Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Residential roof condition assessment system using deep learning

Fan Wang
John P. Kerekes
Zhuoyi Xu
Yandong Wang

SPIE.

Fan Wang, John P. Kerekes, Zhuoyi Xu, Yandong Wang, “Residential roof condition assessment system using deep learning,” *J. Appl. Remote Sens.* **12**(1), 016040 (2018), doi: 10.1117/1.JRS.12.016040.

Residential roof condition assessment system using deep learning

Fan Wang,[a],* John P. Kerekes,[a] Zhuoyi Xu,[b] and Yandong Wang[c]
[a]Rochester Institute of Technology, Chester F. Carlson Center for Imaging Science,
Rochester, New York, United States
[b]Independent Researcher, Haidian District, Beijing, China
[c]EagleView Technologies, Rochester, New York, United States

Abstract. The emergence of high resolution (HR) and ultra high resolution (UHR) airborne remote sensing imagery is enabling humans to move beyond traditional land cover analysis applications to the detailed characterization of surface objects. A residential roof condition assessment method using techniques from deep learning is presented. The proposed method operates on individual roofs and divides the task into two stages: (1) roof segmentation, followed by (2) condition classification of the segmented roof regions. As the first step in this process, a self-tuning method is proposed to segment the images into small homogeneous areas. The segmentation is initialized with simple linear iterative clustering followed by deep learned feature extraction and region merging, with the optimal result selected by an unsupervised index, Q. After the segmentation, a pretrained residual network is fine-tuned on the augmented roof segments using a proposed k-pixel extension technique for classification. The effectiveness of the proposed algorithm was demonstrated on both HR and UHR imagery collected by EagleView over different study sites. The proposed algorithm has yielded promising results and has outperformed traditional machine learning methods using hand-crafted features. [DOI: 10.1117/1.JRS.12.016040]

Keywords: roof condition assessment; deep learning; segmentation; classification.

Paper 170749 received Aug. 23, 2017; accepted for publication Mar. 5, 2018; published online Mar. 21, 2018.

1 Introduction

The emergence of high resolution (HR) and ultra high resolution (UHR) airborne remote sensing imagery is enabling humans to move beyond traditional land cover analysis applications to the detailed characterization of surface objects. In particular, as an example, the inspection of roof condition is an important step in damage claim processing in the insurance industry and represents an analysis problem that would significantly benefit from automated processing of aerial imagery. Currently, roof inspections are performed by humans and are an expensive, time-consuming, and unsafe process. Thus, this topic of automated roof condition assessment from remotely sensed images is the topic explored in this work.

Relatively little research has been published specifically on the topic of roof condition assessment, particularly for recognizing relatively minor damage, e.g., can occur from a hailstorm. The closest area in the remote sensing analysis literature is the general problem of building damage assessment. According to the European Macroseismic Scale 1998 (EMS98),[1] building damage is classified into one of five damage grades: slight damage, moderate damage, heavy damage, very heavy damage, and destruction. There are several studies that have concentrated on building damage detection, which is aimed at the location of heavily damaged buildings in imagery. Sirmacek and Unsalan[2] proposed a damage measure derived from rooftop and shadow detection to discriminate between damaged and undamaged buildings. Brunner et al.[3] proposed a damage detection algorithm that detects buildings destroyed in an earthquake

*Address all correspondence to: Fan Wang, E-mail: wangfanhit@gmail.com using pre-event very high resolution (VHR) optical and postevent detected VHR synthetic aperture radar imagery.

Compared to traditional building damage assessment, the assessment of the condition of residential roofs having relatively minor damage is a more sophisticated task. Bignami et al.[4] studied the sensitivity of textural features with respect to damage levels using pre- and postevent QuickBird data. Gerke and Kerle[5] proposed a supervised classification algorithm to discriminate intact roofs from destroyed roofs and generated a damage score using oblique Pictometry data. An accuracy of 63% for building damage assessment was achieved.[5] Samsudin et al.[6] proposed spectral indices to generate degradation status maps of concrete and metal roofing materials using multispectral imagery. In general, according to the review of Dong and Shan,[7] the heavy damage grades, such as destruction in EMS98, can be addressed by traditional techniques available today. However, the dentification of lower damage grades is still a barrier, even with a submeter resolution data.[7] It is the identification of these lower grades of damage that serves as the objective of this work.

Recently, deep learning (DL) techniques have demonstrated excellent performance on various image analysis tasks and have drawn increasing attention in the remote sensing community. One of the most popular DL techniques is the convolutional neural network (CNN), which, among many applications, has been employed to classify hyperspectral data using spectral information with experimental results demonstrating better performance than traditional methods, such as support vector machines (SVM).[8] A saliency-guided unsupervised feature learning framework using deep network was proposed for scene classification.[9] Romero et al.[10] proposed using greedy layerwise unsupervised sparse features with a CNN for pixel classification. A systematic review of the state-of-the-art DL-based methods used in remote sensing image analysis was made by Zhang et al.[11] Motivated by the success of DL on various tasks in remote sensing, this study is the first to our knowledge to apply DL techniques for residential roof condition assessment.

The imagery used in our work was 1-in. ground resolution HR imagery and less than 1-in. ground resolution UHR imagery collected by EagleView Technologies. These superior resolution air photos can provide detailed information and enable the characterization of roof condition. Considering the nonhomogeneity of residential roofs, roof condition assessment methods using features derived from the entire rooftop will likely not provide promising results. Instead of treating the roof as a uniform object, a better approach is proposed in this paper. We propose first segmenting the roof into more homogenous regions and then characterizing the roof based on the categorization of individual regions.

Thus, the task is divided into two stages: (1) roof segmentation, followed by (2) classification of roof segments. The following outlines our proposed method. A self-tuning segmentation method for the roof condition assessment is proposed in this paper. The algorithm begins with an oversegmentation yielded by the simple linear iterative clustering (SLIC) superpixel method.[12] Our proposed deep learned feature, color-holistcally nested edge detection (HED) histogram, is then extracted to represent each superpixel. A similarity measure is defined to measure the feature similarity. In the region merging process, the most similar adjacent regions are merged iteratively. An unsupervised evaluation metric Q is incorporated into the merging process to select the optimal result. After the roof segmentation, roof images are divided into homogeneous regions. Our proposed k-pixel extension technique is then applied to expand the training data, which enables the implementation of DL techniques on the limited data. Pretrained deep residual networks (ResNet)[13] were fine-tuned on the augmented roof regions to yield final classification results.

The paper is organized as follows. Section 2 introduces the EagleView imagery dataset. Section 3 presents the proposed segmentation algorithm followed by the classification algorithm for the roof segments. Section 4 provides experimental results. Discussion on the limitations of the proposed algorithm and recommendations for future research is provided in Sec. 5. Conclusions are presented in Sec. 6.

2 Data

A leading provider of aerial imagery, EagleView Technologies, provided two sets of color airborne imagery for this study: 1-in. ground resolution HR imagery and less than 1-in. ground

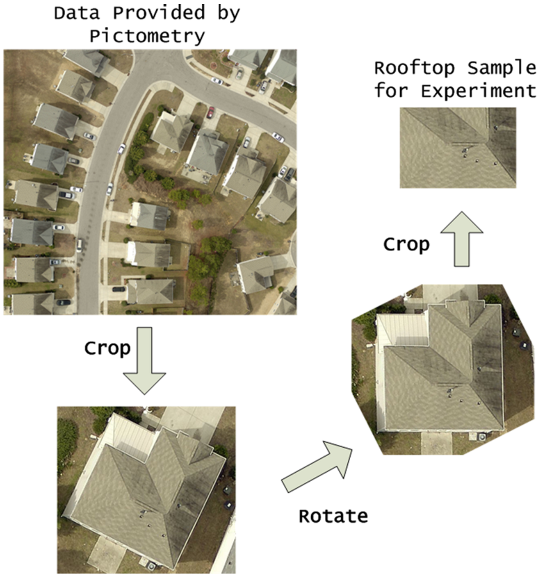

Fig. 1 Example of manual residential roof extraction on HR imagery.

resolution UHR imagery. We manually extracted sample roof images, as shown in Fig. 1. Rooftops of interest were extracted from the imagery and a rotation was followed by additional cropping to obtain the final experimental sample images. Rotation was applied before and after the cropping operation for the convenience of postprocessing. Thus, there was no background areas introduced into our experiment sample images. Also, the rotation to align the building outline with the image sides was not always necessary, if the orientation of the building was due north or due west.

