Multisource Region Attention Network for Fine-Grained Object Recognition in Remote Sensing Imagery

Gencer Sumbul, *Student Member, IEEE*, Ramazan Gokberk Cinbis, and Selim Aksoy, *Senior Member, IEEE*

Abstract—Fine-grained object recognition concerns the identification of the type of an object among a large number of closely related subcategories. Multisource data analysis that aims to leverage the complementary spectral, spatial, and structural information embedded in different sources is a promising direction toward solving the fine-grained recognition problem that involves low between-class variance, small training set sizes for rare classes, and class imbalance. However, the common assumption of coregistered sources may not hold at the pixel level for small objects of interest. We present a novel methodology that aims to simultaneously learn the alignment of multisource data and the classification model in a unified framework. The proposed method involves a multisource region attention network that computes per-source feature representations, assigns attention scores to candidate regions sampled around the expected object locations by using these representations, and classifies the objects by using an attention-driven multisource representation that combines the feature representations and the attention scores from all sources. All components of the model are realized using deep neural networks and are learned in an end-to-end fashion. Experiments using RGB, multispectral, and LiDAR elevation data for classification of street trees showed that our approach achieved 64.2% and 47.3% accuracies for the 18-class and 40-class settings, respectively, which correspond to 13% and 14.3% improvement relative to the commonly used feature concatenation approach from multiple sources.

Index Terms—Deep learning, fine-grained classification, image alignment, multisource classification, object recognition.

I. INTRODUCTION

NEW generation sensors used for remote sensing have allowed the acquisition of images at very high spatial resolution with rich spectral information. A challenging problem that has been enabled by such advances in sensor technology is *fine-grained object recognition* that involves the identification of the type of an object in the domain of a large number of closely related subcategories. This problem differs from the traditional object recognition and classification tasks predominantly studied in the remote sensing literature in at least three main ways: 1) differentiating among many similar categories can be much more difficult due to low between-class variance; 2) difficulty of accumulating examples for a large number of similar categories can greatly limit the training set sizes for some classes; and 3) class imbalance can cause the conventional supervised formulations to overfit to more frequent classes and ignore the ones with limited number of samples. Such major differences lead to an uncertainty in the applicability of existing approaches developed based on traditional settings. Thus, the development of methods and benchmark data sets for fine-grained classification is an open research problem, whose importance is likely to increase over time.

Fine-grained object recognition has received very little attention in the remote sensing literature. Oliveau and Sahbi [1] proposed an alternating optimization procedure that iteratively learned a dictionary-based attribute representation and a support vector machine (SVM) classifier based on these attributes for the classification of image patches into 12 ship categories. Branson *et al.* [2] jointly used aerial images and street-view panoramas for fine-grained classification of street trees. They concatenated the feature representations computed by deep networks independently trained for the aerial and ground views and fed these features to a linear SVM for classification of 40 tree species. In [3], we studied the more extreme zero-shot learning (ZSL) scenario where no training example exists for some of the classes. First, a compatibility function between image features extracted from a convolutional neural network (CNN) and auxiliary information about the semantics of the classes of interest was learned by using samples from the seen classes. Then, recognition was done by maximizing this function for the unseen classes. Experiments were done by using an RGB image data set of 40 street tree categories.

New approaches that aim to learn classifiers under the presence of low between-class variance, small sample sizes for rare classes, and class imbalance can overcome these problems by enriching the data sets so that increased spectral and spatial content provides potentially more identifying information that can be exploited for discriminating instances of fine-grained classes. However, these two types of information do not necessarily come together in the same data source. Thus, multisource remote sensing is a promising research

direction for fine-grained recognition. For example, very high spatial resolution RGB data provides texture information, whereas multispectral (MS) and hyperspectral images contain richer spectral content. Furthermore, LiDAR-based elevation models can provide complementary information about the heights and other structural characteristics of the objects.

Multisource image analysis [4] has been a popular problem in remote sensing with a wide range of solutions including dependence trees [5], kernel methods [6], copula-based multivariate statistical model [7], active learning [8], and manifold alignment [9], [10]. Combining information from multiple data sources has also been the focus of data fusion contests [11]–[14] for land cover/use classification. Similar to their popularity in general classification tasks, deep learning-based approaches also received interest in multisource analysis. Deep networks have typically been used in the classification stage where raw optical bands and LiDAR-based digital surface models (DSM) [15], [16] as well as hand-crafted features from hyperspectral and LiDAR data [17] were concatenated and given as input to a CNN classifier, or in the feature extraction stage where independently learned deep feature representations from hyperspectral and SAR data [18] or hyperspectral and LiDAR data [19] were concatenated to form the input of a separate classifier. The output of a fully convolutional network trained on optical data and a logistic regression classifier trained on LiDAR data were also used in decision-level fusion by using a conditional random field [20].

A common assumption in all of these approaches is that the data sources are georeferenced or coregistered so that concatenation can be used for pixel-wise classification. Potential registration errors may not cause a problem during learning if one considers a small number of relatively distinct classes with many samples, or during testing where evaluation is done by pixels sampled from the inside of large regions labeled by land cover/use classes. Even though various approaches have been proposed for the registration of multisensor images [21], [22], finding pixel-level correspondences between images acquired from different sensors with different spatial and spectral resolution may not be error-free due to differences in the imaging conditions, viewing geometry, topographic effects, and geometric distortions [23]. Furthermore, errors at the level of a few pixels may not matter when the goal is to classify pixels sampled from land cover/use classes such as road, building, vegetation, soil, and water but can be very significant for fine-grained object recognition when the objects of interest, e.g., individual trees as in this paper, can appear as small as a few pixels even in very high spatial resolution images.

