Weakly Supervised Instance Attention for Multisource Fine-Grained Object Recognition with an Application to Tree Species Classification

Bulut Aygunes[a], Ramazan Gokberk Cinbis[b], Selim Aksoy[a,*]

[a]*Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey*
[b]*Department of Computer Engineering, Middle East Technical University, Ankara, 06800, Turkey*

Abstract

Multisource image analysis that leverages complementary spectral, spatial, and structural information benefits fine-grained object recognition that aims to classify an object into one of many similar subcategories. However, for multisource tasks that involve relatively small objects, even the smallest registration errors can introduce high uncertainty in the classification process. We approach this problem from a weakly supervised learning perspective in which the input images correspond to larger neighborhoods around the expected object locations where an object with a given class label is present in the neighborhood without any knowledge of its exact location. The proposed method uses a single-source deep instance attention model with parallel branches for joint localization and classification of objects, and extends this model into a multisource setting where a reference source that is assumed to have no location uncertainty is used to aid the fusion of multiple sources in four different levels: probability level, logit level, feature level, and pixel level. We show that all levels of fusion provide higher accuracies compared to the state-of-the-art, with the best performing method of feature-level fusion resulting in 53% accuracy for the recognition of 40 different types of trees, corresponding to an improvement of 5.7% over the best performing baseline when RGB, multispectral, and LiDAR data are used. We also provide an in-depth comparison by evaluating each model at various parameter complexity settings, where the increased model capacity results in a further improvement of 6.3% over the default capacity setting.

Keywords: Multisource classification, fine-grained object recognition, weakly supervised learning, deep learning

1. Introduction

Advancements in sensors used for remote sensing enabled spectrally rich images to be acquired at very high spatial resolution. Fine-grained object recognition, which aims the classification of an object as one of many similar subcategories, is a difficult problem manifested by these improvements in sensor technology (Oliveau and Sahbi, 2017; Branson et al., 2018; Sumbul et al., 2018). The difficulty of distinguishing subcategories due to low variance between classes is one of the main characteristics of this problem that differs from traditional object recognition and classification tasks studied in the remote sensing literature. Other distinguishing features of fine-grained object recognition are the difficulty of collecting samples for a large number of similar categories, which can cause the training set sizes to be very limited for some classes, and the class imbalance that makes the traditional supervised learning approaches to overfit to the classes with more samples. This makes it necessary to develop new methods for fine-grained classification that could cover the shortfalls of the traditional object recognition methods regarding these problems.

One way to help decrease the confusion inherent to the data in fine-grained classification is to gather complementary information by utilizing multiple data sources, which can provide more distinguishing properties of the object of interest. For example, a high-resolution RGB image can give details about texture, color, and coarse shape, whereas a multispectral (MS) image can provide richer spectral content and LiDAR data can yield information about the object height. However, the question of how to combine the data from multiple sources does not have a straightforward answer. Therefore, it is an open research problem to find a method to benefit from the distinct contents of the sources as effectively as possible.

The common assumption of most multisource image analysis methods is that the data sources are georeferenced or coregistered without any notable errors that may prevent the pixel- or feature-level fusion of the sources. This can be a valid assumption for tasks like land cover classification in which the classes of interest (e.g., water, forest, impervious surfaces) are significantly larger compared to the registration errors (Chen et al., 2017). However, for multisource tasks that involve the classification of relatively small objects such as trees—similar to the problem we focus on in this paper—even the smallest registration errors can introduce high uncertainty among different sources and between the sources and the ground truth labels. Furthermore, it is not always possible to mitigate this uncertainty by trying to discover pixel-level correspondences between the sources due to the fine-grained nature of the problem and for other reasons such as differences in the imaging conditions, viewing geometry, topographic effects, and geometric

*Corresponding author. Tel: +90 312 2903405; fax: +90 312 2664047.
Email addresses: bulut.aygunes@bilkent.edu.tr (Bulut Aygunes), gcinbis@ceng.metu.edu.tr (Ramazan Gokberk Cinbis), saksoy@cs.bilkent.edu.tr (Selim Aksoy)

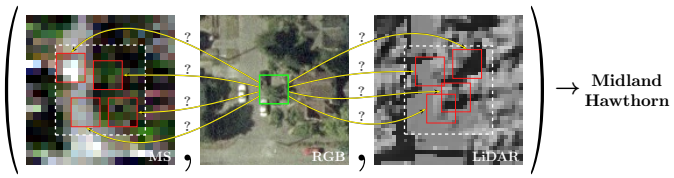

Figure 1: Illustration of our multisource fine-grained object recognition problem. The sources are only approximately registered, therefore, it is unclear which pixels in the low-resolution sources (MS and LiDAR) correspond to the object of interest centered in the high-resolution reference source (RGB). Our goal is to implicitly tackle the registration uncertainties through instance attention to correctly predict the object class using information from all sources.

distortions (Han et al., 2016).

The fine-grained recognition problem studied in this paper involves the classification of street trees using RGB, MS, and LiDAR data. Although the very high-resolution RGB images are manually inspected with respect to the reference tree locations in the GIS-based ground truth, there is still high location uncertainty in the MS and LiDAR data that contain trees of 4×4 and 8×8 pixels, respectively, due to the aforementioned reasons. To cope with this uncertainty introduced by registration errors and small sizes of the target objects, we crop tiles larger than the object sizes around the reference locations given by the point GIS data to ensure that each tree falls inside its corresponding tile. With such images covering larger neighborhoods than the typical object size, the problem becomes a weakly supervised learning (WSL) problem in the sense that the label of each image provides information about the category of the object it contains, but does not yield any information regarding its location in the neighborhood. The problem is illustrated in Figure 1.