Our dataset consisted of 110 intact and 164 impaired roofs manually extracted from the HR data, and 98 intact and 162 impaired roofs extracted from UHR imagery. Examples of the roof images are shown in Fig. 2. For the HR data, Fig. 2(a) shows an intact roof and Fig. 2(b) shows an impaired roof with a large area of cosmetic damage. For the UHR data, Fig. 2(c) shows an intact roof covered by tree and shadow. Figure 2(d) shows an impaired roof. Since the resolution of the UHR data is less than 1-in., detailed information like the roof fan system with a rusty pipeline can be seen clearly. A small impaired region is located at the end of the rusty pipeline.

As stated in Sec. 1, relatively little research has been published specifically on the identification of lower damage grade roofs. The main reason is that previous datasets could not provide sufficient information to support this application. For this study, the large amount of detailed information provided by the HR and UHR data makes characterization of roof condition using aerial imagery feasible.

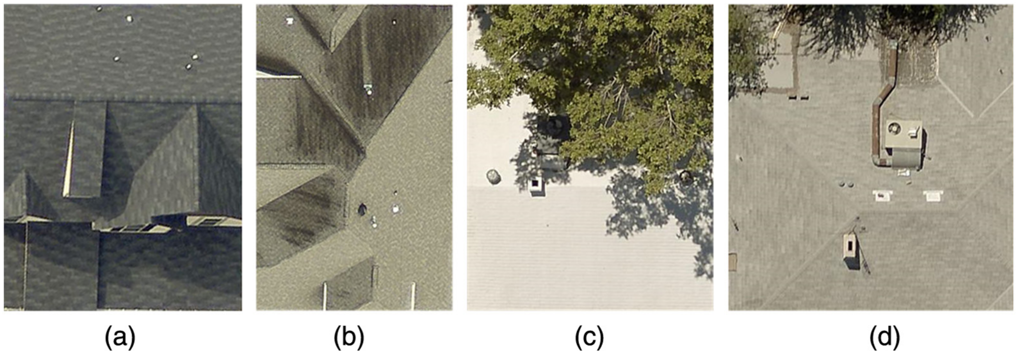

Fig. 2 Examples of roof images used in study: (a) HR intact roof, (b) HR impaired roof, (c) UHR intact roof, and (d) UHR impaired roof.

3 Method

The main challenge of roof condition assessment was the high within-class diversity in both intact and impaired roofs. Features extracted from the entire rooftop were easily seen not to be sufficient to characterize the roof condition. Instead of treating the roof as a whole, our proposed method divided the roof into parts and characterized the roof based on the categorization of individual parts.

3.1 *Roof Segmentation*

In this section, we will detail the steps of our proposed segmentation algorithm. First, the SLIC superpixel algorithm was applied to partition the image into homogeneous regions. Then, our proposed features, color-HED histogram, were extracted for region representation. A similarity measure of color-HED features was defined. The image was then represented by a region adjacency graph (RAG). The next step was region merging, where most similar neighboring regions were merged at each iteration.

Conventional region merging methods would stop when the similarity between any two adjacent regions is less than a preset threshold.[14] The preset threshold is fixed to a certain value and is expected to give satisfactory performance for images similar to those that were used to tune the parameter.[15] However, for our study, there is no single parameter value, which can result in the best possible segmentation for all the images due to the high diversity of roof images. It is not practical and unwise to perform manual parameter adjustment for each roof sample image. Thus, a self-tuning region merging segmentation method is proposed. An unsupervised evaluation metric Q is introduced and used to quantify the merging steps into a score list. The result of our algorithm corresponds to the segmentation achieved at the merging step with the minimal Q.

3.1.1 *Oversegmentation using SLIC superpixel*

The proposed region merging method is based on an initial oversegmentation using the SLIC superpixel method, which groups pixels into small homogeneous regions. The SLIC superpixel algorithm is selected because of its excellent boundary adherence,[12] which is a necessary prerequisite to eventually obtain the accurate shape and area of minor damage areas, such as missing shingles or cosmetic damage. Meanwhile, SLIC is fast and easy to use.[12]

SLIC is often intended to be applied to images in the CIELAB color space.[12] In our algorithm, the roof image was first converted to CIELAB format before SLIC superpixel was performed. Instead of determining the number of desired superpixels in each roof image, we set the nominal size of the superpixel to 15×15. The compactness parameter is used to control the tradeoff between superpixel compactness and boundary adherence,[12] which was empirically set to 7 in this study.

3.1.2 *Color-HED feature extraction*

Feature representations learned through DL techniques often outperform traditional hand-engineered features. Thus, a deep learned feature, color-HED histogram, is proposed for region representation.

Instead of extracting image features using traditional methods, features can be extracted using a CNN. Each layer of CNN produces a response to an input image. The layer at the beginning of the network learns features similar to a Gabor filter and color blobs.[16] Thus, those features are not specific to a particular task. The features computed by the last layer combine all the basic features into a richer one and thus depend greatly on the chosen task.

Our study has limited data, thus a fully supervised deep architecture would generally overfit the training data. Rather than learning a full deep representation, an easy way is to use a pretrained CNN learned from related tasks as feature extractor. HED[17] was chosen for our task. It was first proposed as a DL architecture for edge detection in natural images.[17] It is a system inspired by fully CNNs with additional fine-tuning on top of VGGNet.[18] The HED networks comprise a single stream deep network with multiple side outputs.

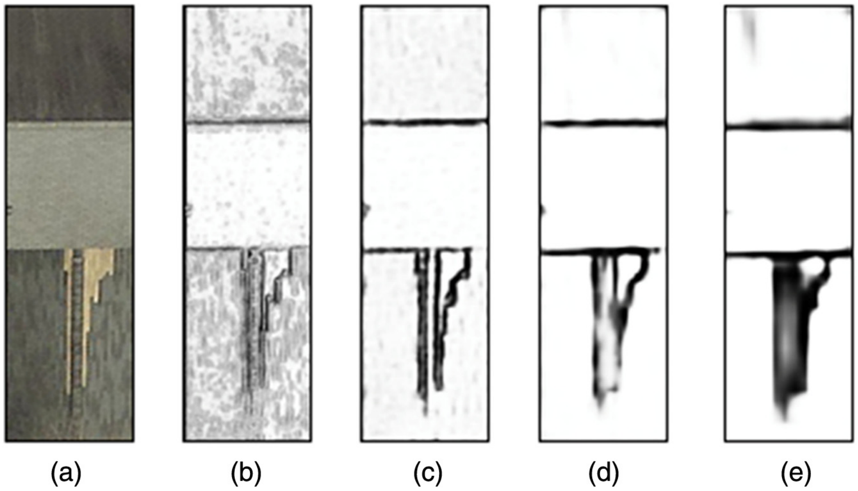

Fig. 3 Examples of HED side outputs on HR roof image: (a) original image, (b) HED side output 1, (c) HED side output 2, (d) HED side output 3, and (e) HED side output 4.

Each side-output produces a corresponding edge map at different scale levels. Combining the side outputs, a weighted-fusion layer yields the final result.

As an example shown in Fig. 3, the pretrained HED network is applied to the roof image. As we go deeper in the HED network, the later outputs, such as side outputs 3 and 4 shown in Figs. 3(d) and 3(e), are more like edge probability maps and cannot be used as region features. As can be seen, activation early in the network, such as side output 1 shown in Fig. 3(b), is more suitable for region feature representation.

Thus, a feature, color-HED histogram, is defined in this study. The color channels in RGB space and side output 1 of HED network constitute a four-dimensional feature map. In our algorithm, the image was divided into many primitive regions after the SLIC segmentation. Color-HED histograms were then computed to represent each region. Each channel in color-HED space was quantized into 16 bins. Each region was then represented by a vector of dimension $16^4 = 65536$.

3.1.3 *Similarity measure*

The order of merging is one of the essential issues in region merging segmentation.[19] It is controlled by the similarity measure between adjacent regions.