In this paper, we propose a multisource fine-grained object recognition methodology that aims to simultaneously learn the *alignment* of the images acquired from different sources and the *classification* model in a unified framework. We illustrate this framework in the fine-grained categorization of 40 different types of street trees using data from RGB, MS, and LiDAR sensors. Classification of urban tree species provides a suitable test bed for this challenging fine-grained recognition problem because of the difficulty of finding field examples for training data, fine-scale spatial variation, and
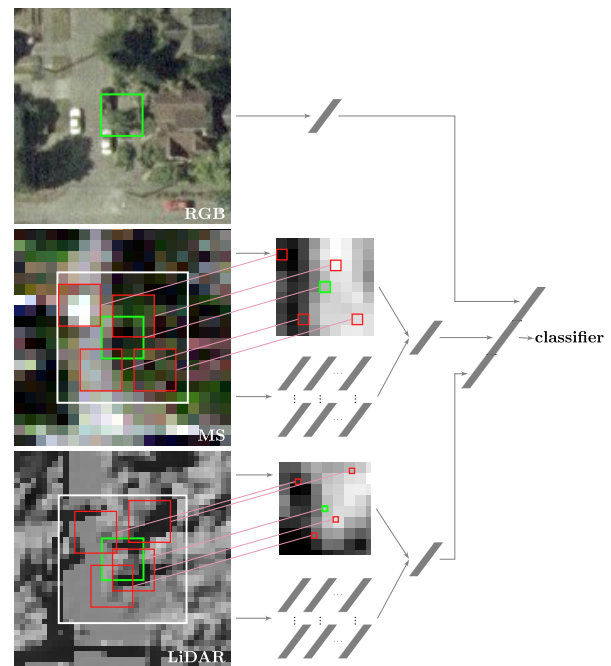

Fig. 1. Attention-driven representation of multisource data with imperfect pixel-level alignment. The RGB image, where a 25×25 pixel region (green box) centered at a verified ground truth tree location, is considered as the *reference* source. The corresponding regions in the MS and LiDAR data occupy 4×4 and 8×8 pixels (green boxes), respectively. The proposed *multisource region attention network* (MRAN) evaluates candidate regions (red boxes) sampled within a larger neighborhood (white box) around the expected object position, computes *attention scores* (shown as a grayscale map) that represent the confidence about each candidate containing the object, uses these scores to form the *attention-driven multisource representation*.

high species diversity [24]. Furthermore, appearance variations with respect to scale and spectral values, the difficulty of coregistration of small objects of interest in multiple sources, and rareness of certain species resemble the typical problems in fine-grained object recognition [25]. Consequently, differentiating the subcategories can be a very difficult task even with visual inspection using very high spatial resolution imagery. Classification of tree species has been previously studied in the remote sensing literature by using specialized approaches via fusion of hyperspectral and LiDAR data with linear discriminant [26], nearest neighbor [27], SVM [28], or random forest [24], [28], [29] classifiers. However, all of these approaches were specialized to tree classification, and none of them considered the alignment problem.

Our main contributions in this paper are as follows. The first contribution is a novel *multisource region attention network (MRAN)* that simultaneously learns to *attend* regions of source images that are likely to contain the object of interest and to *classify* the objects by using an *attention-driven multisource deep feature representation*. The problem flow is illustrated in Fig. 1. We assume that the objects exist in all sources but their exact positions are unknown except one particular source, the *reference*, which is verified with respect to the ground truth. The proposed network learns to: 1) compute per-source deep feature representations; 2) assign attention scores to candidate regions sampled around the expected object locations in remaining sources by relating them to

the reference; and 3) classify objects by using an attention-driven multisource representation that combines the feature representations and the attention scores from all sources. A deep neural network architecture is presented to realize the components of this framework and learn them in an end-to-end fashion. The second contribution is the detailed evaluation of this framework by using different combinations of source images from RGB, MS, and LiDAR sensors in fine-grained categorization of 40 different types of tree species. To the best of our knowledge, the proposed methodology is the first example for a generic unified framework for fine-grained object recognition by using any number of sources with different spatial and spectral resolutions via simultaneous learning of alignment and classification models.

The rest of this paper is organized as follows. Section II introduces the fine-grained data set. Section III describes the details of the methodology. Section IV presents the experiments. Section V provides the conclusions.

II. DATA SET

The data set in [3] contained 48 063 instances of street trees belonging to 40 categories. Each instance was represented by an aerial RGB image patch of 25×25 pixels at 1-ft spatial resolution, centered at points provided in the point GIS data. The names of the classes and the number of samples can be found in [3]. We use both an 18-class subset (named the supervised set in [3]) and the full set of 40 classes here.

This paper extends that data set with an 8-band WorldView-2 MS image (courtesy of DigitalGlobe, Inc.) and a LiDAR-based DSM with 2 m and 3-ft spatial resolution, respectively. Consequently, each tree instance corresponds to a 4×4 pixel patch in the MS data and an 8×8 pixel patch in the LiDAR data. Since the RGB image has the highest spatial resolution and the corresponding annotations were verified by visual inspection in [3], we consider it as the *reference* source. Even though each source image was previously georeferenced, precise pixel-level alignments among these sources were not possible as shown in Fig. 1. Thus, the proposed methodology in Section III aims to find the true, yet unknown, matching patch of 4×4 pixels in the neighboring region of 12×12 pixels in the MS data, and the corresponding patch of 8×8 pixels within a 24×24 pixel region in the LiDAR data. The neighborhood sizes are selected empirically using validation data. Using larger neighborhoods risks the inclusion of other trees that can confuse the attention mechanism, and smaller neighborhoods may not contain sufficient number of candidates.

III. METHODOLOGY

In this section, we first introduce the multisource object recognition problem and present a baseline scheme for it. Then, we explain our MRAN approach, followed by the details of the network architecture.

