INTERNATIONAL JOURNAL OF REMOTE SENSING
2022, VOL. 43, NO. 1, 75–94
https://doi.org/10.1080/01431161.2021.2003467

Gradient convolutional neural network for classification of agricultural fields with contour levee

Abolfazl Meyarian[a,*], Xiaohui Yuan[a,*], Lu Liang[b], Wencheng Wang[c] and Lichuan Gu[d]

[a]Department of Computer Science and Engineering, University of North Texas, Denton, Texas, USA; [b]Department of Geography and the E nvironment, University of North Texas, Denton, Texas, USA; [c]College of Information and Control Engineering, Weifang University, Weifang, China; [d]School of Information and Computer, Anhui Agricultural University, Hefei, China

ABSTRACT

To better understand how contour-levee irrigation practice impacts water resources for formulating effective water management policies, it is important to obtain its application on large-scale data sets, e.g. state-wide. Automatic classification of contour levee croplands from high-resolution aerial images is of great potential given the success of deep neural networks and the availability of high-resolution remote sensing imagery. This paper proposes a gradient CNN model to classify fields with contour levees from remote sensing images. Our model produces high-quality segmentation masks that are refined with superpixel-based segmentation post-processing. Our method is evaluated using images by the National Agriculture Imagery Program (NAIP) for the counties in Arkansas. A comparison with the state-of-the-art methods demonstrates the improved performance of our proposed method. Our method demonstrates superior performance in the classification of challenging cases and achieves an overall 3.08% of accuracy improvement and 28.57% BER error reduction, compared to the second-best method. The p-value with respect to the second-best method is 0.005, which indicates great statistical significance. In addition, the results for data of different counties demonstrate the exceptional generalization ability of our method.

1. Introduction

Water used for irrigation accounts for nearly 65% of the world's freshwater withdrawals excluding thermoelectric power. The most conventional and dominant irrigation method is contour levees, where water flows by gravity from upper to lower paddies, and levees are used to maintain flood between them. To better understand how contour-levee irrigation practice impacts water resources for formulating effective management policies, it is critical to obtain its spatially explicit information. However, the existing irrigation maps only depict the irrigation status and are produced from images of coarse spatial resolution. Differentiating fields with contour levees from other types is a critical task to

CONTACT Xiaohui Yuan xiaohui.yuan@unt.edu University of North Texas, Denton, Texas, USA

*Authors have equal contributions.

understand the impact of agricultural practice on water resources and formulate effective policies for management. Automatic classification of contour levee croplands from high-resolution aerial images is of great potential given the success of deep neural networks and the availability of high-resolution remote sensing imagery.

Contour levees are usually depicted as curved lines in aerial images with uneven spaces in between. Given the importance of line features for differentiating contour levees, we use gradients as inputs for the network model. Image gradients are considered a sparse representation of an image that has non-zero values assigned to pixels where the intensity changes. In this sparse representation, colour features are suppressed from the gradient images. The model deals with a more distinct set of features, allowing it to focus only on the line features.

Figure 1 illustrates examples of fields with contour levees. There are many line features in the aerial images of croplands besides levees, such as field boundaries and tractor tracks. In addition, levees vary in spacing, visibility from the image, and texture of the background. A challenge is to deal with the great variety of gaps between levees that make correct detection of fields with contour levees a non-trivial task, as there is a high chance of classifying cropland as a non-contour levee class if the levees are not represented in the image due to the large gaps between them when the images are sub-sampled from high-resolution imagery. Therefore, providing enough context in an image for the model is essential. CNN models have been commonly used to perform the task of semantic segmentation by providing more contextual information through multi-scale feature extraction using convolution units (Chen et al. 2017; Qiao, Yuan, and Elhoseny 2020; Zhuang, Yuan, and Wang 2020).

Due to the similarity in colour and texture of croplands, it is possible that many small areas in an image are misclassified as contour levee. The boundaries produced by the model are not exactly a match to the actual boundaries. Hence, we used a boundary-guided post-processing method to refine the predictions, using the superpixel maps representing semantic units in the images. This method provides a chance to perform majority voting on the existing classes in prediction corresponding to each cropland and select the most dominant class in terms of frequency as the label for the whole cropland.

Our contributions of this paper include an encoder-decoder network equipped with a deep supervision mechanism that uses image gradients to classify croplands with contour levees. The network leverages a superpixel-aided image sampling as well as a post-processing technique. The sampling method maximizes the possibility of having only levees inside the images by selecting the samples from inside each cropland. The post-processing method uses the boundary information to perform majority voting in each cropland and improve the predictions.

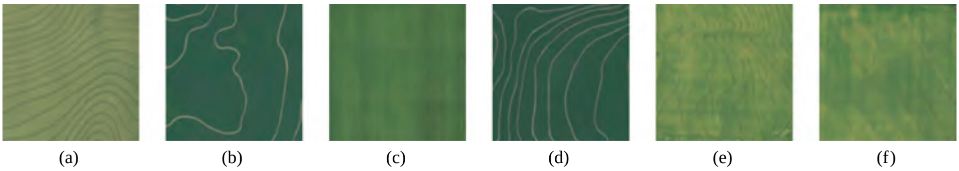

Figure 1. Fields with contour levees. (a) and (b) present samples of different levee spacing. (c) and (d) show weakly and strongly visible levees, respectively. (e) and (f) depicts a share of texture in croplands with different irrigation systems.

The organization of the rest of this paper is as follows: Section 2 reviews the related work for land segmentation and classification. Section 3 presents the proposed network for contour levee classification and post-processing using superpixels to refine the predictions. Section 4 discusses the experimental results of classifying high-resolution aerial images for fields with contour levees. Section 5 concludes this paper with a summary and future work.

2. Related work

There have been many studies on the classification of remote sensing images for crop type (Phalke et al. 2020; Shen, Liu, and Yuan 2017) and land cover classification (Fang et al. 2018; Qiao and Yuan 2021). Two commonly used strategies in these applications include pixel-based methods and region-based methods (Yuan, Shi, and Gu 2021). Pixel-based methods predict the class of each pixel at a time. Teluguntla et al. (2018) proposed a Landsat-derived cropland extent method for agricultural lands. The proposed technique uses Random Forest as the classifier and provides eight bands of pixel features for the model. Xie et al. (2019) worked on the semi-automatic classification of irrigated croplands provided in Landsat-based dataset (LANID-US 2012). A total of 112 features are given to a Random Forest model to perform the classification. Löw et al. (2013) investigated the effect of the feature space provided to a Support Vector Machine (SVM) on its performance on pixel-level crop mapping. The proposed method calculates the importance of different features based on the number of decrements in Out-Of-Bag (OOB) error for a Random Forest model in the classification of the crop types. Afterward, the highly informative features are chosen, and the corresponding subset of the dataset is given to the SVM. The pixel-based classification makes predictions for each pixel independent from the others in the neighbourhood. That is, such methods often fail to ensure spatial integrity.

With the recent advancements in deep learning, CNNs have been studied and applied to many computer vision tasks. Seferbekov et al. (2018) proposes Feature Pyramid Network (FPN) for multi-class land segmentation. The FPN extends ResNet50 by having two additional parallel pathways to extract features from the aerial images. The first stream down-samples the image and produces the deep features, while the second stream takes the feature map of each stage and applies 1×1 convolution to reduce the depth. The modified feature maps of each stage are passed to two convolution layers followed by upsampling to create feature maps of the same dimensionality as the input image. All feature maps of different stages are concatenated and used to generate the final prediction. (Mohammadimanesh et al. 2019) designed a model with an encoder and decoder network (referred to as FCN-M in this paper). The encoder uses a series of convolution, batch normalization, and ReLU activation along with inception and residual module to extract the features from images. The decoder also tries to retrieve the segmentation mask by means of conventional and transpose convolution. Multi-scale feature processing is not considered in the architecture of this network. Although region-based methods improve the consistency among local neighbourhoods, the results are often inaccurate near the boundary of the objects, which are in the forms of false positive or false negative.