In our algorithm, the Bhattacharyya coefficient[20] was adopted to measure the color-HED feature similarity Sim between adjacent regions p and q:

$$\mathrm{Sim} = \mathrm{BC}(p, q) = \sum_{i=1}^{n} \sqrt{\mathrm{CH}_p^i \cdot \mathrm{CH}_q^i}, \tag{1}$$

where CH_p^i and CH_q^i are the normalized color-HED histograms of regions p and q. The superscript i represents the i'th element and n is the dimension of the color-HED histogram.

3.1.4 *Maximal similarity merging process*

After the distance measures Sim_c between all adjacent region pairs were computed, the image was represented as a RAG $G = (V, E, W)$.[21] Regions produced by SLIC superpixel algorithm were denoted as a set of nodes $v_i \in V$. The edge $(v_i, v_j) \in E$ between adjacent nodes had a corresponding weight $w(v_i, v_j) \in W$ to measure the dissimilarity of two nodes. The minimum dissimilarity corresponds to the maximal similarity denoted by distance Sim. The region merging segmentation was then performed by iteratively merging the most similar connected regions. After each merging, the color-HED features, RAG, and similarity ranking were updated.

3.1.5 *Self-tuning segmentation with unsupervised segmentation evaluation*

To design a self-tuning segmentation algorithm, the first thing we need to know is what is a good segmentation. Segmentation evaluation is usually done by visual inspection or supervised evaluation using a manually derived reference.[22] However, unsupervised segmentation evaluation is needed for a self-tuning segmentation algorithm. In the literature, relatively little research effort has been devoted to segmentation evaluation as compared to the development of segmentation algorithms.[23] There has been some fundamental research about unsupervised segmentation evaluation using intraregion homogeneity and inter-region disparity to access the segmentation result. However, no metric could handle well the semantic relationships presented in a complex scene. Thus, researchers rarely use unsupervised evaluation compared to supervised methods. For our study, the semantic relationships are not necessary to be considered, which makes it possible for us to incorporate unsupervised evaluation into our segmentation algorithm.

The next problem is which metric to use. An extensive evaluation of unsupervised evaluation metrics was presented in the survey of Zhang et al.[24] Based on the results of their experiments examining the performance of metrics on segmentation produced by the same algorithm with varying numbers of segments, the best performing metric Q[25] is selected for our algorithm.

Q is defined by

$$Q = \frac{\sqrt{R}}{10000(N \times M)} \sum_{i=1}^{R} \left\{ \frac{e_i^2}{1 + \log A_i} + \left[\frac{R(A_i)}{A_i} \right] \right\}, \tag{2}$$

where $N \times M$ is the size of the image, R is the number of regions, A_i and e_i^2 are the area in number of pixels, and the squared color error of the i'th region v_i, respectively. The squared color error of the i'th region v_i is defined as follows:

$$e_i^2 = \sum_{p \in v_i} [C(p) - \hat{C}_i]^2, \tag{3}$$

where $C(p)$ denotes the value of pixel p and $\hat{C}_i$ is the average value of i'th region. $R(A_i)$ represents the number of regions that have an area equal to A_i.[25] Since $R(A_i)/A_i$ typically has a very small value as compared to the first term in the summation,[26] $R(A_i)$ is fixed as 1 during the implementation. Lower Q values mean better segmentation quality. Note that 10,000 is replaced by 1000 during implementation. This change does not affect the Q's function on segmentation assessment. It only changes the scale of Q for plotting purposes.

Next, we demonstrate how to incorporate Q into our algorithm. The sample roof image was segmented into superpixels using the SLIC algorithm. The Q metric was computed and updated during each merging step. The series of merging steps were compiled into a score list. The selected segmentation result corresponded to the step with minimal Q value.

However, the quantitative score Q is known to be biased toward undersegmentation.[24] For some relatively simple roof images, Q would select the final step of region merging processing, where there is no segmentation at all. To handle this limitation of Q, a similarity threshold is introduced. The merging stops at this preset similarity threshold to get rid of those undersegmented candidates. There are two requirements that need to be satisfied for the selection of a similarity threshold. On one hand, we need to ensure that the segmentation is performed to some extent, so there are enough segmentation candidates for Q to make a right decision. On the other hand, the threshold needs to make sure that the small regions, such as a chimney or an area of impaired region, would not be merged into the roof background, so those undersegmented results will not be there to confuse Q. The similarity threshold was empirically selected to be 0.4 because it gave satisfactory performance on both aspects for most images that were used for tuning.

Q can be incorporated into the algorithm from the beginning of region merging processing. However, for computational efficiency, the region merging process was divided into two steps, as shown in Fig. 4, and Q is only introduced during the postregion merging. Preregion merging was stopped at a preset number of regions. The number of regions is selected based on the resolution of dataset and the hypothesis that the optimal segmentation of residential roof under this resolution contains fewer regions than the preset number. For the 1-in. resolution HR dataset,

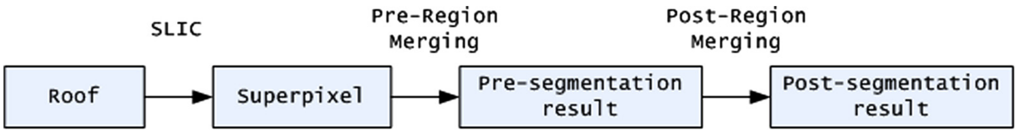

Fig. 4 Framework of self-tuning segmentation algorithm.

25 regions were selected to provide a sufficient number for the final region merging processing while speeding up the algorithm. For the UHR data with less than 1-in. resolution, this number of regions is increased to a relatively safe value of 50.

3.2 *Classification*

After roof segmentation, the roof images were divided into homogeneous regions. To characterize the roof condition, roof segment classification was applied. Then, the roof condition assessment can be performed based on the categorization of the roof segments.

3.2.1 *Data for classification*

For this study, we only consider segments with an area larger than 400 pixels. The reason is that the average area of our residential roof data is more than 60,000 (200 × 300) pixels. Those segments with an area less than 400 take less than 0.67% of the whole roof area. The information contained in those tiny regions was determined not to be sufficient for classification.

Our proposed segmentation algorithm was performed on 274 HR and 260 UHR roof images provided by EagleView Technologies. From these images, 1105 HR and 1120 UHR roof segments were generated and used for this study.

Examples of typical roof segment images are shown in Fig. 5. As shown in Fig. 5, the roof segments can be divided into nine classes, including impaired region, intact region, "structure" region, fan, shadow, tree, chimney, ridge, and other. The "other" class includes solar panels, ladder, and other clutters. They are combined into one "other" class since they were too few to reliably classify separately. Most of the images in the fan and the "other" class were extracted from the UHR imagery.

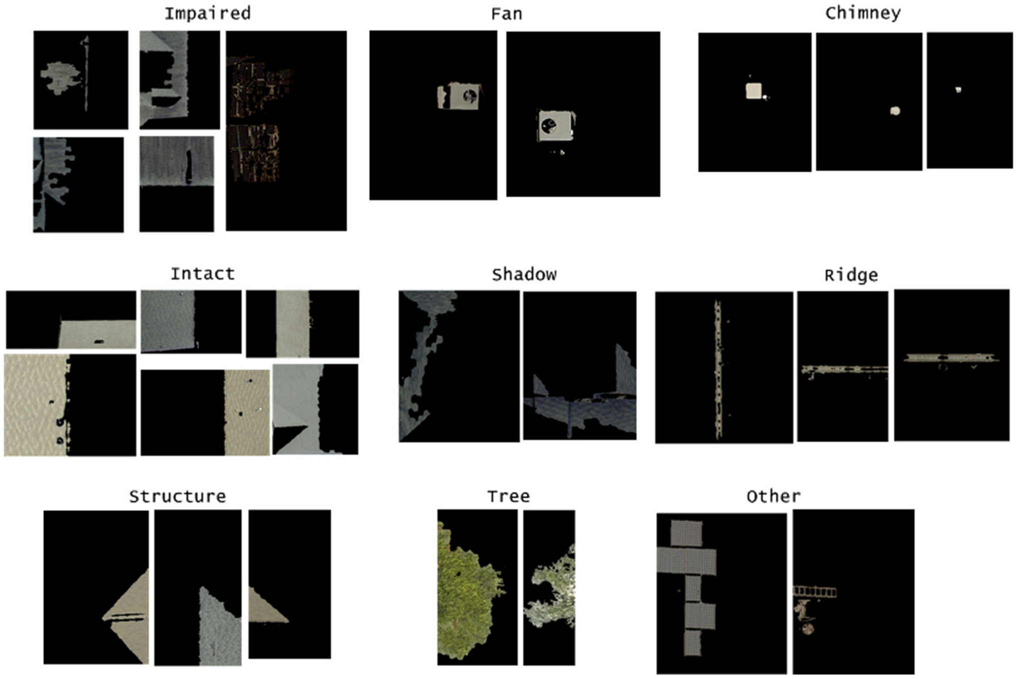

Fig. 5 Labeled roof segment images (mixture of HR and UHR data).

