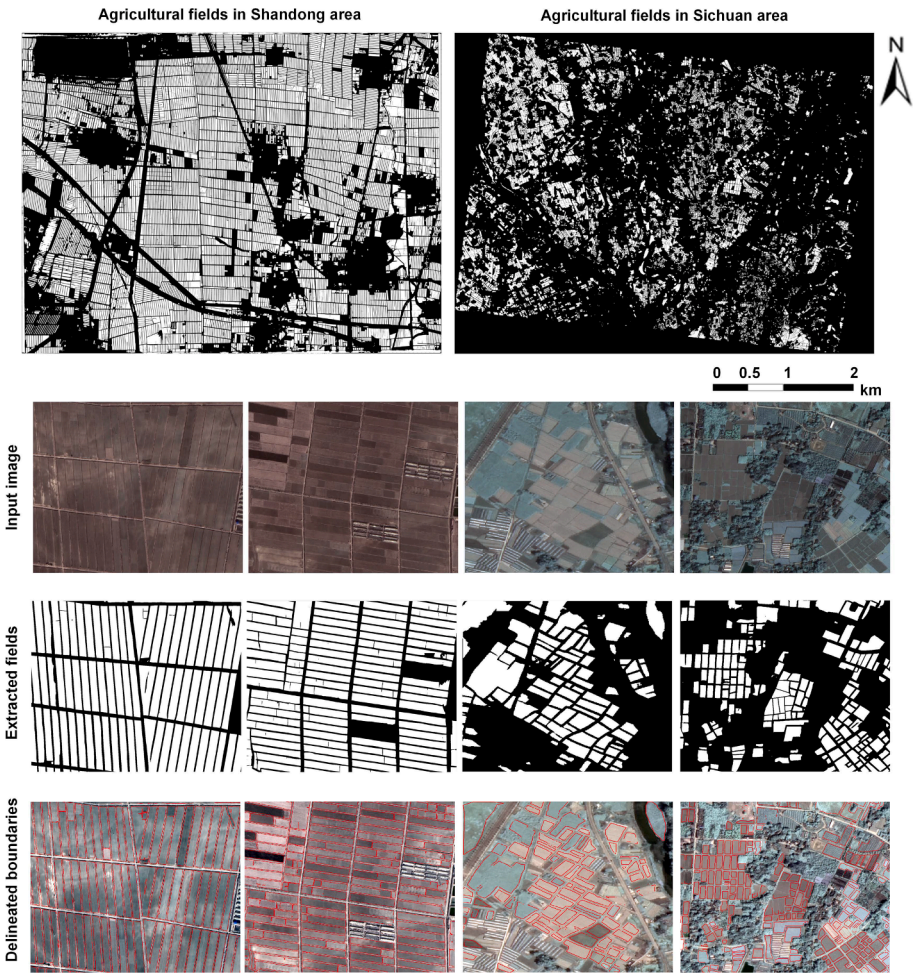

Fig. 12. Results of agricultural fields and field boundaries extractions using BsiNet on the Shandong and Sichuan datasets.

object boundaries (Yuan et al., 2020), which still requires further research.

To improve the identification of small fields, we added the SGE attention module to the multi-task network. Our results show that this emphasizes the learning of field-related features (e.g., field shape and boundaries) and ignores unrelated features, thus generating well-delineated fields. The boundaries of small fields are ambiguous, and extracting them is a challenging task. Recent studies have explored combining edge detection methods to extract small fields (Persello et al., 2019). This paper capitalized on the advantages of SGE attention in small object identification, applying it to field extraction and the results show its potential for extracting small fields.

We investigated the effect of the backbone network on BsiNet. The findings suggest that this did not improve the accuracy of the results. We also found that the models with a pre-trained network backbone (e.g., $BsiNet_{LN}$ and $BsiNet_{DN}$) failed to improve field extraction accuracy, which may be related to weight initialization strategy and loss function settings. For this, we still need further research.