Martins et al. (2020) proposed a Multi-OCNN composed of six networks for land cover classification. The method performs the classification task at the pixel level. For each pixel in the image patches of different scales are extracted. For each scale, an individual network is trained to provide a label for the centre pixel. This method can be regarded as a hybrid pixel-based/region-based model. The inference time and computational cost are dramatically high, as the model performs the classification for each pixel at a time, by collecting the predictions from all networks. Mboga et al. (2020) proposed a fully convolutional neural network that fuses feature maps of different levels to produce the final segmentation map. This mechanism for feature fusion helps the model to involve features at a low and high level in inferring the label of pixels. The transformer is another type of model that was designed for natural language processing and has recently been adapted for computer vision tasks. Swin-UNet Cao et al. (2021) is an UNet-like Transformer, designed for image segmentation. The image is divided into patches of size 4 by 4 that are fed to the encoder network, which performs a series of hierarchical Swin Transformer with a sliding window. Skip connections are used between the encoder and the decoder networks to capture the global and local relations between the patches and different scales of images. Despite the much-improved performance, the pixel-wise classification makes Swin-UNet facing the spatial integrity issue.

3. Gradient convolutional neural network

3.1. Network architecture

Colour images have been vastly used in many recent image segmentation and object detection (Yuan, Shi, and Gu 2021; Lu et al. 2019) methods. However, agricultural croplands share highly similar colours regardless of the irrigation system type. On the other hand, contour levees are often distinguished with irregular curved lines of uneven spacing. The presence of such patterns in cropland is a necessary condition for deciding a contour levee. Therefore, if a model is able to distinguish other types of lines such as boundary lines from levees, there is a potential to achieve high performance in the detection of contour levees.

Based on this observation, we designed a convolutional neural network that uses image gradients as the input to detect fields with contour levees from high-resolution images. Image gradient provides the model with key information about where intensity level changes drastically. Figure 2 shows the architecture of our proposed network with its two main components: an encoder network and a decoder network. The encoder network extracts the convolutional features. It is a CNN with eight convolutional layers. The second convolution layer of the first three stages has a stride of two, which achieves an effect of a down-scale ratio of eight. Hence, the encoder extracts image features at different scales. The decoder network generates the prediction for each pixel from the extracted features using a mask generator. The network leverages the skip connections to pass the feature maps from three levels of the encoder to their counterparts in the decoder. The fine-grain features from the corresponding layers in the encoder network are concatenated with the convolutional results from the coarse resolution. A deconvolution process is used to restore the spatial resolution.

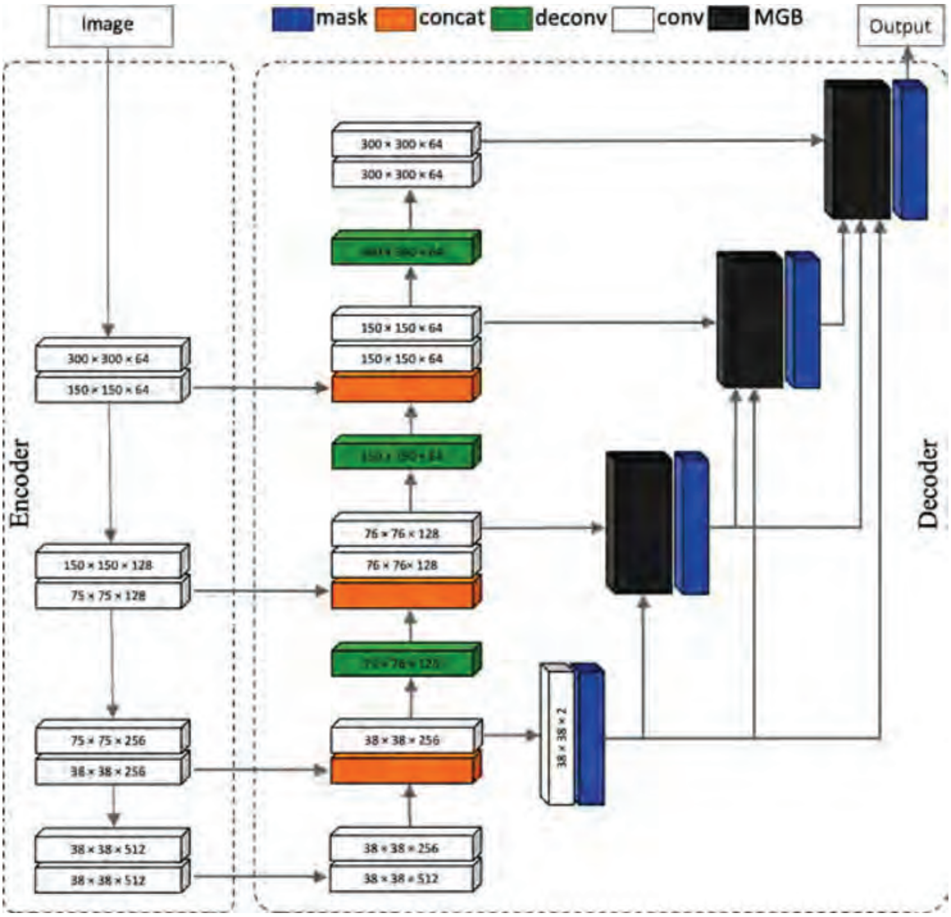

Figure 2. The architecture of the proposed gradient convolutional neural network.

The integration of the low-level features from the encoder alongside the high-level features of the decoder makes it possible to use the detailed features as well as coarse features. Hence, the network produces a high-resolution prediction map. In our implementation, as the size of feature maps is not always divisible by two, the feature of 75 × 75 × 128 in the encoder network is resized to 76 × 76 × 128 in the decoder network to ensure a compatible size with the deconvolution output. The resize is achieved with padding and in the next deconvolution process, the padded row/column is removed.

Following the idea of deep supervision (Xie and Tu 2015), we designed Mask Generator blocks (MGBs) to produce the segmentation masks in different stages of our network. The structure of an MGB is shown in Figure 3. At each stage, a segmentation mask is generated given the last feature map passed forwards from the last convolution layer of that stage. Starting from the second stage, the production of a mask is achieved in a two-step procedure inside the MGBs in each stage: generation of an initial mask, which is concatenated with the mask from the previous stage. The concatenated masks are fed into

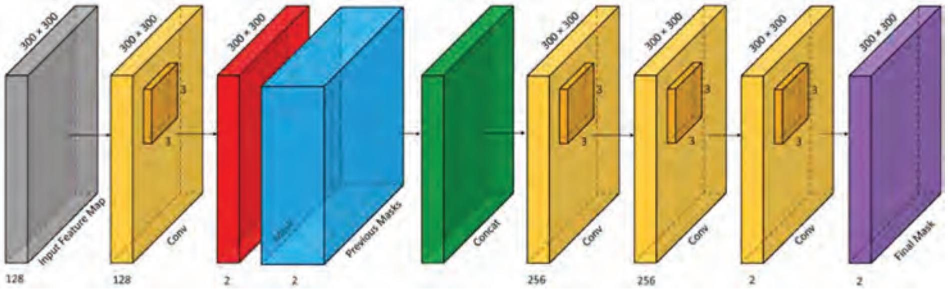

Figure 3. The mask generator block in the decoder.

the convolution layers. This operation is performed to improve the quality of the current mask based on the prior results. To ensure the consistent size of the masks in each stage with the ground truth, these masks are up-sampled.