3.2.2 *Roof classification via hand-crafted features*

For comparison with existing techniques, we implemented a traditional classification method using hand-crafted features. Several hand-crafted features were designed and extracted from each roof segment. In our previous work, for example, we have used color histograms for spectral information. Gabor filters[27] were adopted to extract texture features from each segment. Several statistical indices were calculated to represent the statistical features inside each segment, including variance, standard deviation, uniformity, and entropy. We also considered the location information for each image segment. Normalized X and Y means were computed. Shape features, extent, and eccentricity were also considered.

Regions with a large area pack in much more information than the small ones. The information contained in those small regions may not be sufficient for classification. For example, even a human cannot differentiate between some impaired roof regions and shadowed areas based only on the segmented image. Thus, besides some classical hand-crafted features, such as color, texture, statistical, and shape features, similarity features were designed to include neighbor information for classification. The Bhattacharyya coefficient was adopted to measure the similarity of color histograms of two neighboring regions. The algorithm considers all adjacent segments for each segment being processed. The maximal and minimal similarity between the current segment to its neighbors was used as features for classification, as well as the number of its neighbors.

The optimal parameters for an SVM with a radial basis function (RBF) kernel were found through a grid search with fourfold cross-validation. The training data were partitioned into four equal sized folds. The folds preserve the percentage of samples for each class. A model was trained using threefold and validated on the remaining single fold. The cross-validation process was repeated four times. Each of the fourfold was used once as the validation set. The performance estimation was the average of four results obtained in the loop. Grid-search on C and γ was performed using fourfold cross-validation. Exponentially growing sequences of C and γ values are tried and the one with the best cross-validation score is picked. The performance of the trained classifier was tested on the test set.

3.2.3 *Roof classification via deep learning*

In recent years, CNNs have produced stellar results and their capabilities were illustrated in the ImageNet large scale visual recognition challenge.[28] The performance achieved is now close to humans. Motivated by the success of DL on various tasks, DL techniques were implemented for roof segment classification in this study.

In 2015, deep ResNet took the DL world by storm and won the ImageNet competition with an error score of 3.57%.[13] By adding skip connections that bypass a few convolution layers and learning a residual mapping, ResNet ensures a fluent information flow, allowing neural networks that are over 100 layers deep to be effectively trained.[13] Thus, ResNet was selected as the DL model for our study.

In our traditional method implementation, similarity features were designed to include neighbor information for classification. For our DL method, for small roof regions, neighbor information was introduced during data augmentation for training. All roof segments with adjacent background were interpolated to 321×321 pixels. For roof segments with an area less than 4000 pixels, the smallest rectangle containing the roof region was detected. Then, a larger rectangle mask was created as a k-pixel extension in all directions from the detected rectangle. The extension would stop if it touched the edge of the roof image. The portion of the original roof image specified by the k-pixel extension was added into the training dataset as the augmented training data. Random horizontal and vertical flipping was implemented for the augmentation. For small roof regions with area less than 4000, a series of k-pixel extensions were performed while k was specified as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 17, 20, 25, 30, 35, and 40.

An example of k-pixel extension on a shadowed region is shown in Fig. 6. From the original image shown Fig. 6(a), we can see that the segment image shown in Fig. 6(b) is fan system's shadow. Its extensions for training are shown in Figs. 6(c)–6(f) with k specified as 1, 10, 20, and 40.

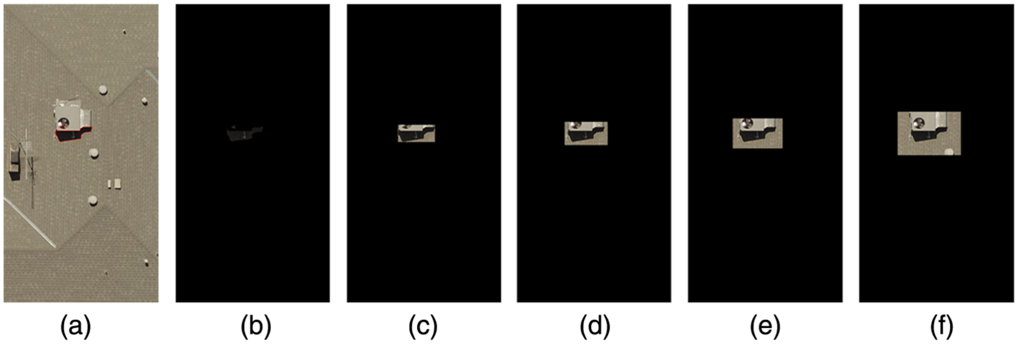

Fig. 6 Example of *k*-pixel extension technique on UHR data: (a) original image where the target region is highlighted by red outline, (b) segment image for training, (c) 1-pixel extension, (d) 10-pixel extension, (e) 20-pixel extension, and (f) 40-pixel extension.

The proposed data augmentation technique greatly expands the training set allowing us to implement the DL technique on this limited dataset and introduces different levels of neighbor information into the training processing.

Since the training data were limited, learning a full deep representation from scratch can be ineffective and time consuming.[29] Instead, it is common to pretrain a base network on a very large dataset like ImageNet, which contains 1.2 million images with 1000 categories.[16] Earlier features of a network have sufficient representational power and generalization ability. Thus, we can keep some of the earlier layers fixed for a target network, and the higher-level portion of the network is then randomly initialized and trained toward the target task.[16]

For our study, we used the ResNet with 50 layers pretrained on ImageNet as a starting point. We replace the 1000-way ImageNet classifier in the last layer with a classifier having as many targets as the number of roof segment classes. The models are trained for up to 5000 iterations. Then, we run back propagation on the network to fine-tune all layers for 1000 steps.

4 Results

4.1 *Segmentation*

In this section, we will demonstrate the segmentation results of our proposed algorithm on manually extracted roof sample images in the 1-in. resolution EagleView dataset.

The segmentation algorithm consists of steps of forming the SLIC superpixel, preregion merging, and postregion merging segmentation. To demonstrate the results of each step, a typical impaired roof sample image, shown in Fig. 7, is selected as an example. As shown in Fig. 7, the roof contains five tiny missing shingles, a region of cosmetic damage, a white chimney, and a triangular "structure." An ideal segmentation would be able to depict the general boundary

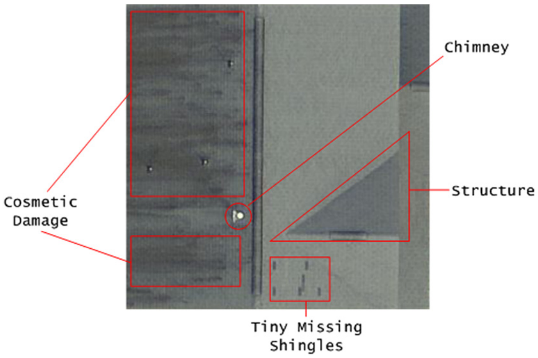

Fig. 7 Typical impaired roof (HR data).

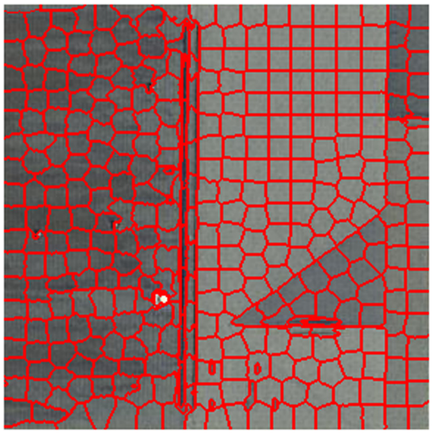

Fig. 8 SLIC superpixel result (HR data).

of the cosmetic damage area. Meanwhile, it would also isolate the tiny missing shingles into individual areas.