A. Multisource Object Recognition Problem

In the multisource object recognition problem, we assume that there exists M different source domains, where the space of samples from the mth domain is represented by $\mathcal{X}^m$. Our goal is to learn a classification function that maps a given object represented by a tuple of input instances from the source domains $(x^1 \in \mathcal{X}^1, \ldots, x^M \in \mathcal{X}^M)$ to one of the classes $y \in \mathcal{Y}$ where $\mathcal{Y}$ is the set of all classes.

In this paper, we focus on the problem of object recognition from multiple source images, where each source corresponds to a particular sensor, such as RGB, MS, LiDAR, and so on. We are particularly interested in the utilization of overhead imagery, where the samples are typically collected from cameras with different viewpoints, elevations, resolutions, dates, and time of day. Such differences in imaging conditions across the data sources make the precise spatial alignment of the images very difficult. The image contents may also differ due to changes in the area over time and occlusions in the scene.

In the following, we first present a simple baseline approach toward utilizing such multiple sources, and then, we explain our approach for addressing these challenges in a much more rigorous way.

B. Multisource Feature Concatenation

A simple and commonly used scheme for utilizing multiple images in classification is to extract features independently across the images and then concatenating them later, which is often called *early fusion*. More precisely, for each source m, we assume that there exists a feature extractor $\phi_m(x^m)$ which maps the input x^m to a d_m-dimensional feature vector. In this approach, it is presumed that each multisource tuple $x = (x^1, \ldots, x^M)$ consists of the images of the same object from all sources, and these images are spatially registered. Then, the multisource representation $\phi(x)$ is obtained by concatenating per-source feature vectors

$$\phi(x) = [\phi_1(x^1)^\top, \ldots, \phi_M(x^M)^\top]^\top. \tag{1}$$

Once the multisource representation is obtained, the final object class prediction is given by a classification function. This approach is illustrated using plate notation[1] in Fig. 2(a).

The main assumption of the simple feature concatenation approach is that the representation obtained independently of each source successfully captures the characteristics of the object within the target region. However, registration across the sources is usually imprecise, which requires choosing relatively large regions to ensure that all images within a tuple contain the same object instance. In this case, however, the features extracted from these relatively large regions are likely to be dominated by background information, which can greatly degrade the accuracy of the final classification model.

This problem is tackled by the proposed MRAN, explained in Section III-C.

C. Multisource Region Attention Network

A central problem in multisource remote sensing is the difficulty of registration of source images, as discussed earlier. To address this problem, we propose a deep neural network

---

[1]The plate notation represents the variables that repeat in the model where the number of repetitions is given by the number on the bottom-right corner of the corresponding rectangle enclosing these variables.
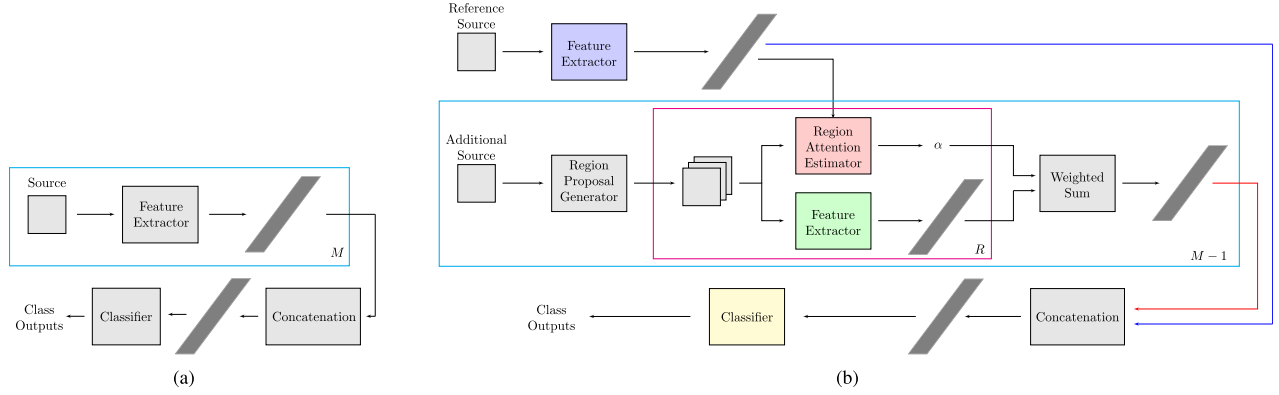
Fig. 2. (a) Basic multisource model. The feature representations independently obtained from each source are concatenated as the object representation. (b) Proposed MRAN. One source is chosen as the reference with verified selection of its object locations according to the ground truth. The remaining sources with the possibility of imprecise alignment of object locations are considered as additional sources. For each such source x^m, region proposals (candidate patches) $\{x_1^m, \ldots, x_R^m\}$ are generated, feature representations $\phi_m^{\text{reg}}(x_r^m)$ are obtained, and the attention scores (weights) $\alpha_m(x_r^m)$ are computed with the help of the feature representation of the reference source. The final representation for each additional source is the weighted sum $\phi_m^{\text{att}}(x^m)$ of its proposal regions' representations, and the final representation $\phi^{\text{att}}(x)$ used for class prediction is obtained by concatenation. Even though alternative weighting schemes such as filter weighting [30] are possible, we focus on the weighting of the region proposals in this paper.

that learns to *attend* regions of source images such that the resulting multisource representation is most informative for recognition purposes. We refer to this approach as MRAN.

In our approach, we presume that there is (at least) one source for which feature extraction is reliable, i.e., the representation for this source is not dominated by noise and/or background information. We refer to this source as the *reference*. Since the reference source typically has a higher spatial resolution, it is often possible to annotate objects in it with high spatial fidelity, either by georegistering it with some form of ground truth or by visually inspecting the images.

Our goal is to enhance recognition by leveraging additional sources. While we presume that all images within a tuple contain the same object instance, we do not expect a precise spatial alignment among them, i.e., the exact position of an object is locally unknown in the sources other than the reference. In addition, the images of additional sources may potentially contain other object instances belonging to different classes. In this realistic setting, therefore, extracting features independently at each source is likely to perform poorly.