3.2. Objective function

We formulate the cost function as a weighted sum of the individual cost function, comparing the generated prediction at different stages with the ground-truth masks, as follows:

$$\sum_{i=1}^{4} \lambda_i \Gamma(G, M_i) + L_r \tag{1}$$

where G is the ground-truth for a given image, M_i is the mask generated by i th stage of the decoder, λ_i is the weight associated with the mask for each level of prediction in the decoder and $\Gamma(\cdot)$ is a Softmax cross-entropy function for binary classification. Softmax calculates the probability of each pixel belonging to every class. $\Gamma(\cdot)$ is computed as follows:

$$\Gamma(G, M_i) = \frac{1}{N} \sum_{x,y} -(G(x,y) \log(M_i(x,y)) + (1 - G(x,y)) \log(1 - M_i(x,y)) \tag{2}$$

where N is the total number of pixels, $G(x,y)$ and $M_i(x,y)$ are the ground-truth label and the predicted probability for the pixel at (x,y), respectively. The value for the weight parameters λ_i is determined empirically by testing different cases of giving more weight to the last prediction, early prediction, and evenly distributed among all stages.

To regularize the network parameters, a weighted sum of l2-norm is adopted to the network weights as follows:

$$L_r = \lambda_R \sum_{i=1}^{N} w_i^2 \tag{3}$$

where λ_R is the regularization coefficient and w_i denotes single weight parameters of the network.

3.3. Crop field segmentation and decision fusion

Superpixels provide a perceptual representation of the entities inside an image, in which pixels are grouped in terms of spatial, colour, or texture attributes (Liu et al. 2018; Mi and Chen 2020; Yang et al. 2020). Superpixels are a key component in our model, as they are used in both pre- and post-processing phases. To prepare the training images, we perform image sampling inside the superpixels with a single class type. Ideally, a superpixel is one crop field. Moving the sampling window inside croplands guarantees that the borderlines and contents from a nearby field are not captured. Hence, levees are the only type of line features in the selected samples.

Segmentation of an image into superpixels is semantically imperfect. Most of the algorithms rely on colour and spatial similarities to decide which pixels can be grouped into a superpixel. Under these circumstances, these models may divide a semantic unit into smaller or elongated superpixels.

Superpixels representing the croplands with contour levee systems or other types of irrigation systems often have a low level of elongation and have a compact shape, which means the distribution of pixels in all directions is almost uniform. However, superpixels covering roads and boundary regions that separate large croplands have usually a prolonged shape.

To select the superpixels of crop fields, we use the number of pixels in a field and the elongation of the region to rank the superpixels with respect to their shape. The elongation E (Stojmenović and Žunić 2008) of a superpixel (region) I is computed as follows:

$$E(I) = \frac{\bar{m}_{2,0}(I) + \bar{m}_{0,2}(I) + \sqrt{4(\bar{m}_{1,1}(I))^2 + (\bar{m}_{2,0}(I) - \bar{m}_{0,2}(I))^2}}{\bar{m}_{2,0}(I) + \bar{m}_{0,2}(I) - \sqrt{4(\bar{m}_{1,1}(I))^2 + (\bar{m}_{2,0}(I) - \bar{m}_{0,2}(I))^2}} \tag{4}$$

where $\bar{m}_{p,q}(I)$ are centralized moments:

$$\bar{m}_{p,q}(I) = \sum_{x,y} I(x,y)(x-\bar{x})^p(y-\bar{y})^q \tag{5}$$

where x, y indicate the position of a pixel in the image and p, q are the order of the moment, $\bar{x}$ and $\bar{y}$ are the mass centres, $m_{p,q}$, as follows:

$$m_{p,q}(I) = \sum_{x,y} I(x,y)x^p y^q, \text{where} \bar{x} = \frac{m_{10}}{m_{00}} \text{and} \bar{y} = \frac{m_{01}}{m_{00}} \tag{6}$$

The score D of a superpixel I is the ratio of the number of pixels in I, denoted with $N(I)$ and its the elongation factor $E(I)$:

$$D(I) = \frac{N(I)}{E(I)}. \tag{7}$$

Our method classifies pixels of fields with contour levee. Ideally, all pixels of a crop field have the same class label. However, as the semantic information is missing, we often see a single field consisting of pixels classified into different classes, and a small group of pixels is put into one class that is different from the surroundings. The left panel

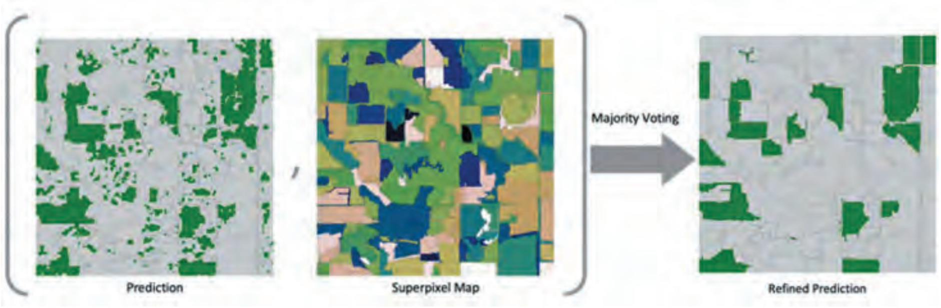

Figure 4. Classification refinement using superpixels.

(prediction) of Figure 4 illustrates the output of the classification model. Pixels of contour levee class are depicted in green, and the rest are in gray. There exist many small regions that are classified as contour levees, which are misclassifications. The middle panel (Superpixel map) illustrates the superpixels that capture the crop fields. Comparing the superpixels with the classification outputs, we see fields with a mixture of contour levee pixels and non-contour levee pixels. To rectify such errors, we leverage the boundary information of the superpixel map and perform majority voting of all pixels within each superpixel. The final classification is decided to the class with the greatest count.

4. Experimental results and discussion

4.1. Dataset and settings

The National Agriculture Imagery Program (NAIP) acquires imagery during the agricultural growing seasons in the continental U.S. Images are acquired at 1-metre resolution. The dataset used in our study is a subset of NAIP 2015 that includes 138 image tiles that cover Lonoke County in Arkansas. Each image tile has a size of $5,000$ by $5,000$. Among all samples, 16 tiles are annotated into two classes (contour levee and non-contour levee) and are used for training and testing. The spatial distribution of the tiles is shown in Figure 5.

Among all samples, only 23.75% of the pixels are labelled as contour levees and 76.25% of them are considered as background. The original images are rotated by $[5, 10, \ldots, 180]$ degrees as a way to augment the samples of all classes. More than 1.7 million samples are generated, and 88 thousand samples are randomly selected for training the network.

For each network, the initial parameters are chosen randomly from a normal distribution with $\mu = 0$ and $\sigma = 0.1$. The technique used for optimization of the network is Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We assigned equal weight to the output of each level for the deep supervision mechanism in our network, with values of $[\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1, \lambda_4 = 1]$. The learning rate of the networks is initialized with $\eta = 10^{-3}$. After 65,000 iterations, the learning rate is reduced to $\eta = 10^{-4}$. After 75,000 iterations, it is reduced again to $\eta = 10^{-5}$.