The SLIC superpixel algorithm was applied in the CIELAB color space. The nominal size of the superpixel was set to 15×15. To isolate tiny missing shingles, the compactness, which controls the ability of boundary adherence, was set to 7. As shown in Fig. 8, the boundaries of the five tiny missing shingles were well depicted.

Preregion merging processing started from the SLIC superpixel result. Color-HED histograms were computed to represent each region. The Bhattacharyya coefficient was used to measure the similarity between all adjacent region pairs. The most similar connected regions were merged iteratively until the number of regions was equal to 25 for HR data. The roof sample image, shown in Fig. 7, is a relatively complex roof scene containing tiny missing shingles, cosmetic damage, a chimney, and "structure." As shown in Fig. 9, the preregion merging result was still an overly segmented result.

Postregion merging segmentation started from the premerging result. Unlike premerging-processing, the unsupervised segmentation evaluation score, Q, was computed and updated during each step. The merging steps were quantified by the Q score, as shown in Fig. 9. Like conventional region merging segmentation approaches, the merging stopped when the preset similarity threshold of 0.4 was reached. The result of the last step is also shown in Fig. 9. The preset threshold was fixed to 0.4, because it gave satisfactory performance on most images that were used for tuning. However, for this specific roof, the result achieved by the conventional method is not perfect. Only two tiny missing shingles are isolated. The result of our proposed initial algorithm corresponds to the step with minimal Q value. At this step, all five tiny missing shingles were isolated. The boundary of the cosmetic damage area is also depicted.

To better illustrate the results of our proposed algorithm, additional representative results using HR data are provided in Fig. 10. The proposed modified algorithm is compared to the well-known compression-based texture merging (CTM) method.[30] Human segmentation truth boundary maps are also provided. To fairly compare these methods, the original NCuts superpixel method in the CTM software was replaced by SLIC for better performance. For CTM, we needed to adjust the threshold γ for a satisfactory segmentation. We ran CTM with parameter γ chosen at intervals in [2, 7] and found that $\gamma = 5$ gave good overall performance.

A "texture" roof is shown in the first row of Fig. 10. It shows that the CTM algorithm produced oversegmentation around the white chimney. It also created a sinuous edge around the ridge. The modified algorithm produced a clear and reasonable result. An intact roof example is shown in the second row of Fig. 10. The proposed algorithm produced a clearer boundary map while CTM generates weird boundaries around the white object on the roof. A "structure" roof is shown in the third row of Fig. 10. CTM fails to isolate the cosmetic damage area and produces a weird boundary around the chimney. A better boundary of the cosmetic damage area is generated by the proposed algorithm. A roof with cosmetic damage and missing shingles is shown in the fourth row of Fig. 10. It can be seen that the missing shingles are isolated well by both algorithms and a better shape of the cosmetic damage area is produced by our algorithm.

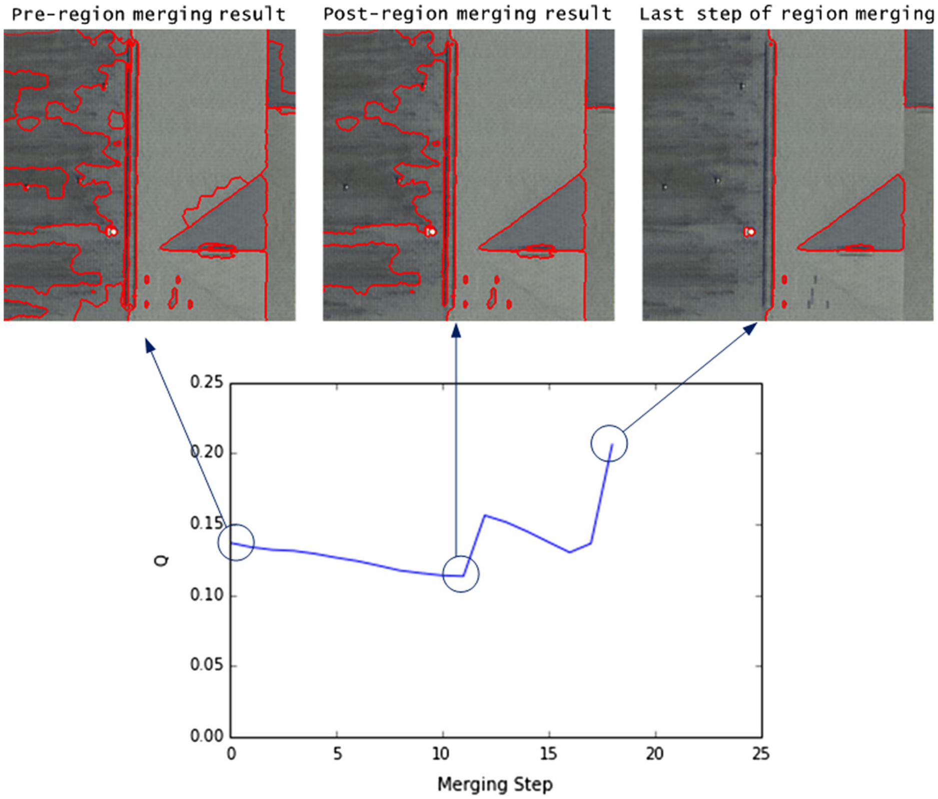

Fig. 9 Pre- and postregion merging results (HR data).

In summary, the proposed algorithm produced better visual segmentation results compared to CTM method. The CTM algorithm suffers from an oversegmentation problem and produces sinuous boundaries around ridges and chimneys. Meanwhile, CTM is not robust. It isolated cosmetic damage areas on some data but failed on other data.

4.2 *Classification*

In this section, we will demonstrate the roof segment classification results. After roof segmentation, roof images were segmented into homogeneous regions. For our HR data, the manually labeled roof segments were divided into 884 roof segments for training and 221 segments for testing. These 1105 roof segments can be divided into six classes, including impaired regions, intact regions, "structure" regions, ridges, and shadow regions. For our UHR data, 1120 roof segments were used for training and 285 for testing. Since the UHR data are collected from a different study site and with higher resolution compared to HR data, these 1405 roof segments were divided into nine classes. In addition to the six classes included in HR dataset, another three classes—chimney, fan, and other—are added.

Hand-crafted features—including color, texture, statistical, location, shape, and similarity features—were extracted from each segment. Grid search with fourfold cross-validation was applied on the training set to identify the best parameter pair for the SVM classifier with RBF kernel.

For our HR data, $C = 1380$ and $\gamma = 0.00287$ were found as the optimal parameter pair for the SVM with RBF kernel. The overall accuracy was 69.68%. A segment-level confusion matrix shown in Table 1 is used to describe the performance of the traditional method on HR data. From the confusion matrix, we can see that 11 impaired regions were misclassified as shadow regions since their texture was similar. For roof condition assessment, we are more concerned if the segment is intact or not. Thus, intact, structure, shadow, ridge, and tree classes can be combined into one, a pristine roof class. So, ignoring the intraclass error inside pristine class, the combined accuracy is 82.35%.