In our approach, we aim to overcome these difficulties via a deep network including a conditional attention mechanism that selectively assigns importance scores to regions in each one of the $M-1$ sources, i.e., in those other than the reference. Without loss of generality, we assume that the reference source is the very first one, and there are R candidate regions in each one of the other source images, denoted by $x_1^m, \ldots, x_R^m$ where $m \in \{2, \ldots, M\}$. In our experiments, we obtain these candidate regions (proposals) by regularly sampling overlapping patches of fixed size within a larger neighborhood around the expected position of the object obtained by a simple transformation from the reference source (see Section IV).

To formalize the proposed conditional attention mechanism, we define the *region attention estimator* $\omega_m(x_r^m, \phi_1(x^1))$, which takes the rth candidate region from the mth source and the feature representation of the corresponding reference image, and maps to a nonnegative attention score. The attention score represents the confidence that the region contains (a part of) the object of interest.

We then leverage these scores to obtain a multisource representation that focuses on the regions containing the object of interest within each source. For this purpose, we define the *attention-driven source representation* $\phi_m^{\text{att}}(x^m), m = 2, \ldots, M$, as a weighted sum of per-region representations

$$\phi_m^{\text{att}}(x^m) = \sum_{r=1}^{R} \alpha_m(x_r^m) \phi_m^{\text{reg}}(x_r^m) \qquad (2)$$

where ϕ_m^{reg} is the region-level feature extractor, and the weighting term α_m is the normalized attention score of the region

$$\alpha_m(x_r^m) = \frac{\omega_m(x_r^m, \phi_1(x^1))}{\sum_{r'=1}^{R} \omega_m(x_{r'}^m, \phi_1(x^1))}. \qquad (3)$$

Our final *attention-driven multisource representation* is obtained by concatenation of attention-driven source representations

$$\phi^{\text{att}}(x) = \left[\phi_1(x^1)^\top, \phi_2^{\text{att}}(x^2)^\top, \ldots, \phi_M^{\text{att}}(x^M)^\top\right]^\top. \qquad (4)$$

The resulting attention-driven multisource representation can be fed to a classifier to recognize the object of interest

$$C(x) : \phi^{\text{att}}(x) \to \mathbb{R}_{\geq 0}^K \qquad (5)$$

where C is the classifier function that outputs a confidence score for each of the K classes. An illustration of our MRAN framework can be found in Fig. 2(b).

In Section III-D, we present the proposed deep architecture that implements the complete MRAN model by realizing the region-level feature extractors $\phi_m, m = 1, \ldots, M$, the per-source conditional attention estimators $\omega_m, m = 2, \ldots, M$, and the classifier C by using deep neural networks, and explain how we jointly learn these networks in an end-to-end fashion.

D. MRAN Architecture Details

We utilize a deep neural network in order to realize our MRAN framework. Our goal here is to: 1) jointly process spatial and spectral information in the source images; 2) implement an effective conditional region attention mechanism;
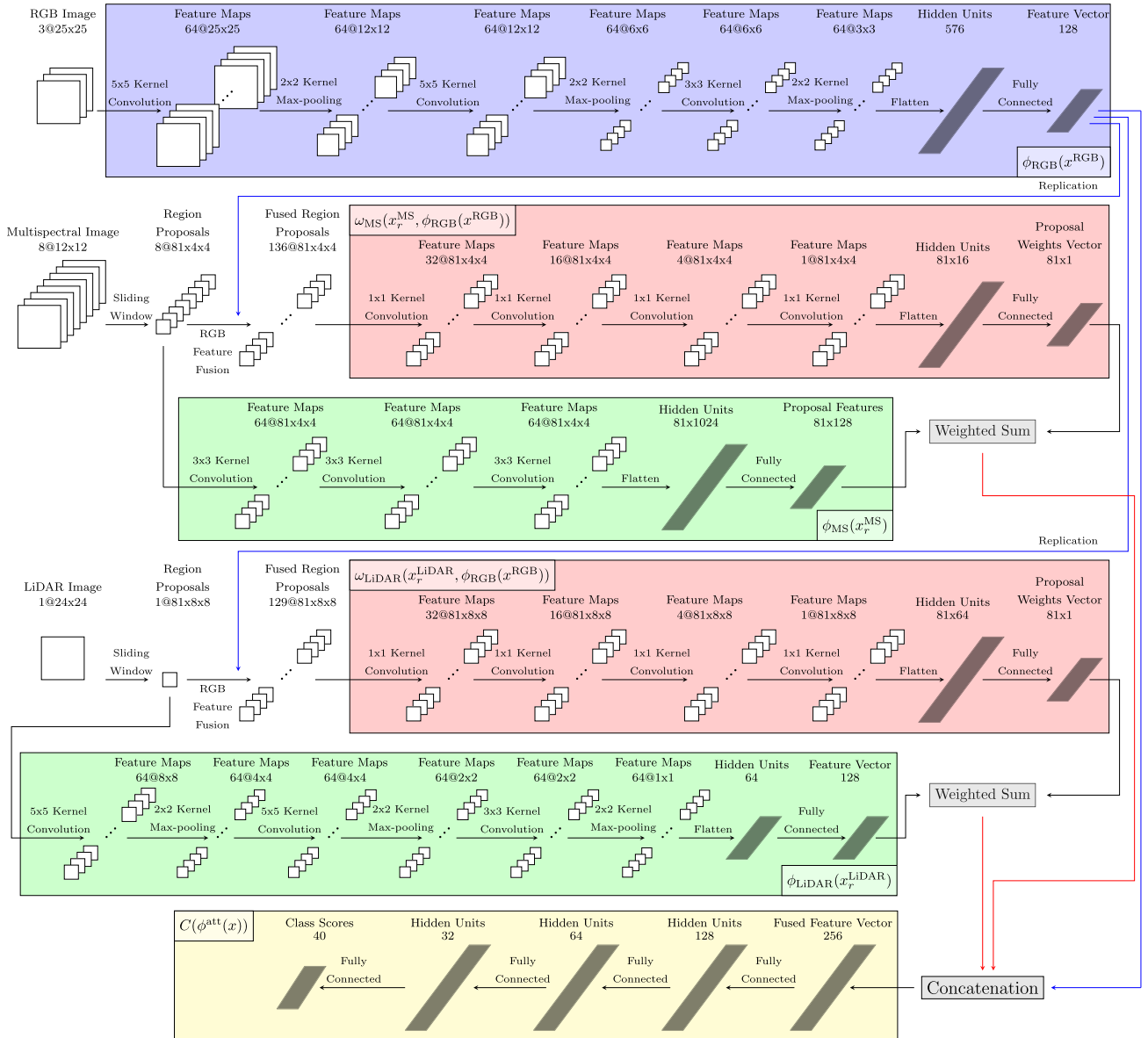
Fig. 3. Realization of the proposed MRAN architecture by using three source domains. The colors of the branches correspond to the boxes in Fig. 2(b). The first branch ϕ_{RGB} acts as the feature extractor for the reference domain, the aerial RGB image x^{RGB}. It contains three convolutional layers containing 64 filters with sizes 5×5, 5×5, and 3×3, respectively, followed by an FC layer containing 128 neurons. We apply max-pooling after each convolutional layer. The second branch ω_{MS} estimates the attention scores for the region proposals x_r^{MS} of the MS image, with the help of the feature representation of the RGB image. This branch contains four convolutional layers with 32, 16, 4, and 1 filters each with size 1×1, followed by an FC layer containing 16 and 1 neurons, respectively. The third branch ϕ_{MS} computes the feature representations of the MS region proposals and includes three convolutional layers each containing 64 filters with size 3×3, followed by an FC layer containing 128 neurons. Similarly, the fourth branch ω_{LiDAR} estimates the attention scores for the LiDAR data, and the fifth branch ϕ_{LiDAR} computes the LiDAR feature representations. The last branch C calculates the class scores from the concatenation of the feature representations of all three sources ϕ^{att}. It consists of four FC layers containing 128, 64, 32, and 40 neurons, the last one giving the class scores. Note that the feature map sizes and descriptive names are stated at the top of each layer.