In our evaluation, we use precision, recall, accuracy, F1-Score, Mean-IoU, and Balanced Error Rate (BER) (Le et al. 2018; Wang, Li, and Yang 2018). The balanced error rate computes the average error of classification with respect to each class:

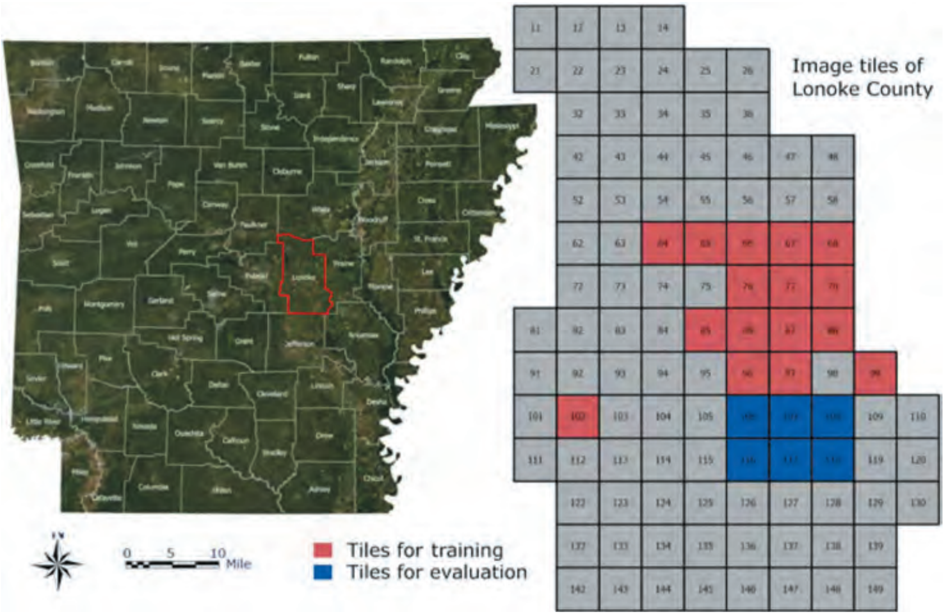

Figure 5. Arkansas state and the layout of images presenting the Lonoke county. The left sub-figure shows the boundaries of the Arkansas state and other countries. Lonoke county is shown with red boundaries. The right sub-figure shows the spatial location of all tiles with training tiles shown in red and 6 additional test tiles in blue color. The red tiles are also used for testing in each iteration of the k-fold cross-validation.

$$BER = 1 - \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{8}$$

where the first term in the parenthesis is known as *Sensitivity* and the second term is called *Specificity*. In all of the evaluations, a high value of accuracy and a low value for BER is desired.

4.2. Comparison of superpixel segmentation methods

A robust and accurate superpixel segmentation algorithm plays a crucial role in improving the performance of our method. Superpixel segmentation methods are generally categorized into graph-based and gradient-based classes. Superpixel segmentation methods such as *Felzenswalb and Huttenlocher*(Felzenszwalb and Huttenlocher 2004) model the problem using a graph-based similarity measurement technique. *Quick Shift*(Vedaldi and Soatto 2008) uses both color and position attributes to measure the similarity of pixels presented in a form of (X, Y, R, G, B) with an approximation of kernel-based mean-shift, where X, Y present the position of the pixel and R, G, B are the pixel colors. Supported by many recent works in object-based classification and segmentation of remote sensing data, (Gao et al. 2021; Abd Manaf et al. 2018), to have a fair comparison between segmentation methods of all categories, we choose to evaluate *SLIC*(Achanta et al. 2010), *Compact Watershed*(Neubert and Protzel 2014), *Quick Shift*(Vedaldi and Soatto

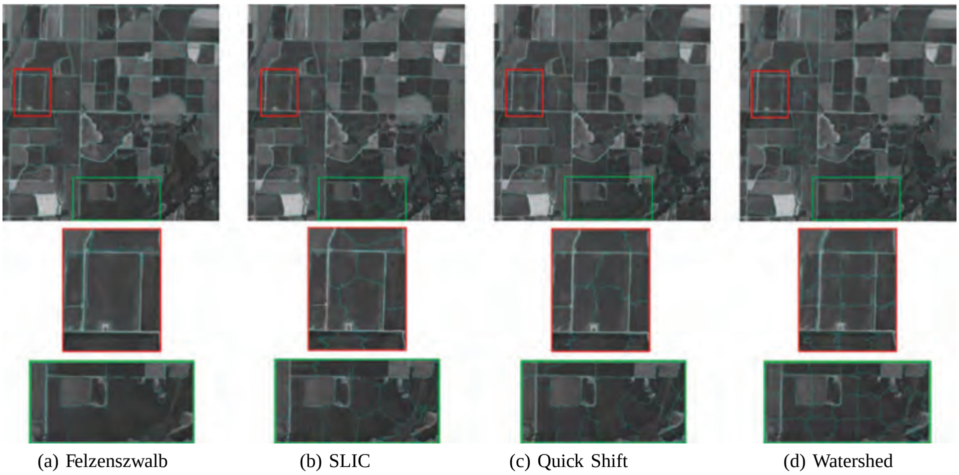
(a) Felzenszwalb (b) SLIC (c) Quick Shift (d) Watershed

Figure 6. A comparison of superpixel segmentation methods. The resulted segmentation boundaries are overlaid on the image in blue.

2008), and *Felzenswalb and Huttenlocher*(Felzenszwalb and Huttenlocher 2004) in identifying the boundaries of the croplands, among which, the first three methods are gradient-based and the last method is a graph-based technique.

The visualization of the segmentation using each method on a 5000 × 5000 pixel tile from Lonoke County is presented in Figure 6. In the figure, the detected boundaries are overlaid with blue colour on the NAIP image. Moreover, two sample regions are shown on the top of each test image for the models. In the first case, shown in the green for each method, the Felzenszwalb segmentation technique has identified the boundaries of the wood in a more accurate way, and the boundaries aligned with the semantic structure of the land, compared to the other three methods. In the second case, which belongs to agricultural cropland with a simple texture structure (shown in red), Compact Watershed (CW), Quick Shift, and SLIC have divided the land into several small regions, which can introduce inaccuracies if these segmentation boundaries are used for the post-processing step. This issue occurs as many small regions in the image can have different irrigation types in the prediction. The only potential way to resolve this issue is to perform the majority voting with the correct boundaries of the whole cropland not separately on the small partitions. In this case, also Felzenszwalb performed the best and was able to recognize the boundaries most accurately.

The effect of using the majority voting on three prediction tiles, each from Lonoke, Woodruff, and Arkansas counties with and without the four selected methods is provided in Table 1. The average value of each metric on all test samples is reported with the standard deviation in the parentheses. The best results are highlighted in boldface font and the second-best results are underlined. Using the Felzenszwalb method improves the overall accuracy with a 2.13% margin compared to the second-best method, as well as 6.92% compared to the case of using the raw predictions. An improvement of 14.81% in precision over the cases without using the superpixels demonstrates the effectiveness of majority voting in restoring the misclassified false-negative predictions.

Table 1. Classification performance with and without superpixel for post-processing.

Model	Accuracy	BER	Precision	Recall	F1-Score	M-IoU
w/o sup. pix.	0.86 (0.04)	0.16 (0.03)	0.81 (0.04)	0.80 (0.05)	0.80 (0.02)	0.73 (0.04)
CW	0.88 (0.04)	0.13 (0.02)	0.86 (0.05)	0.81 (0.01)	0.84 (0.02)	0.77 (0.06)
SLIC	0.89 (0.04)	0.12 (0.03)	0.88 (0.03)	0.84 (0.04)	0.79 (0.06)	**0.86** (0.02)
Quick Shift	0.90 (0.05)	0.11 (0.04)	0.86 (0.04)	**0.87** (0.02)	0.86 (0.03)	0.80 (0.08)
Felzenszwalb	**0.92** (0.03)	**0.10** (0.02)	**0.93** (0.01)	0.84 (0.04)	**0.88** (0.02)	0.83 (0.05)

4.3. Superpixel generation

Using superpixel help improves the integrity of the classification of crop fields. We compute the scores for the superpixels following Eq. (7). According to our score, small scores are often indicative of small or extremely elongated regions. Examples of such superpixels include roads between fields. These superpixels should be eliminated. A crop field often appears as a rectangle region that is close to square in shape with a reasonable size. Hence, an appropriate threshold should follow these observations.

Figure 7 shows the heatmap of superpixel scores of a sample image. We can observe that most low-score samples are the connecting roads between the croplands. The superpixels with median-range scores are the ones with a fairly large number of pixels and a small elongation rate. The difference between superpixels with large and median-range scores is the number of pixels. The elongation factor among them is very similar.

To decide the threshold, we create the histogram of the superpixel score from all training images as shown in Figure 8.