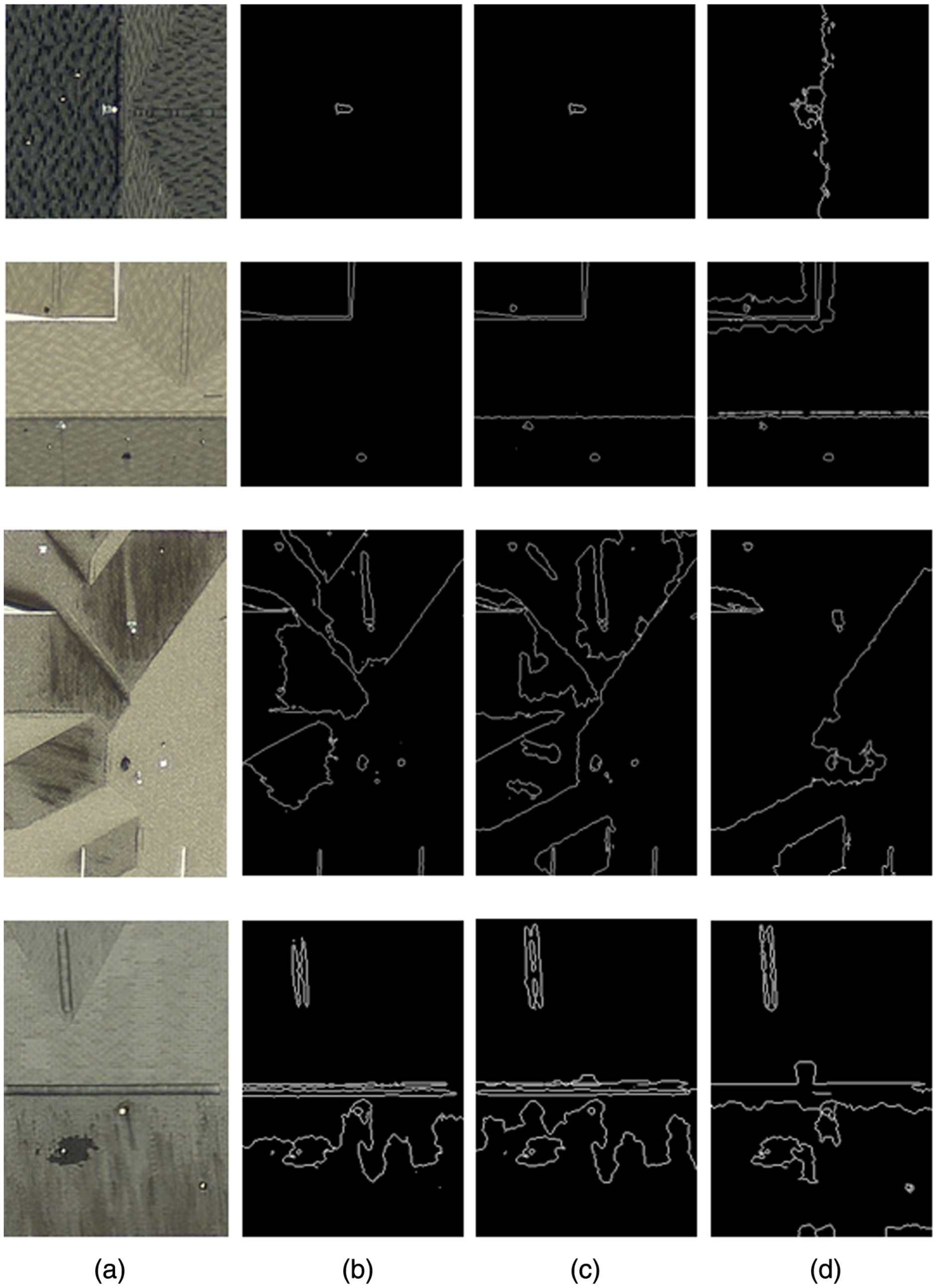

Fig. 10 Segmentation results (HR data) by human, proposed algorithms and CTM: (a) original image, (b) human segmentation, (c) proposed algorithm, and (d) CTM result.

For our proposed DL method, the pretrained ResNet with 50 layers was used as input. The 1000-way ImageNet classifier in the last layer was replaced with a six-class classifier, and RMSprop with a batch size of 32 was used. The learning rate started from 0.01 and a weight decay of 0.00004 were selected. The models were trained for 5000 steps. Then, all the layers were fine-tuned for another 1000 steps with a learning rate of 0.001 and a weight decay of 0.00004. The experiments were run using a single NVIDIA GeForce GTX 1080Ti GPU and implemented with TensorFlow. This fine-tuning process took $\tilde{4}$ h. The overall accuracy was 82.35%, more than 12% higher than the traditional method.

A segment-level confusion matrix shown in Table 2 is used to describe the performance of the DL method on HR data. Compared to the results achieved by the traditional method in Table 1, 40 impaired regions were classified correctly compared to 34 by traditional method and only 4 of impaired regions were misclassified as shadow compared to 11 by traditional method. For intact

Table 1 Confusion matrix for roof segments classification via traditional method on HR data.

		Prediction					
		Impaired	Intact	Structure	Shadow	Ridge	Tree
Truth	Impaired	34	2	4	11	0	0
	Intact	14	79	16	3	1	0
	Structure	3	2	24	2	0	0
	Shadow	4	3	0	6	0	0
	Ridge	1	0	0	1	9	0
	Tree	0	0	0	0	0	2

Table 2 Confusion matrix for roof segments classification via DL method on HR data.

		Prediction					
		Impaired	Intact	Structure	Shadow	Ridge	Tree
Truth	Impaired	40	6	0	4	0	1
	Intact	7	104	2	0	0	0
	Structure	0	9	21	1	0	0
	Shadow	3	4	0	6	0	0
	Ridge	1	0	0	0	10	0
	Tree	0	0	0	1	0	1

regions, our DL method correctly identified 104 intact roof regions compared to 79 by the traditional SVM. Only 7 of the intact regions were misclassified as impaired segment compared to 14 by the SVM. For our roof condition application, intact, structure, shadow, ridge, and tree classes can be combined into one pristine roof class. The combined accuracy was then 90.04%, upto 7.6% compared to the traditional SVM method.

For our UHR data, $C = 1372$ and $\gamma = 0.00268$ were selected though grid search with four-fold cross-validation for the SVM classifier with RBF kernel. The overall accuracy was 74.38%, which is better than its performance on HR data. A segment-level confusion matrix shown in Table 3 is used to describe its performance. The 41 of 46 shadowed regions were correctly labeled compared to 6 of 13 for HR data by traditional SVM method. So, we can say that, with the increase of data resolution and training data, identification of shadow seems to be easier. From the point of view of roof condition assessment, intact, structure, shadow, ridge, tree, chimney, fan, and other classes can be combined into one pristine roof class. The combined accuracy was 88.42%. The increase of data resolution improved the traditional method's performance to some degree.

For our DL method on UHR data, the 1000-way ImageNet classifier in the last layer was replaced with a nine-class classifier. Even though the traditional SVM method did a better job on UHR data compared to its performance on HR data, the overall accuracy of our DL method on UHR data was 82.80%, 8.42% higher than the traditional method. A segment-level confusion matrix shown in Table 4 is used to describe the performance of our DL method on the UHR data. From the confusion matrix, we can see that the performance on most classes was improved compared to the traditional method except for the shadow class. Ignoring the intraclass error inside pristine class, the combined accuracy is 91.57%, 3.05% higher than the traditional method.

Table 3 Confusion matrix for roof segments classification via traditional method on VHR data.

		Prediction								
		Impaired	Intact	Shadow	Tree	Chimney	Ridge	Fan	Structure	Other
Truth	Impaired	16	5	1	4	4	0	2	2	0
	Intact	6	65	0	0	1	1	0	7	0
	Shadow	2	0	41	2	0	1	0	0	0
	Tree	1	1	1	38	0	1	0	0	1
	Chimney	1	0	0	0	16	0	6	0	1
	Ridge	3	0	0	0	1	17	2		0
	Fan	1	0	0	0	8	0	10	0	0
	Structure	1	3	0	0	0	0	0	8	0
	Other	0	0	0	1	1	0	0	0	1

Table 4 Confusion matrix for roof segments classification via DL method on VHR data.

		Prediction								
		Impaired	Intact	Shadow	Tree	Chimney	Ridge	Fan	Structure	Other
Truth	Impaired	18	9	2	1	3	1	0	0	0
	Intact	4	73	0	1	0	0	0	2	0
	Shadow	1	0	39	3	0	3	0	0	0
	Tree	2	0	0	41	0	0	0	0	0
	Chimney	0	1	1	0	20	0	2	0	0
	Ridge	1	0	1	0	0	22	0	0	0
	Fan	0	0	1	0	6	1	11	0	0
	Structure	0	2	0	0	0	0	0	10	0
	Other	0	0	0	1	0	0	0	0	2

Some examples of correct classification by our DL method are shown in Figs. 11 and 12. Figure 11 shows examples of pristine segments with correct classification. Figure 11(a) is a typical intact segment with a large area and uniform texture, which is the most common intact roof segment in our dataset. The texture of Fig. 11(b) is rough and caused some trouble for the traditional SVM method using hand-crafted features. Figure 11(c) is a shadowed region, which can be easily misclassified as an impaired region. This is a problem for the traditional method and even for a human. Figure 11(d) is a small intact segment with a regular shape. Its texture is similar to Fig. 11(b). Figure 11(e) is an intact segment with an irregular shape. These kinds of segments are most difficult to identify. When we manually label them, we have to rely on the background information. All of them were correctly classified by our fine-tuned ResNet.