The effectiveness of BsiNet was tested for agricultural fields delineated on the Xinjiang GF1 and Fujian, Shandong, and Sichuan GF2 images. Most Xinjiang and Shandong fields have a regular shape and a clear boundary, leading to excellent extraction results. For the Fujian and Sichuan study areas, the fields are irregular and have ambiguous boundaries, and field extraction can be more challenging (Fig. 1). Based upon our analysis, our proposed method is applicable to both scenarios. In addition, we did transferability tests, namely transferring the BsiNet trained with the Sichuan GF2 image to extract fields on the Fujian GF2 image, and transferring the BsiNet trained with the Xinjiang GF1 image to extract fields on the Shandong GF2 image (referring to Supplementary document). We observe that BsiNet has a high transferability, promoting the use of this method to wider applications. Moreover, we found that the existing ResUNet-a D7, which was developed based upon medium-resolution satellite images, failed to produce satisfactory results from high-resolution images. In addition, we also conducted extra experiments to evaluate the applicability of BsiNet for medium-resolution Sentinel-2 images (referring to Supplementary document). Our results showed that the proposed BsiNet also has a high potential to be applied to medium-resolution satellite images like Sentinel-2 images.

Conventional convolutional neural networks have difficulty in obtaining closed field boundaries. We used the multi-task BsiNet to identify agricultural fields, and obtained satisfactory extraction results on the Xinjiang and Shandong datasets. While our proposed method extracted fields with clear boundaries and shapes, it yielded non-closed field boundaries on the Fujian and Sichuan datasets which have irregular and complex fields. To deal with such an area, a possible solution is to extract field boundaries incorporating edge detection models or instance segmentation networks like Mask R-CNN (He et al., 2017). Studies have shown that they can produce closed boundaries in one segmentation, thereby avoiding complex post-processing work (Ruiz-Santaquiteria et al., 2020).

6. Conclusions

This study developed a multi-task neural network, called BsiNet, to extract agricultural fields using high-resolution satellite images. It combines the single encoder-decoder Psi-Net with a spatial group-wise

enhancement (SGE) module to improve the accuracy and efficiency of field delineation. We conclude that:

(1) The proposed BsiNet achieved the highest extraction accuracy compared with 10 BsiNet variants (with different distance measures, attention modules, and backbone structures) and three existing methods (i.e., Psi-Net, ResUnet-a D7, and UNet). Moreover, BsiNet is a lightweight nework with fewer model parameters and faster implementation than the existing methods;
(2) The use of auxiliary tasks (i.e., boundary and distance maps) improves the geometric accuracy of extracted fields compared with single task UNet, which only uses a mask task. The additional attention module SGE further improves the extraction of small fields. BsiNet is effective for delineating farm fields with irregular shapes and small sizes;
(3) BsiNet can be used as a multi-task model for remote sensing information extraction, characterized by a flexible structure and is easy to expand. The results suggest that it offers a potential solution for agricultural applications needing delineating agricultural fields.

Acknowledgments

The research was funded by the National Natural Science Foundation of China (NSFC) with No. 42001283.

Supplementary document

A supplementary document is provided for this article. It contains four additional experiments and related analysis on (1) applying BsiNet to delineate agricultural fields in Xinjiang and Fujian areas using high-resolution images acquired from Google Earth; (2) transferability test on BsiNet, i.e., applying BsiNet that was trained on one dataset to delineate agricultural fields on another dataset; (3) the impacts of different input image sizes on delineation results; (4) applying BsiNet to delineate agricultural fields from medium-resolution Sentinel-2 images.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jag.2022.102871.