This histogram illustrates the distribution of the scores that range from zero up to over 250,000. The majority of the superpixels are in the lower end of the distribution. The box plot underneath the histogram in Figure 8 shows the outliers and quartiles of the distribution. The red bars in the histogram plot highlight the first quartile, i.e. 25th percentile, and the median. Due to the highly skewed distribution, the average score is in between the first quartile and median. Based on our empirical analysis, we use the second quartile as the range for selecting the thresholds, i.e. [4,000, 20,000]. Note that the lower and upper bound of this range is round to thousands. In our study, to be more inclusive yet avoid unlikely crop fields, our threshold is set to 4,000.

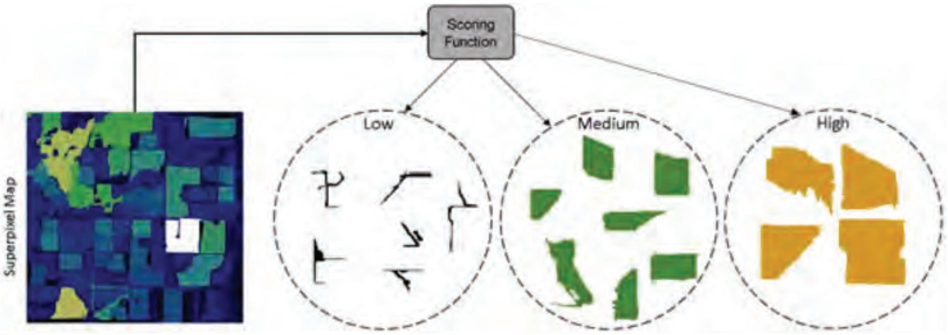

Figure 7. Geometric superpixel score of a superpixel map. A sample image colored according to the scores and groups of superpixels with low, medium, and high scores.

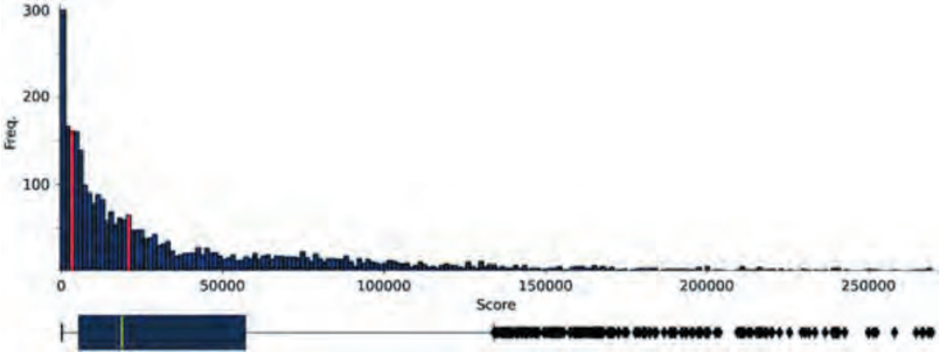

Figure 8. The histogram of superpixel scores. The red bars indicate the range used to choose the threshold for selecting superpixels.

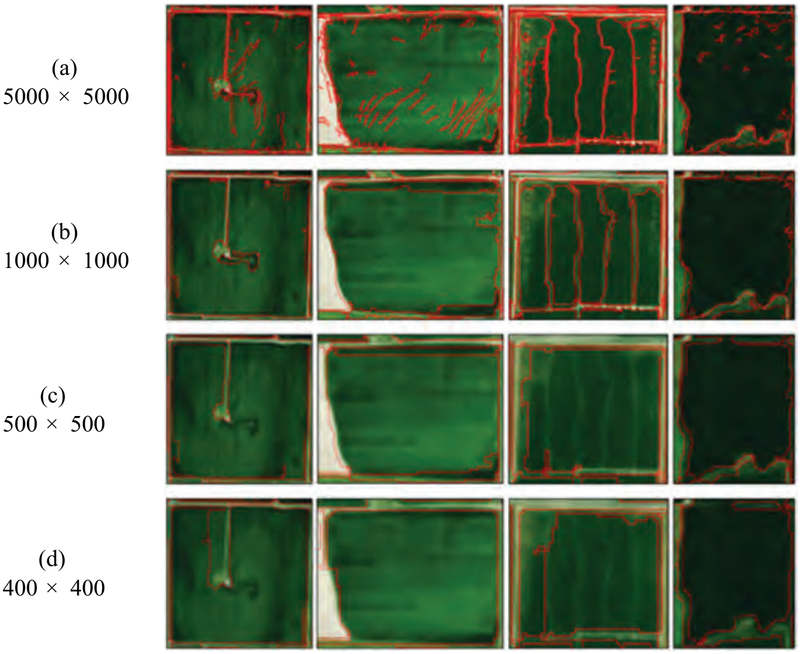

Figure 9. Results of Felzenswalb on images of different resolutions. Each column shows a different case.

Figure 9 depicts the result of the Felzenswalb method using images of different resolutions as input. The boundary of superpixels is highlighted in red and is overlaid on the image.

When a high-resolution image is used as input, the method could result in overly fine superpixels. For example, Figure 9(a) and (b) show the results of images of 5000 × 5000 and 1000 × 1000, respectively, and each contains many small segmentations that split one crop field into small pieces. Hence, such overly fine superpixels are inappropriate for crop field classification. A low-resolution input (e.g. Figure 9(c) and (d)) gives plausible

Table 2. Accuracy (%) with respect to the image down-scale size for gradient extraction.

Class	Contour Levee			Non-Contour Levee			Overall		
Size	400	500	1000	400	500	1000	400	500	1000
Accuracy	85.32	85.79	78.71	94.68	94.59	91.41	93.3	93.3	89.52
STD	11.11	10.45	10.19	4.48	4.14	5.32	2.56	2.62	2.83

results. The superpixels are mostly aligned with the boundary of the fields. However, as we reduce the resolution, staircase artifacts are visible and under segmentation appears more often.

In addition, we evaluate the impact of superpixels on classification accuracy. Table 2 presents the average classification accuracy using superpixels generated from different resized images of sizes 400, 500, and 1000. It is clear that when we use superpixels generated from images of resolution 1000 × 1000, the accuracy dropped. The accuracy of using the size of 400 and 500 is very similar. The overall accuracy of each case is at 93.3% with slightly better accuracy for contour levees at the resolution of 500. This is consistent with our observation of superpixel segmentation illustrated in Figure 9. Considering the overall accuracy and the focus of classifying contour levees, we resized images to 500 × 500 to produce the superpixels in the rest of our experiments.

4.4. Performance analysis and comparison study

We compare our proposed Gradient CNN with a number of classical and state-of-the-art methods including Random Forest (Teluguntla et al. 2018; Xie et al. 2019), FCN-ATR-SKIP, FCN-M (Mohammadimanesh et al. 2019), Feature Pyramid Network and Swin-UNet (Cao et al. 2021). These methods serve as the baseline in our comparison. Among the compared methods, we also applied our method to the RGB image that is denoted as RGB-Network. In this implication, the only difference between our proposed method and the RGB-Network is the inputs. The RGB-Network takes the raw image of the three colour bands. The model relies on mostly the colour features instead of gradient features. Using images from Lonoke County, 16-fold cross-validation was conducted. Table 3 presents the average performance of the compared methods. The numbers in the parentheses are the standard deviation. For each metrics, the best performance is highlighted with the boldface, and the second-best is marked with an underscore. Large values of accuracy, precision, and recall indicate better performance, while a small value of BER is desired.

Among all the models of FCN-ATR-SKIP we trained, only two were successfully completed, the results of which are in the table for the comparisons. In terms of all metrics, our proposed Gradient CNN exhibited the best performance. Our method

Table 3. A performance comparison.