Figure 12 shows examples of impaired segments, which were correctly classified by our trained ResNet. Figure 12(a) is a typical impaired segment with cosmetic damage. It has a regular shape and covers half of the rooftop. Figure 12(b) is an impaired segment with irregular shape. It suffers less cosmetic damage than Fig. 12(a). Figures 12(c) and 12(d) are relatively small segments with cosmetic damage. All of them were identified correctly by ResNet.

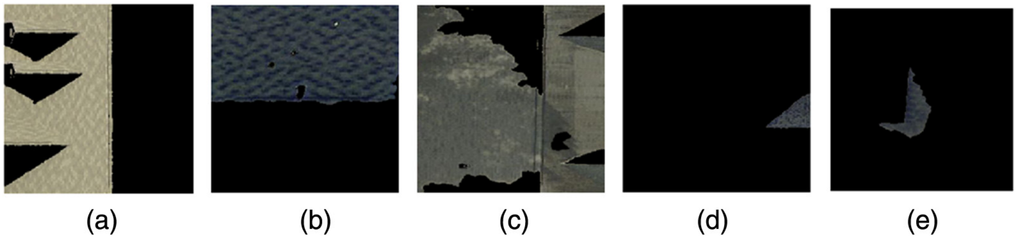

Fig. 11 Examples of intact segments, which are correctly classified by ResNet (HR data): (a) typical intact roof region, (b) intact region with rough texture, (c) shadowed region, (d) small intact region with regular shape, and (e) small intact region with an irregular shape.

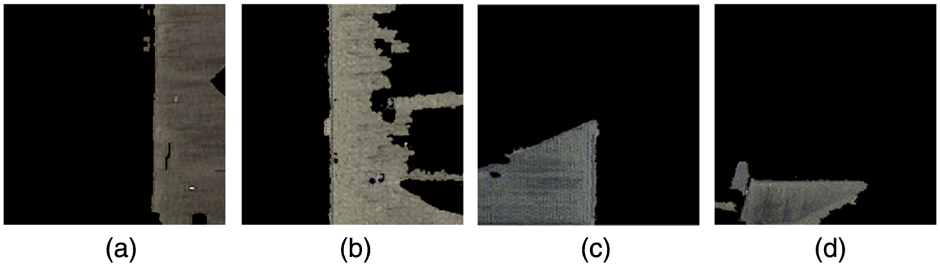

Fig. 12 Examples of impaired segments, which are correctly classified by ResNet (HR data): (a) and (b) large impaired roof segments, (c) and (d) small impaired roof segments.

In summary, our DL method outperformed a traditional method using hand-craft features in both our HR and UHR datasets, even though the traditional method benefited from the increase of data resolution. Furthermore, our DL method provided stable performance independent of the resolution of data.

5 Discussion

In this section, typical misclassification examples of our proposed method are reviewed and the potential causes of the mistakes are discussed. Limitations of the proposed algorithm are investigated and recommendations for future research are proposed.

One misclassification case is shown in Fig. 13. Our method misclassified a shadow region as an impaired roof region. Distinguishing impaired roof areas from shadows often was the hardest

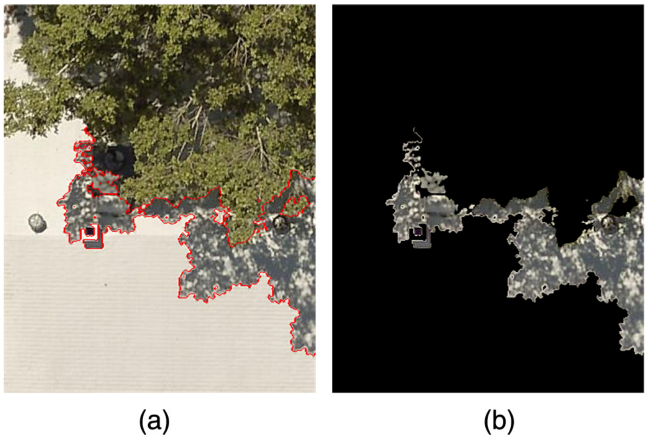

Fig. 13 Example of a shadow region (UHR data), which was misclassified as an impaired area by our method: (a) original image where the target region are highlighted by red outline and (b) segment image for classification.

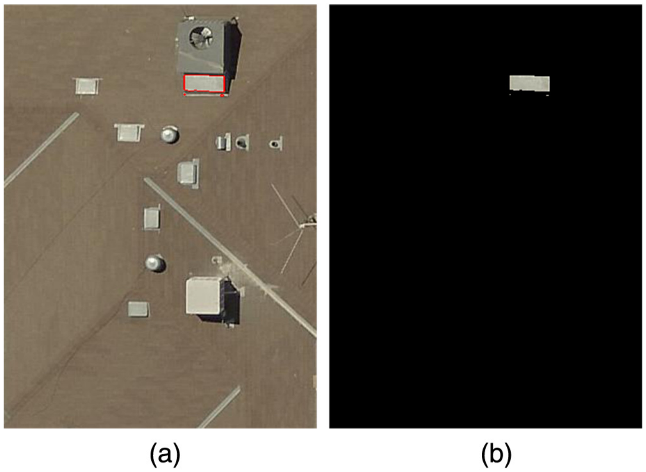

Fig. 14 Example of a fan region (UHR data), which is misclassified as chimney by our method: (a) original image where the target region is highlighted by red outline and (b) segment image for classification.

part of this task because these two classes share similar texture and shape features. By observing the texture of Fig. 13(b) carefully, humans can still recognize it as a shadow area because of the dappled pattern. Unfortunately, our trained ResNet did not recognize as well as a human.

One way to improve the performance of our method in these cases would be to increase the size of dataset. The current dataset used for this study contained around 2000 labeled segments, which is still too small to realize the full potential of deep networks. However, manual region labeling takes significant resources and time. Until larger datasets become available, our recommendation is to use the existing trained deep network to classify data and then manually correct erroneous results.

Another two typical misclassification errors are shown in Figs. 14 and 15. For Fig. 14, our method misclassified the fan region as a chimney. Relying only on Fig. 14(b), it is nearly impossible even for a human to decide that the region is belonged to a fan system or a chimney. With the help of the original data shown in Fig. 14(a), a judgment can be made that the segment is the pipeline connecting fan system to roof.

For Fig. 15, our method misclassified an intact roof region as an impaired one. From the information provided by Fig. 15(b), the shape of the segment is irregular and the area is not big in which features are associated with an impaired area. Referring the original data shown in

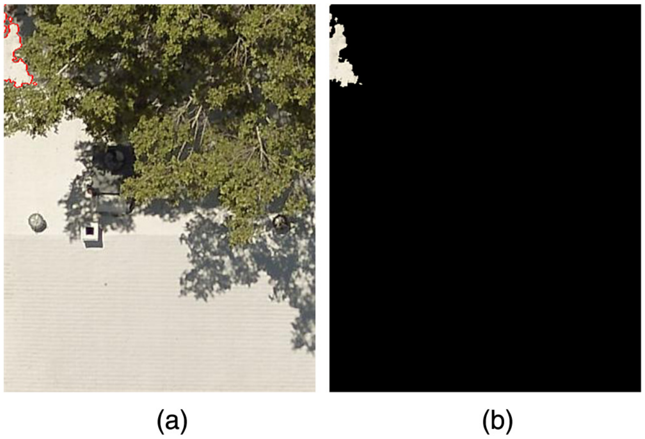

Fig. 15 Example of an intact area (UHR data), which is misclassified as impaired region by our method: (a) original image where the target region is highlighted by red outline and (b) segment image for classification.