References

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote Sens. Environ. 204, 509–523.
Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, pp. 1–4.
Garcia-Pedrero, A., Lillo-Saavedra, M., Rodriguez-Esparragon, D., Gonzalo-Martin, C., 2019. Deep learning for automatic outlining agricultural parcels: Exploiting the land parcel identification system. IEEE access 7, 158223–158236.
Graesser, J., Ramankutty, N., 2017. Detection of cropland field parcels from landsat imagery. Remote Sens. Environ. 201, 165–180.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969.
Hossain, M.D., Chen, D., 2019. Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective. ISPRS journal of photogrammetry and remote sensing 150, 115–134.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141.
Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv:1801.05746.
Laben, C.A., Brower, B.V., Jan. 4 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. US Patent 6,011,875.
Li, M., Bijker, W., Stein, A., 2015. Use of binary partition tree and energy minimization for object-based classification of urban land cover. ISPRS journal of photogrammetry and remote sensing 102, 48–61.
Li, X., Hu, X., Yang, J., 2019. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv:1905.09646.
Masoud, K.M., Persello, C., Tolpekin, V.A., 2020. Delineation of agricultural field boundaries from sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks. Remote sensing 12 (1), 59.
McCarty, J., Neigh, C., Carroll, M., Wooten, M., 2017. Extracting smallholder cropped area in tigray, ethiopia with wall-to-wall sub-meter worldview and moderate resolution landsat 8 imagery. Remote Sens. Environ. 202, 142–151.
Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M., 2019a. Conv-mcd: A plug-and-play multi-task module for medical image segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 292–300.
Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M., 2019b. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 7223–7226.
Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999.
Paglieroni, D.W., 1992. Distance transforms: Properties and machine vision applications. In: CVGIP: Graphical models and image processing, 54, pp. 56–74.
Persello, C., Bruzzone, L., 2009. A novel protocol for accuracy assessment in classification of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48 (3), 1232–1244.
Persello, C., Tolpekin, V., Bergado, J., de By, R., 2019. Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sens. Environ. 231, 111253.
Ruder, S., 2017. An overview of multi-task learning in deep neural networks. arXiv: 1706.05098.
Ruiz-Santaquiteria, J., Bueno, G., Deniz, O., Vallez, N., Cristobal, G., 2020. Semantic versus instance segmentation in microscopic algae detection. Eng. Appl. Artif. Intell. 87, 103271.
Singh, V., Devgan, V., Anand, I., 2020. Determining image similarity with quasi-euclidean metric. arXiv:2006.14644.
Tan, C., Zhao, L., Yan, Z., Li, K., Metaxas, D., Zhan, Y., 2018. Deep multi-task and task-specific feature learning network for robust shape preserved organ segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 1221–1224.
Volpi, M., Tuia, D., 2018. Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. ISPRS J. Photogram. Remote Sens. 144, 48–60.
Wagner, M.P., Oppelt, N., 2020. Deep learning and adaptive graph-based growing contours for agricultural field extraction. Remote sensing 12 (12), 1990.
Waldner, F., Diakogiannis, F.I., 2020. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. Remote Sens. Environ. 245, 111741.
Waldner, F., Diakogiannis, F.I., Batchelor, K., Ciccotosto-Camp, M., Cooper-Williams, E., Herrmann, C., Mata, G., Toovey, A., 2021. Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images. Remote Sens. 13 (11), 2197.
Wassie, Y., Koeva, M., Bennett, R., Lemmen, C., 2018. A procedure for semi-automated cadastral boundary feature extraction from high-resolution satellite imagery. J. Spatial Sci. 63 (1), 75–92.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19.
Xia, L., Luo, J., Sun, Y., Yang, H., 2018. Deep extraction of cropland parcels from very high-resolution remotely sensed imagery. In: 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics). IEEE, pp. 1–5.
Yuan, Y., Xie, J., Chen, X., Wang, J., 2020. Segfix: Model-agnostic boundary refinement for segmentation. In: European Conference on Computer Vision. Springer, pp. 489–506.
Zhang, H., Liu, M., Wang, Y., Shang, J., Liu, X., Li, B., Song, A., Li, Q., 2021. Automated delineation of agricultural field boundaries from sentinel-2 images using recurrent residual u-net. Int. J. Appl. Earth Obs. Geoinf. 105, 102557.
Zhou, L., Zhang, C., Wu, M., 2018. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 182–186.