Model	Accuracy	BER	Precision	Recall	F1-Score	M-IoU
FCN-ATR-SKIP	0.74 (0.08)	0.26 (0.03)	0.71 (0.18)	0.58 (0.08)	0.62 (0.04)	0.55 (0.09)
Random Forest	0.76 (0.05)	0.24 (0.04)	0.48 (0.15)	0.74 (0.06)	0.57 (0.12)	0.56 (0.07)
FCN-M	0.79 (0.05)	0.23 (0.05)	0.51 (0.15)	0.73 (0.13)	0.59 (0.12)	0.58 (0.06)
Swin-UNet	0.80 (0.06)	0.30 (0.06)	0.57 (0.15)	0.50 (0.16)	0.52 (0.13)	0.56 (0.05)
RGB-Network	0.89 (0.05)	0.18 (0.04)	0.76 (0.12)	0.69 (0.10)	0.73 (0.12)	0.71 (0.07)
Feature Pyra. Net	0.91 (0.03)	0.14 (0.04)	0.77 (0.09)	0.79 (0.11)	0.78 (0.07)	0.76 (0.04)
Gradient CNN	**0.93** (0.03)	**0.10** (0.04)	**0.83** (0.08)	**0.85** (0.10)	**0.85** (0.08)	**0.82** (0.06)

achieved a 3.08% improvement in terms of accuracy and a 28.57% error reduction in terms of BER compared to the second-best method. The high accuracy of our model for all classes is consistent with the low rate of BER. Moreover, an improvement of 7.7%, 7.5%, 8.9%, and 7.9% in terms of precision, recall, F1-score, and Mean-IoU, respectively, compared to the second-best confirms the superior performance of the Gradient CNN in classifying fields with contour levees. The standard deviation of our method is in the lower range, which demonstrates the consistency of the Gradient CNN.

We conducted a one-way ANOVA analysis of the results of our method and Feature Pyramid Network (the method exhibited the second-best performance). The test statistic is the F value of 8.97. Using an α of 0.05, we have $F_{0.05} = 4.171$. Since the test statistic is much larger than the critical value, we reject the null hypothesis of equal population means of these two methods and conclude that there is a statistically significant difference. The p-value for 8.97 is 0.005, so the test statistic is significant. Hence, the improvement by our method is statistically significant.

Figure 10 depicts the classification results of a sample image. Figure 10(a) shows the input image with the ground-truth superimposed as a semi-transparent layer. The crop fields with contour levees are in the green shade. Four fields are highlighted in coloured bounding polygons. A zoom-in view of the highlighted fields is shown in Figure 10(b). The fields with contour levees are depicted in green.

Figure 10(c) shows the result of our method, which is very close to the ground truth with two false-positive fields close to the right side of the image. The result of Feature Pyramid Network shown in Figure 10(d) also exhibits competitive performance. However, many false negatives exist in the resulted field classification. In addition, it failed to keep

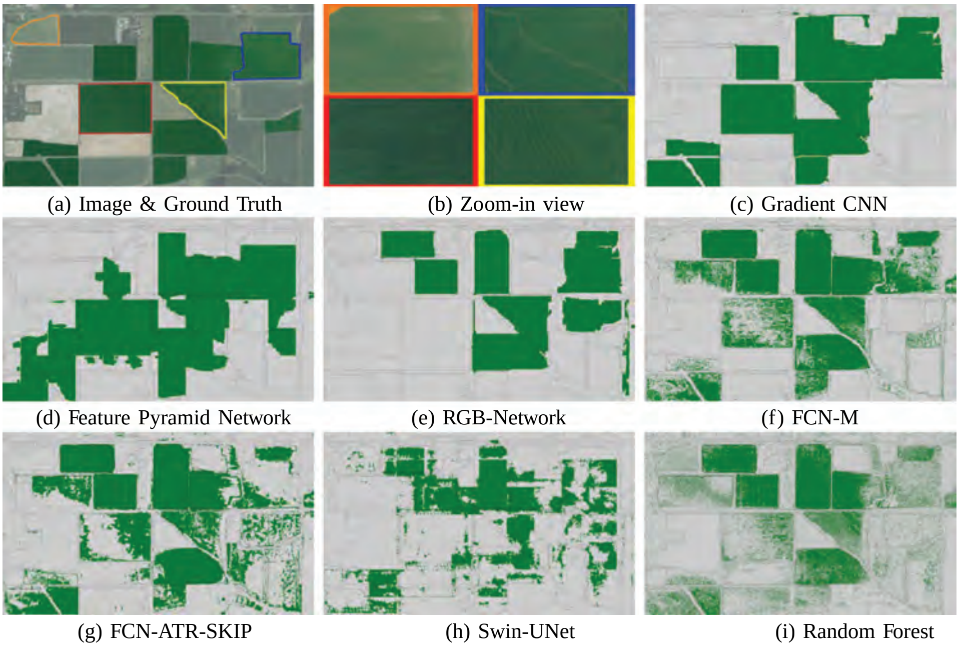
(a) Image & Ground Truth (b) Zoom-in view (c) Gradient CNN
(d) Feature Pyramid Network (e) RGB-Network (f) FCN-M
(g) FCN-ATR-SKIP (h) Swin-UNet (i) Random Forest

Figure 10. Classification results of the compared methods.

the roads out of the classification, which heavily contributes to the false positives of its results. The network using RGB as inputs, as shown in Figure 10 (e), failed to detect several large crop fields with contour levees, e.g. the field in the red bounding box, and resulted in a sizable false negative as well. Despite the variations of classification performance among these three methods, they maintain the integrity of the crop fields. That is, the pixels within one crop field share the same class label. Our method demonstrates a meticulous classification of contour levees that retains the integrity of the croplands.

It is evident that the performance of FCN-M, FCN-ATR-SKIP, Swin-UNet, and Random Forest, as shown in Figure 10(f)-(i), is inferior to Gradient CNN. A more severe issue is the lack of field integrity. The predictions of these methods are mostly independent of the context. There are many fields that consist of pixels classified into contradicting classes.

Figure 10(a) highlighted four exemplar cases using coloured polygons and a zoom-in view of part of the fields is shown in Figure 10(b). The field enclosed by orange polygon contains no contour levees and the other three fields are representative ones with uneven large gaps (blue), inverted levee colour (yellow), and vague levees (red). Almost all methods correctly classified the field in orange. Both Swin-UNet and Random Forest, however, resulted in spotty misclassification in that field. The fields in red and blue are most confusing and only Gradient CNN and Feature Pyramid Network reached correct classification. Even if some pixels were correctly classified by the other methods, it is impossible to reach a correct label using majority voting. The classification of the field in yellow bounding polygon is mostly satisfactory by all methods. For FCN-M, FCN-ATR-SKIP, Swin-UNet, and Random Forest, most of the pixels in the field are labelled correctly.

Figure 11 depicts the classification results of a large region in Lonoke County using our Gradient CNN. The detected fields are superimposed to the original image. The fields without contour levees are coloured in a blue shade. This illustration consists of six tiles of the Lonoke County highlighted in blue in Figure 5. The classified crop fields with contour levees are intact with satisfactory accuracy.

4.5. Model generalization

One of the most important aspects of performance evaluation in the classification of contour levee is assessing the generalization of a model on sample images different from the training set. The dataset used in this study is collected from Lonoke County. The dataset used to train all the models covers only a small section of Lonoke County with 16 tiles out of 138 tiles. In addition to Lonoke County, there are many other counties in Arkansas state in which the use of contour levees is common. To evaluate the generalization ability of our proposed method in comparison with state-of-the-art ones, we extend our evaluation using image tiles from the other sample images of Lonoke County as well as sample images from Arkansas County and Woodruff County.

The additional image tiles of Lonoke County, as well as the image tiles of the other two counties: Woodruff and Arkansas, were not used in the K-fold cross-validation training. The land coverage in these two counties differs from that of Lonoke County. Six image tiles were randomly selected from each county. Figure 12 illustrates the average performance of our model and the state-of-the-art methods together with the standard deviation (the coloured bars show the average value of the corresponding metric and the error bars show the magnitude of the standard deviation).