Fig. 15(a), this intact region is isolated by a tree and that is why it has a relative small area with an irregular shape.

The root cause of these mistakes is one limitation of our method: when the area of the segment is less than a certain value, the information contained in the segment image may not be sufficient for classification. This is also one of the reasons why tiny segments with an area less than 400 pixels have not been considered in the study.

To overcome this limitation, one possible approach would be to introduce local neighbor information around the target segment region during the classification. In our method, the neighbor information was introduced for training using our proposed k-pixel extension data augmentation technique. However, there was no neighbor information involved when the trained deep network is applied.

Future research should be focused on how to introduce neighbor information into the classification step. For example, during classification, for those segments with area less than certain value, k-pixel extension could also be performed. An original segment and a series of its extension images could be generated as candidates. The trained deep network can provide predicated labels for all the candidates. The challenge of the research will be how to use those labels for generating the final class. Perhaps, a weighted voting algorithm could be used.

6 Conclusion

As the remote sensing image analysis community looks beyond the categorization of the objects with increasing interest in the characterization of objects, this paper has provided an example of what modern DL techniques can offer to the problem of assessing the condition of residential roofs. Instead of treating the object (roof) as a whole, the proposed method divides the object into parts and characterizes the object based on the categorization of individual parts.

In the context of our specific application of roof condition assessment considered in this paper, previous research was found to be concentrated on building damage detection, which is only aimed at identifying heavily damaged buildings. Relative little research has been published specifically on roof condition assessment with the emphasis placed on grading lower damaged roof. According to previous reviews, heavy damage detection can be addressed by traditional techniques. The automated identification of lower damage grades was identified as an open research problem, even with a submeter resolution data. This was the motivation and objective that our proposed method aimed to resolve.

HR data with 1-in. resolution and UHR data with less than 1-in. resolution provided by EagleView Technologies were used for our study. The proposed method operated at the level of an individual roof and divided the task into two stages: (1) roof segmentation, followed by (2) classification of the segmented roof regions. The roof segmentation algorithm began with an oversegmentation result yielded by the SLIC superpixel method. Our proposed color-HED features were extracted to represent each superpixel. The region merging process merged the most similar adjacent regions iteratively. An unsupervised evaluation metric Q was incorporated into the merging process to select the optimal result. After the roof segmentation, the roof images were divided into homogeneous regions. A data augmentation technique, k-pixel extension, was proposed. It expanded the training set to enable the implementation of DL techniques on the limited data while introducing different levels of neighbor information. A pretrained ResNet was fine-tuned on the augmented dataset for classification.

To build an end-to-end system for application, our proposed method should be bundled with a building detection algorithm. The system takes HR or UHR aerial imagery as the input. Individual roof images generated by the building detection algorithm are provided to our algorithm. In the end, a condition report for each roof is generated according to the area of impaired roof region.

The effectiveness of our proposed algorithm was demonstrated on both HR and UHR data. The proposed algorithm provided a promising result and outperformed traditional machine learning method using hand-crafted features on both datasets.

Typical misclassification examples of our proposed method were reviewed and the cause of the mistakes is discussed. We investigated the limitations of the proposed algorithm: the information contained in the segment image having a relatively small area may not be sufficient for classification. Moving forward, our research will seek to prepare additional data for roof condition assessment. More training data will unleash additional potential of deep networks for this application. Furthermore, we will investigate the classification of segments having tiny areas.

References

1. G. Grünthal, "Conseil de l' Europe," in *European Macroseismic Scale (EMS)*, European Centre for Geodynamics and Seismology, Luxembourg (1998).
2. B. Sirmacek and C. Unsalan, "Damaged building detection in aerial images using shadow information," in *4th Int. Conf. on Recent Advances in Space Technologies*, pp. 249–252, IEEE (2009).
3. D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.* **48**(5), 2403–2420 (2010).
4. C. Bignami et al., "Objects textural features sensitivity for earthquake damage mapping," in *Joint Urban Remote Sensing Event (JURSE)*, pp. 333–336, IEEE (2011).
5. M. Gerke and N. Kerle, "Automatic structural seismic damage assessment with airborne oblique pictometry© imagery," *Photogramm. Eng. Remote Sens.* **77**(9), 885–898 (2011).
6. S. H. Samsudin, H. Z. Shafri, and A. Hamedianfar, "Development of spectral indices for roofing material condition status detection using field spectroscopy and Worldview-3 data," *J. Appl. Remote Sens.* **10**(2), 025021 (2016).
7. L. Dong and J. Shan, "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques," *ISPRS J. Photogramm. Remote Sens.* **84**, 85–99 (2013).
8. W. Hu et al., "Deep convolutional neural networks for hyperspectral image classification," *J. Sens.* **2015**, 1–12 (2015).
9. F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.* **53**(4), 2175–2184 (2015).
10. A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.* **54**(3), 1349–1362 (2016).
11. L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: a technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.* **4**(2), 22–40 (2016).
12. R. Achanta et al., "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2274–2282 (2012).
13. K. He et al., "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
14. S. Navlakha, P. Ahammad, and E. W. Myers, "Unsupervised segmentation of noisy electron microscopy images using salient watersheds and region merging," *BMC Bioinf.* **14**(1), 294 (2013).
15. B. Peng and O. Veksler, "Parameter selection for graph cut based image segmentation," in *Proc. of the British Machine Vision Conf.*, Vol. **32**, pp. 42–44 (2008).
16. J. Yosinski et al., "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, pp. 3320–3328, Curran Associates, Inc. (2014).
17. S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 1395–1403 (2015).
18. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. on Learning Representations*, arXiv:1409.1556, http://arxiv.org/abs/1409.1556 (2014).
19. B. Peng and D. Zhang, "Automatic image segmentation by dynamic region merging," *IEEE Trans. Image Process.* **20**(12), 3592–3605 (2011).
20. T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.* **15**(1), 52–60 (1967).

21. R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," in *Technical Symp. East*, pp. 2–9, International Society for Optics and Photonics (1985).
22. R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 929–944 (2007).
23. P. L. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Trans. Image Process.* **12**(2), 186–200 (2003).
24. H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: a survey of unsupervised methods," *Comput. Vision Image Understanding* **110**(2), 260–280 (2008).
25. M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognit. Lett.* **19**(8), 741–747 (1998).
26. H. Zhang, J. E. Fritts, and S. A. Goldman, "An entropy-based objective evaluation method for image segmentation," *Proc. SPIE* **5307**, 38–49 (2003).
27. J. Movellan, "Tutorial on Gabor filters," Technical Report, MPLab Tutorials, University of California, San Diego, California (2005).
28. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, Curran Associates, Inc. (2012).
29. J. Donahue et al., "DeCAF: a deep convolutional activation feature for generic visual recognition," in *Int. Conf. on Machine Learning*, Vol. **32**, pp. 647–655 (2014).
30. A. Y. Yang et al., "Unsupervised segmentation of natural images via lossy data compression," *Comput. Vision Image Understanding* **110**(2), 212–225 (2008).

Fan Wang received his PhD degree in imaging science in 2017 from the Carlson Center for Imaging Science at Rochester Institute of Technology, Rochester, New York. He is currently working in the Image Quality Group in GoPro Inc., San Mateo, California. His research interests include computer vision and machine learning.

John P. Kerekes received his BS, MS, and PhD degrees from Purdue University in 1983, 1986, and 1989, respectively, all in electrical engineering. From 1989 to 2004, he was a technical staff member at Lincoln Laboratory, Massachusetts Institute of Technology. In 2004, he joined the Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, where he is currently a professor and a director of the Digital Imaging and Remote Sensing Laboratory.

Zhouyi Xu completed her master's degree from the China Academy of Space Technology in 2013 and her undergraduate study at Harbin Institute of Technology in 2010. She is currently working on data mining in Tencent. Her research interests lie in the field of computer vision, deep learning, and data mining.

Yandong Wang has conducted extensive research work in photogrammetry and remote sensing, including automatic triangulation of aerial images and oblique images, automatic generation of point cloud and automatic extraction of building information from digital images. He has authored and coauthored a number of journal and conference papers in photogrammetry. His interests include automatic extraction of 3-D information of objects from images, automatic feature extraction, and image understanding.