and 3) learn the whole recognition pipeline in an end-to-end fashion. While the proposed architecture can easily be adapted to various combinations of sources, we assume that the reference source is RGB imagery, and the additional sources are obtained using MS and LiDAR sensors in this presentation.

For this purpose, we define an architecture that is formed by the combination of five deep CNN branches and a block of fully connected (FC) layers as shown in Fig. 3. The first branch extracts the image feature representation of the reference RGB data. We adopt this architecture from our previous work [3]. The second and fourth branches take the region proposals of the images from the additional sources and append the

feature vector of the reference source to the end of each pixels' input channels via replication. Four convolutional layers with 1×1 dimensional filters and an FC layer estimates the attention scores of region proposals. The third branch in which the feature representation is computed for each region proposal in the MS data differs from the first branch for the RGB data by using smaller filters and not using max-pooling because of the difference in spatial resolution. The fifth branch is the feature extractor for the LiDAR data and is similar to the first branch. Finally, the concatenation of attention-driven source representations and the reference source representation goes to the last branch in which four FC layers implement the classifier

that gives the final class scores. Stride for all convolutional layers is set at 1 to prevent information loss. We use zero-padding to avoid reduction in the spatial dimensions over convolutional layers.

The number of filters for each convolutional layer in the first, third, and fifth branches is selected as 64 in order to find a balance between model capacity and preventing overfitting. However, for the attention score estimator branches, we prefer to use decreasing number of filters from 32 to 1 in order to have correct number of scores at the end. Finally, although we have experimented with deeper and wider models, we reached the best performance with the presented network.

The particular instantiation of the network in Fig. 3 uses the RGB data as the reference and MS and LiDAR data as the additional sources. Note that the *region attention estimator* branches (red boxes) are the same for all sources, and the *feature extractor* branches (blue and green boxes) differ only slightly in terms of the number of layers according to the spatial resolutions of the sources and the sizes of the region proposals. The design in Section III-C and the abstraction in Fig. 2(b) are generic so that any number of reference and additional sources with any spatial and spectral resolution can be handled in the proposed framework by selecting an appropriate feature extractor model for each source.

Training the model is carried out over the classes by employing the cross-entropy loss, corresponding to the maximization of label log-likelihood in the training set. For enhancement of training, we benefited from dropout regularization and batch normalization. Additional training details and a comparison of our model are provided in Section IV.

IV. EXPERIMENTS

In this section, we present the experimental setup, results when the sources are used individually and in different combinations, and comparisons with the baseline approach.

A. Experimental Setup

We follow the same class split in [3] and evaluate our method using both 18 and 40 classes. For both cases, we split images from all sources into *train* (60%), *validation* (20%), and *test* (20%) subsets. Based on our previous observations, we add perturbations to training images by shifting each one randomly with an amount ranging from zero to 20% of height/width.

For all experiments, training is carried out on the train set by using stochastic gradient descent with the Adam method [31] where the hyperparameters are tuned on the validation set. All network parameters are initialized randomly and are learned in an end-to-end fashion. The initial learning rate of Adam, minibatch size, and ℓ_2-regularization weight are set to 10^{-3}, 100, and 10^{-5}, respectively, as in [3].

We use normalized accuracy as the performance metric where the per-class accuracy ratios are averaged to avoid biases toward classes with a large number of examples.