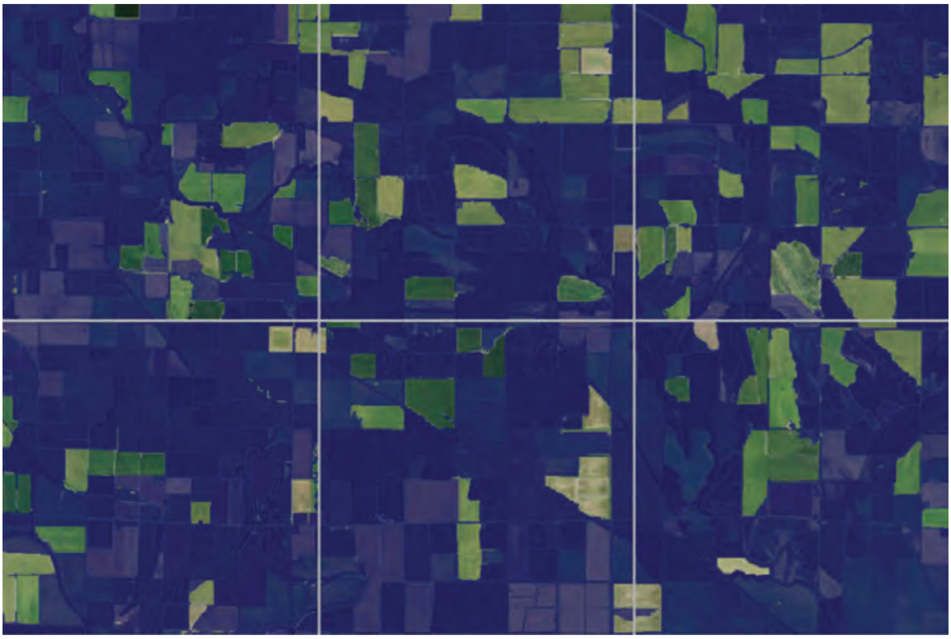

Figure 11. Illustration of segmentation results for six tiles of the Lonoke county. Fields without contour levees are colored in blue.

By comparing the results for Lonoke County with the results presented in Table 3, we see a highly similar trend for all performance metrics. Almost for all study areas, our proposed method remains mostly stable. The stable performance of our method can be attributed to the network structure and the employment of image gradients. The variation of image gradients is much less compared to that of the colour and texture features among image tiles from different counties. Hence, our method achieves a very competitive performance when it is evaluated with images of Woodruff and Arkansas Counties. It consistently exhibits the best performance in almost all cases.

However, when the models of the compared methods were applied to the image tiles of the other two counties, the performance varies. The performance of the Feature Pyramid Network degrades when it was applied to a different set of data. FCN-ATR-SKIP also exhibits an unstable performance with a slight drop of performance for samples of Woodruff County. Among all models, Random Forest shows the highest level of instability when it is applied to the additional samples of Lonoke, which is confirmed by the error bars in the sub-figure for Lonoke County.

4.6. Computational cost

Table 4 presents FLOPS (floating-point operations per second) and the number of parameters of the compared methods. FLOPS of a specific model determines how costly and time-consuming it is for that model to produce the segmentation mask given an input image. Therefore, smaller values are desired. On the other hand, the number of parameters helps to measure how much the complexity of a model is as well as a rough

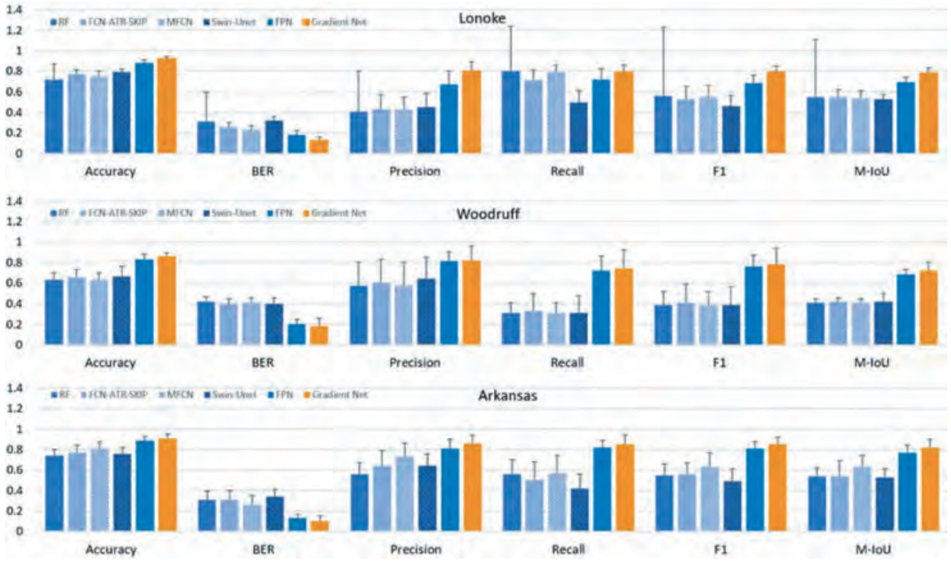

Figure 12. Performance comparison using images from three counties: Lonoke, Woodruff, and Arkansas.

estimation of the memory usage. Hence, smaller values indicate a simpler model with a lower level of memory consumption. The best performing method for each of the metrics is shown in boldface font and the second-best method is underlined.

FCN-ATR-SKIP exhibits the least FLOPS and number of parameters among all methods, whereas Swin-UNet has the greatest FLOPs due to the use of Multi-Layer Perceptron. Additionally, Feature Pyramid Network has the second-highest FLOPs and number, and the greatest number of parameters, this is partly because Feature Pyramid Network uses the ResNet50 backbone in the structure. Our proposed network is capable of producing the segmentation maps with 94.51% and 95.71% improvement in terms of FLOPs over Feature Pyramid Network and Swin-UNet, respectively. In our proposed model, we used 62.61% parameters less than Feature Pyramid Network while keeping the FLOPs close to a lightweight model such as FCN-ATR-SKIP with only six layers. The accuracy and efficiency of Gradient CNN become an important feature as it helps to provide the raw segmentation maps for the post-processing faster and reduce the overall process time to obtain a refined segmentation.

Table 4. FLOPS and number of parameters of the compared methods.

Model	FLOPS (Giga)	# of Para. (Million)
Random Forest	–	–
FCN-ATR-SKIP	**0.38**	**0.19**
FCN-M	1.23	5.05
Swin-UNet	17.28	26.60
Feature Pyramid Net	13.49	28.65
Gradient CNN	0.74	10.71

5. Conclusion

Detection of fields with contour levee irrigation system is a key component in providing an opportunity to preserve the water resources and to prevent the change of water regime in agricultural areas. Differentiation of contour levee from other irrigation practices is a challenging task due to the similarities in terms of texture and specific line features representing different practices. Given the importance of the line features and contextual information, we develop a region-based network to classify contour levees from other types of irrigation practices based on image gradients. We used a deep supervision mechanism to improve the quality of the predictions and to prevent gradient-vanishing. Having the boundaries of croplands generated by Felzenszwalb superpixel segmentation, we proposed a boundary-based majority voting to improve the quality of our results.

Our experimental results demonstrate that the proposed method achieved an overall 3.08% of accuracy improvement and 28.57% BER error reduction rate, compared to the second-best method. More importantly, our method demonstrates superior performance in the classification of challenging cases such as croplands with a low level of levee visibility, varying space between levees, and fields with high contrast between levees and the background. Our results from data of Woodruff and Arkansas counties demonstrate the improved performance of our proposed method with a margin of 5.16%, 1.17%, and 0.02% for samples of Lonoke, Woodruff, and Arkansas counties, respectively. From the aspect of computational expense, Gradient CNN performs the classification with high accuracy while maintaining the computational complexities at a very low level, compared to methods such as Feature Pyramid Network. Integration of the post-processing module with the network is a potential point of improvement that we consider in our next work. This enables us to train the model in an end-to-end manner and reduce the time needed for majority voting.

In our experiments, we observe confusion among contour levees and some straight levees. The vague line features of the straight levees introduce false uneven spaces between levees. In our future work, we plan to integrate the metric of straightness to assist the differentiation of contour levees.

Acknowledgements

We appreciate the comments and assistance by George Mihaila and Jawad Ahmed in the preparation of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID

Xiaohui Yuan http://orcid.org/0000-0001-6897-4563
Lu Liang http://orcid.org/0000-0002-9892-8346

References

Abd Manaf, S., Mustapha, N., Sulaiman, M.N., Husin, N.A., Shafri, H.Z.M. and Razali, M.N. 2018. "Hybridization of SLIC and extra tree for object based image analysis in extracting shoreline from medium resolution satellite images."

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S. 2010. *SLIC superpixels*. Technical report, Ecole Polytechnique Fedrale de Lausanne.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M. 2021. "Swin-Unet: Unet-like pure transformer for medical image segmentation." *arXiv Preprint* arXiv:2105.05537.

Chen, L. C., Papandreou, G., Schroff, F. and Adam, H. 2017. "Rethinking atrous convolution for semantic image segmentation." *arXiv Preprint* arXiv:1706.05587.

Fang, F., X. Yuan, L. Wang, Y. Liu, Z. Luo. 2018. "Urban land-use classification from photographs." *IEEE Geoscience and Remote Sensing Letters* 15 (12): 1927–1931. DOI:10.1109/LGRS.2018.2864282.

Felzenszwalb, P. F., and D. P. Huttenlocher. 2004. "Efficient graph-based image segmentation." *International Journal of Computer Vision* 59 (2): 167–181. doi:10.1023/B:VISI.0000022288.19776.77.

Gao, H., J. Guo, P. Guo, X. Chen. 2021. "Classification of very-high-spatial-resolution aerial images based on multiscale features with limited semantic information." *Remote Sensing* 13 (3): 364. DOI:10.3390/rs13030364.

Le, H., Vicente, T.F.Y., Nguyen, V., Hoai, M., and Samaras, D. 2018. "A + D net: Training a shadow detector with adversarial shadow attenuation." *In*: *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich, Germany, 662–678.

Liu, X., S. Guo, B. Yang, S. Ma, H. Zhang, J. Li, C. Sun, *et al*. 2018. "Automatic organ segmentation for CT scans based on super-pixel and convolutional neural networks." *Journal of Digital Imaging* 31 (5): 748–760. DOI:10.1007/s10278-018-0052-4.

Löw, F., Michel, U., Dech, S. and Conrad, C. 2013. "Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines". *ISPRS Journal of Photogrammetry and Remote Sensing* 85: 102–119. doi:10.1016/j.isprsjprs.2013.08.007.

Lu, Q., Liu, Y., Huang, J., Yuan, X. and Hu, Q. 2019. "License plate detection and recognition using hierarchical feature layers from CNN." *Multimedia Tools and Applications* 78 (11): 15665–15680. DOI:10.1007/s11042-018-6889-1.

Martins, V. S., Kaleita, A.L., Gelder, B.K., da Silveira, H.L. and Abe, C.A. 2020. "Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution". *ISPRS Journal of Photogrammetry and Remote Sensing* 168: 56–73. doi:10.1016/j.isprsjprs.2020.08.004.

Mboga, N., T. Grippa, S. Georganos, S. Vanhuysse, B. Smets, O. Dewitte, E. Wolff, *et al*. 2020. "Fully convolutional networks for land cover classification from historical panchromatic aerial photographs". *ISPRS Journal of Photogrammetry and Remote Sensing* 167: 385–395. doi:10.1016/j.isprsjprs.2020.07.005.

Mi, L., and Z. Chen. 2020. "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation." *ISPRS Journal of Photogrammetry and Remote Sensing* 159: 140–152. doi:10.1016/j.isprsjprs.2019.11.006.

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Gill, E. and Molinier, M. 2019. "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem". *ISPRS Journal of Photogrammetry and Remote Sensing* 151: 223–236. doi:10.1016/j.isprsjprs.2019.03.015.

Neubert, P., and P. Protzel, 2014. "Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms." *In*: *2014 22nd International Conference on Pattern Recognition*, 996–1001.

Phalke, A. R., M. Özdoğan, P. S. Thenkabail, T. Erickson, N. Gorelick, K. Yadav, R. G. Congalton, *et al*. 2020. "Mapping croplands of Europe, Middle East, Russia, and Central Asia using landsat, random forest, and google earth engine". *ISPRS Journal of Photogrammetry and Remote Sensing* 167: 104–122. doi:10.1016/j.isprsjprs.2020.06.022.

Qiao, Z., and X. Yuan. 2021. "Urban land-use analysis using proximate sensing imagery: A survey." *International Journal of Geographical Information Science* 35 (11): 2129–2148. doi:10.1080/13658816.2021.1919682.

Qiao, Z., X. Yuan, and M. Elhoseny, 2020. "Urban scene recognition via deep network integration." *In*: *International Conference on Urban Intelligence and Applications*, August 14-16, Taiyuan, China, 135–149.

Seferbekov, S., Iglovikov, V., Buslaev, A. and Shvets, A. 2018. "Feature pyramid network for multi-class land segmentation." *In*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 272–275.

Shen, Y., X. Liu, and X. Yuan. 2017. "Fractal dimension of irregular region of interest application to corn phenology characterization." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10 (4): 1402–1412. doi:10.1109/JSTARS.2016.2645880.

Stojmenović, M., and J. Žunić. 2008. "Measuring elongation from shape boundary." *Journal of Mathematical Imaging and Vision* 30 (1): 73–85. doi:10.1007/s10851-007-0039-0.

Teluguntla, P., Thenkabail, P.S., Oliphant, A., Xiong, J., Gumma, M.K., Congalton, R.G., Yadav, K., *et al*. 2018. "A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on google earth engine cloud computing platform". *ISPRS Journal of Photogrammetry and Remote Sensing* 144: 325–340. doi:10.1016/j.isprsjprs.2018.07.017.

Vedaldi, A., and S. Soatto, 2008. "Quick shift and kernel methods for mode seeking." *In*: *European conference on computer vision*. Marseille, France, 705–718.

Wang, J., X. Li, and J. Yang, 2018. "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal." *In*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, Utah, 1788–1797.

Xie, S., and Z. Tu, 2015. "Holistically-nested edge detection." *In*: *Proceedings of the IEEE international conference on computer vision*. Santiago, Chile, 1395–1403.

Xie, Y., Lark, T.J., Brown, J.F. and Gibbs, H.K. 2019. "Mapping irrigated cropland extent across the conterminous United States at 30 m resolution using a semi-automatic training approach on google earth engine". *ISPRS Journal of Photogrammetry and Remote Sensing* 155: 136–149. doi:10.1016/j.isprsjprs.2019.07.005.

Yang, S., Yuan, X., Liu, X. and Chen, Q. 2020. "Superpixel generation for polarimetric SAR using hierarchical energy maximization". *Computers & Geosciences* 135: 104395. doi:10.1016/j.cageo.2019.104395.

Yuan, X., J. Shi, and L. Gu. 2021. "A review of deep learning methods for semantic segmentation of remote sensing imagery." *Expert Systems with Applications* 169: 114417. doi:10.1016/j.eswa.2020.114417.

Zhuang, C., X. Yuan, and W. Wang, 2020. "Boundary enhanced network for improved semantic segmentation." *In*: *International Conference on Urban Intelligence and Applications*. Taiyuan, China, 172–184.