B. Single-Source Fine-Grained Classification

Our initial experiments consist of evaluating the performance of each source individually. For this, we use the first,

TABLE I
SINGLE-SOURCE FINE-GRAINED CLASSIFICATION RESULTS (IN %)

	Random guess	LiDAR 8×8 patches	LiDAR 24×24 patches	RGB 25×25 patches	MS 4×4 patches	MS 12×12 patches
18 classes	5.6	12.1	25.6	34.6	39.0	**47.7**
40 classes	2.5	7.8	18.1	23.9	25.1	**34.6**

third, and fifth branches of the network in Fig. 3 for RGB, MS, and LiDAR data, respectively. We add one more FC layer that maps the output of the last layer in each branch to class scores to obtain three separate CNNs for single-source classification. CNN for RGB data always takes 25×25 pixel patches as input. CNN for MS data takes both 4×4 and 12×12 pixel patches (corresponding to green and white boxes in Fig. 1, respectively) as input in two separate experiments. CNN for LiDAR data is also given 8×8 and 24×24 pixel patches. Each CNN operates on the whole single patch as there is no region proposal in this setup.

Performances of different sources are summarized in Table I. Results show that all settings are clearly better than the random guess baseline (choosing one of the classes randomly with an equal probability). Similar trends are observed for 18-class and 40-class classification, though the latter proved to be a more difficult problem as expected. We also see that the height and limited structure information in LiDAR data cannot cope with the spectral information in the other sources while the MS data outperform all the others. Even though MS has one-sixth of the spatial resolution of the RGB data, its rich spectral content proves to be the most informative for fine-grained classification of trees. We also see that using larger patches results in higher accuracies. Although 4×4 patches for MS and 8×8 patches for LiDAR perfectly coincide with the point-based ground truth locations that were verified with respect to the RGB data, the samples for which these patches could not include most of the target trees due to alignment problems have better predictions when the extended context in larger patches is used. However, using such patches has a risk of including irrelevant details in the feature representation. We show that finding the correct patch with the correct size in the surrounding neighborhood significantly improves the accuracies in Section IV-C.

C. Multisource Fine-Grained Classification

In this section, we evaluate the proposed framework (Section III-C) against the basic multisource model that uses simple feature concatenation (Section III-B). During the end-to-end training of the proposed model, first, the network in Fig. 3 accepts images from all sources and produces class scores, and then, backpropagation is carried out with respect to the calculated loss from class labels. The goal is to simultaneously learn both the spatial distribution of object regions and the mapping from multiple sources to class probabilities.

To identify the locations that are likely to contain an object of interest, our model evaluates sliding windows of region proposals within larger neighborhoods. We experimented with
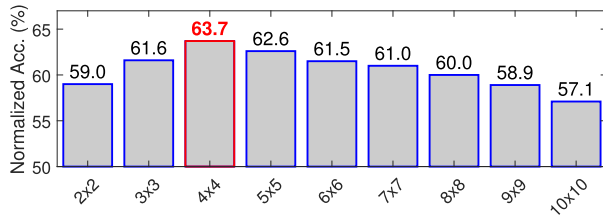
Fig. 4. Effect of region proposal size on classification performance. The *y*-axis shows the normalized accuracy when the RGB data are used with the MS data as the additional source. The *x*-axis shows different sized region proposals within a 12×12 neighborhood.

different region proposal and neighborhood sizes. Fig. 4 shows the details of these experiments when the RGB data and MS data are used together (first, second, third, and sixth branches in Fig. 3). Different sized region proposals were searched within a 12×12 pixel neighborhood in the MS data, and 4×4 pixel regions achieved the best performance. This size is also the matching spatial dimension of objects when the spatial resolution of the RGB data is considered. Note that different sizes lead to different number of region proposals. For the particular case of 4×4 regions within 12×12 neighborhoods, we obtain 81 proposals with a stride of 1 pixel. In the rest of this section, we present the multisource classification results when 4×4 regions in 12×12 neighborhoods are used for the MS data and 8×8 regions in 24×24 neighborhoods are used (with a stride of 2 pixels to similarly obtain 81 proposals) for the LiDAR data in the proposed framework. Figs. 1 and 3 also illustrate this particular setting.

Table II summarizes the results for multisource classification for both 18-class and 40-class settings. We used two versions of the basic multisource model in Fig. 2(a). The version named *basic CNN model* uses the first, third, and fifth branches in Fig. 3 as the feature extractor networks, concatenates the resulting feature representations, and uses an FC layer as the classifier. This model is also learned in an end-to-end fashion. The version named *recurrent attention model* uses a network that learns discriminative region selection and region-based feature representation at multiple scales [32]. It uses a CNN for feature extraction at each scale, and an attention proposal network between two scales predicts the bounding box of the region given as input to the next scale. The network is trained by a multitask loss: an intrascale classification loss that optimizes the convolution layers, and an interscale pairwise ranking loss that optimizes the proposal network. The final multiscale representation is constructed by concatenating the output of a specific FC layer at each scale. A two-scale network is observed to perform better than a three-scale one in our experiments. We use the two-scale architecture to train a feature extractor for each source, concatenate the resulting feature representations, and train an FC layer as the classifier. For both versions, the best performing single-source settings in Table I (12×12 patches for MS and 24×24 patches for LiDAR) are used as the inputs to the model.

As seen in Table II, all settings performed significantly better than the random baseline and the single-source settings. This shows the significance of using multisource data in the challenging fine-grained classification problem. When

TABLE II
MULTISOURCE FINE-GRAINED CLASSIFICATION RESULTS (IN %)

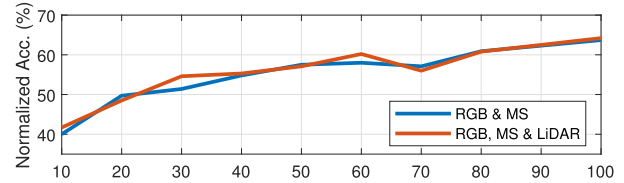|	*18 classes*	*40 classes*
Random guess	5.6	2.5
Basic CNN model (RGB & MS)	55.5	39.1
Basic CNN model (RGB, MS & LiDAR)	56.8	41.4
Recurrent attention model (RGB & MS) [32]	58.7	41.6
Recurrent attention model (RGB, MS & LiDAR) [32]	58.2	42.6
Proposed framework (RGB & MS)	63.7	46.6
Proposed framework (RGB, MS & LiDAR)	**64.2**	**47.3**

Fig. 5. Effect of amount of training data on classification performance. The *y*-axis shows the normalized accuracy and the *x*-axis shows the percentage of the data used during training.

we compare the two additional sources, MS and LiDAR, the contribution of the MS data in the overall accuracy is more significant as the rich spectral content appears to be more useful for discriminating the highly similar fine-grained categories. Overall, we observe that the proposed framework that simultaneously learns the feature extraction, attention, and classification networks performs significantly better than the commonly used basic multisource model with direct feature concatenation. Considering the observation that using larger patches gives higher accuracies for single-source classification in Table I, the difference between the accuracy of the basic model (e.g., 56.8% for 18 classes and 41.4% for 40 classes for RGB, MS, and LiDAR sources) that uses the larger patches and the proposed one (e.g., 64.2% for 18 classes and 47.3% for 40 classes) that assigns object localization confidence scores to smaller sized region proposals via the region attention estimators and uses the resulting attention-driven multisource feature representations confirms the importance of learning both the alignment and the classification models.

We also performed controlled experiments to analyze the effect of the amount of training data on classification performance. Fig. 5 shows the resulting accuracies for the 18-class setting when the amount of training data is reduced from 100% to 10% with 10% decrements. We observe that the accuracy achieved by the proposed framework using 50% of the training data is still higher than that of the basic multisource model using 100% of the training data. The proposed framework using only 20% of the training data also performed better than the best single-source model with 100% of the training data.

When the confusion matrices are considered, we observed that most confusions are among the trees that belong to the same families in higher levels of the scientific taxonomy (given in [3]). For example, among 40 classes, 30% of thundercloud plum samples are wrongly predicted as cherry plum, and 16% are wrongly predicted as blireiana purpleleaf plum.

TABLE III
ZSL RESULTS (IN %)

Image representation	Normalized accuracy
Random guess	6.3
LiDAR 8×8 patches	8.0
LiDAR 24×24 patches	12.1
RGB 25×25 patches [3]	14.3
MS 4×4 patches	15.2
MS 12×12 patches	16.7
Basic CNN model (RGB & MS)	15.8
Basic CNN model (RGB, MS & LiDAR)	17.4
Proposed framework (RGB & MS)	**17.7**
Proposed framework (RGB, MS & LiDAR)	17.0

Similarly, 31% of cherry plum samples are wrongly predicted as thundercloud plum. As other examples for the cases with the highest confusion, 21% of double Chinese cherry are wrongly predicted as Kwanzan flowering cherry and 13% are wrongly predicted as autumn flowering cherry, 21% of common hawthorn are wrongly predicted as English midland hawthorn, 23% of red maple are wrongly predicted as sunset red maple, and 18% of scarlet oak are wrongly predicted as red oak. Since these trees are only distinguished with respect to their subspecies level in the taxonomy and they have almost the same visual appearance, differentiating them even from ground-view images with a high accuracy is too difficult.

D. Fine-Grained Zero-Shot Learning

We also evaluate the proposed approach in the ZSL scenario [3] where new unseen classes are classified using a model that is learned by using an independent set of seen classes. Thus, no samples for the target classes of interest exist in the training data. We follow the same methodology as in [3] except the way how image embeddings are obtained. In place of the feature representation that is obtained from a single CNN that is trained on RGB data in [3], the multisource image embedding in this paper is obtained from the output of the first FC layer in the classifier (last) branch of the network in Fig. 3. We also evaluate the performances of using representations similarly obtained from the networks trained for the basic multisource model and the individual single-source models. The class split (18 training, 6 validation, and 16 test) and the rest of the experimental setup are the same as in [3].

Comparison of different representations evaluated using the 16 ZSL-test classes is shown in Table III. (Additional comparisons with other ZSL models can be found in [3].) When the single-source results are considered, we observe the same trend as in Table I where MS-based representation performed better than LiDAR-based and RGB-based representations, and using larger patches had higher accuracies than smaller patches that have potential alignment problems and limited spatial context. Regarding the multisource results, the best performance of 17.7% was obtained by the proposed model trained using RGB and MS data. The more complex models that use all three sources had slightly lower performances, probably because of the difficulty of learning in the extremely challenging ZSL scenario with very limited number of training samples.

Fig. 6.   Attention scores for example images. For each sample, the 25×25 RGB, 12×12 MS, and 24×24 LiDAR patches, as well as the scores for 4×4 MS and 8×8 LiDAR region proposals within these patches are shown.

Overall, together with the supervised classification results in the previous sections, the ZSL performance presented here highlights the efficacy of our proposed multisource framework.

E. Qualitative Evaluation

The quantitative evaluation results presented above highlight the remarkable performance by the proposed multisource approach compared to the baseline feature-level fusion model commonly used in remote sensing multisource image analysis. Fig. 6 provides qualitative results to investigate how well the model solves the alignment problem by learning to generate meaningful attention scores for the region proposals. These examples show that our model is capable of estimating the correct alignment of images obtained from different sources with imprecise registration while correctly classifying them.

V. CONCLUSION

We studied the fine-grained object recognition problem in multisource imagery, potentially having imprecise alignment with each other and with the ground truth. In order to deal with the complexity of learning many subcategories having subtle differences by using multiple image sources with different spatial and spectral resolutions and with misregistration errors, we proposed a framework that assigns attention scores to local regions sampled around the expected location of an object by comparing their content with the features of the reference source that is assumed to be more reliable with respect to the ground truth, computes a multisource feature representation as the concatenation of attention-weighted feature vectors of the local regions, and classifies the objects using a deep network that learns all of these components in an end-to-end fashion. Experiments using RGB, MS, and LiDAR data showed that our approach achieved 64.2% and 47.3% accuracies for the 18-class and 40-class settings, respectively, when all data sources were used, which correspond to 13% and 14.3% improvement relative to the commonly used feature concatenation approach from multiple sources. Future work includes evaluation of the model in other domains and solving other multisource classification problems in addition to alignment.

REFERENCES

[1] Q. Oliveau and H. Sahbi, "Learning attribute representations for remote sensing ship category classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2830–2840, Jun. 2017.

[2] S. Branson, J. D. Wegner, D. Hall, N. Lang, K. Schindler, and P. Perona, "From Google Maps to a fine-grained catalog of street trees," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 13–30, Jan. 2018.

[3] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 770–779, Feb. 2018.

[4] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[5] M. Datcu, F. Melgani, A. Piardi, and S. B. Serpico, "Multisource data classification with dependence trees," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 3, pp. 609–617, Mar. 2002.

[6] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. L. Rojo-Álvarez, and M. Martínez-Ramón, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.

[7] A. Voisin, V. A. Krylov, G. Moser, S. B. Serpico, and J. Zerubia, "Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3346–3358, Jun. 2014.

[8] Y. Zhang, H. L. Yang, S. Prasad, E. Pasolli, J. Jung, and M. Crawford, "Ensemble multiple kernel active learning for classification of multi-source remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 2, pp. 845–858, Feb. 2015.

[9] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7708–7720, Dec. 2014.

[10] G. Gao and Y. Gu, "Tensorized principal component alignment: A unified framework for multimodal high-resolution images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 46–61, Jan. 2019.

[11] C. Debes *et al.*, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[12] L. Wenzhi *et al.*, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.

[13] M. Campos-Taberner *et al.*, "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest—Part A: 2-D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5547–5559, Dec. 2016.

[14] N. Yokoya *et al.*, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.

[15] S. Morchhale, V. P. Pauca, R. J. Plemmons, and T. C. Torgersen, "Classification of pixel-level fused hyperspectral and lidar data using deep convolutional neural networks," in *Proc. 8th Workshop Hyperspectral Image Signal Process. (WHISPERS)*, Aug. 2016, pp. 1–5.

[16] L. Pibre, M. Chaumon, G. Subsol, D. Lenco, and M. Derras, "How to deal with multi-source data for tree detection based on deep learning," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2017, pp. 1150–1154.

[17] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.

[18] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, 2017, pp. 1–4.

[19] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[20] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1561–1570.

[21] J. Lee, X. Cai, C.-B. Schönlieb, and D. A. Coomes, "Nonparametric image registration of airborne LiDAR, hyperspectral and photographic imagery of wooded landscapes," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6073–6084, Nov. 2015.

[22] D. Marcos, R. Hamid, and D. Tuia, "Geospatial correspondences for multimodal registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5091–5100.

[23] Y. Han, F. Bovolo, and L. Bruzzone, "Edge-based registration-noise estimation in VHR multitemporal and multisensor images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1231–1235, Sep. 2016.

[24] L. Liu, N. C. Coops, N. W. Aven, and Y. Pang, "Mapping urban tree species using integrated airborne hyperspectral and LiDAR remote sensing data," *Remote Sens. Environ.*, vol. 200, pp. 170–182, Oct. 2017.

[25] F. E. Fassnacht *et al.*, "Review of studies on tree species classification from remotely sensed data," *Remote Sens. Environ.*, vol. 186, pp. 64–87, Dec. 2016.

[26] M. Alonzo, B. Bookhagen, and D. A. Roberts, "Urban tree species mapping using hyperspectral and lidar data fusion," *Remote Sens. Environ.*, vol. 148, pp. 70–83, May 2014.

[27] M. Voss and R. Sugumaran, "Seasonal effect on tree species classification in an urban environment using hyperspectral data, LiDAR, and an object-oriented approach," *Sensors*, vol. 8, no. 5, pp. 3020–3036, May 2008.

[28] M. Dalponte, L. Bruzzone, and D. Gianelle, "Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data," *Remote Sens. Environ.*, vol. 123, pp. 258–270, Aug. 2012.

[29] L. Naidoo, M. A. Cho, R. Mathieu, and G. Asner, "Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment," *ISPRS J. Photogramm. Remote Sens.*, vol. 69, pp. 167–179, Apr. 2012.

[30] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[31] D. P. Kingma and J. Ba. (Dec. 2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[32] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4476–4484.

Gencer Sumbul (S'18) received the B.S. and M.S. degrees in computer engineering from Bilkent University, Ankara, Turkey, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Faculty of Electrical Engineering and Computer Science, Technical University of Berlin, Berlin, Germany.

His research interests include computer vision and machine learning, with special interest in deep learning and remote sensing.

Ramazan Gokberk Cinbis received the B.S. degree from Bilkent University, Ankara, Turkey, in 2008, the M.A. degree from Boston University, Boston, MA, USA, in 2010, and the Ph.D. degree from INRIA, Grenoble, France, in 2014.

He is currently an Assistant Professor with METU, Ankara. His research interests include machine learning and computer vision, with special interest on deep learning with incomplete supervision.

Selim Aksoy (S'96–M'01–SM'11) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1996, and the M.S. and Ph.D. degrees from the University of Washington, Seattle, WA, USA, in 1998 and 2001, respectively.

Since 2004, he has been with the Department of Computer Engineering, Bilkent University, Ankara, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning with applications to remote sensing and medical imaging.