To the best of our knowledge, the only related work that studied the multisource fine-grained recognition problem is that of Sumbul et al. (2019), where an attention mechanism over a set of candidate regions was used with guidance by a reference source to obtain a more effective representation of the sources which are fused together for the final classification. However, in that scheme, the attention mechanism simply aims to maximize the discriminative power of the attention-driven representation. While that approach yields promising results, it is susceptible to overfitting to *accidental* correlations appearing in training examples, and may learn to put too much emphasis on background features. In this work, we instead utilize a stronger WSL-based formulation that aims to induce an *instance attention* behavior: instead of estimating pooling weights of candidate regions, we estimate the relevance of each candidate region as a function of its spatial *and* semantic distinctiveness. Here, therefore, we aim to incorporate the prior knowledge that in most cases one (or very few) local regions actually belong to the object of interest in a local neighborhood.

The method proposed in this paper loosely builds upon our preliminary work (Aygüneş et al., 2019), which has shown that weakly supervised learning objective can be repurposed to improve single-source object recognition when the images contain high location uncertainty. In this paper, as our main contribution, we extend this idea to the multisource setting with a number of novel information fusion schemes. We first propose a more generalized version of our WSL-based instance attention model for single-source classification, which can be applied to any source with location uncertainty. Then, the proposed fusion schemes combine multiple additional sources that are processed with this instance attention method and a reference source that is assumed to have no uncertainty and is processed in a fully supervised fashion. Each proposed scheme aims to leverage the reference source to aid the instance attention branches by combining the reference with the additional sources in four different levels: probability level, logit level, feature level, and pixel level. We show that it is possible to benefit from the reference source with all levels of fusion, as they surpass the state-of-the-art baselines. As another contribution, we also propose a methodology to compare different models in a more principled way, by evaluating each model at various parameter complexity settings. The results of this experiment highlight the importance of investigating approaches at various model capacities to make fair comparisons, as comparing different methods each of which having a different single model capacity setting can be misleading. Overall, our results indicate that we obtain significant improvements over the state-of-the-art.

In the rest of the paper, we first present a summary of related work in Section 2 and give information about the data set we use in our experiments in Section 3. We then describe our proposed methods in Section 4. Next, we give details about our experimental setup, and present quantitative comparisons with several baselines and qualitative results in Section 5. Finally, we provide our conclusions in Section 6.

2. Related work

Multisource image analysis. There are many studies in the remote sensing literature that focus on multisource image analysis (Gomez-Chova et al., 2015; Dalla Mura et al., 2015), which has also received the attention of data fusion contests (Debes et al., 2014; Liao et al., 2015; Campos-Taberner et al., 2016; Yokoya et al., 2018). The research includes statistical learning methods such as dependence trees (Datcu et al., 2002), kernel-based methods (Camps-Valls et al., 2008), copula-based multivariate model (Voisin et al., 2014), and active learning (Zhang et al., 2015). Another well-studied problem is manifold alignment (Tuia et al., 2014; Hong et al., 2019; Gao and Gu, 2019) where the goal is to transfer knowledge learned in the source domain to a target domain. The underlying reason that necessitates this transfer is typically the spectral mismatch between the domains. In this paper, the main problem in the multisource analysis is the spatial mismatch among the image sources.

More recently, deep learning-based methods have focused on classification with pixel-level or feature-level fusion of multiple sources. Pixel-level fusion includes concatenation of hyperspectral and LiDAR data preprocessed to the same resolution, followed by a convolutional neural network (CNN) (Morchhale et al., 2016), while in feature-level fusion, hand-crafted (Ghamisi et al., 2017) or CNN-based (Pibre et al., 2017; Hu et al., 2017; Xu et al., 2018; Ienco et al., 2019) features, ob-

tained from different data sources such as multispectral, hyperspectral, LiDAR, or SAR, are processed with convolutional and/or fully-connected layers to obtain the final decision.

Weakly supervised remote sensing. WSL approaches in remote sensing have utilized class activation maps for object localization. For example, Ji et al. (2019) combined the per-class activation maps from different layers of a convolutional network trained with image-level labels to obtain class attention maps and localize the objects. Wang et al. (2020) proposed a modification to the U-Net architecture to enable using image-level weak labels corresponding to the majority vote of the pixel labels instead of pixel-level strong labels during the training for a binary segmentation task. Xu et al. (2019) localized objects by using a combination of two different convolutional layers. Zhang et al. (2019) suggested using gradients of network layers to obtain saliency maps for background and foreground classes. Similarly, Ma et al. (2020) obtained saliency maps by utilizing gradients with respect to the input pixels to localize residential areas in aerial images. Ali et al. (2020) studied destruction detection from only image-level labels where each image was represented using a weighted combination of the patch-level representations obtained from a convolutional network. The weights were learned by using a single fully-connected layer that was trained using a sparsity loss. Li et al. (2020a) introduced a global convolutional pooling layer to build a cloud detection network that was trained using block-level labels indicating only the presence or absence of clouds within image blocks without pixel-level annotations. However, all of these approaches focus on a binary classification scenario. For the multi-class setting, Li et al. (2018) used pairs of images with the same scene-level labels to train a Siamese-like network for learning convolutional weights, and updated this network with a global pooling operation and a fully-connected layer to learn class-specific activation weights. Hua et al. (2019) used a linear combination of all channels in the output of a CNN-based feature extractor network to learn class-specific feature representations that were further combined in a recurrent neural network to learn class dependencies. However, both of these approaches use a global combination of the convolutional channels where a single fully-connected layer is expected to learn the class attention mechanism. Furthermore, none of the approaches above considers the label uncertainty problem in a multisource setting. In our case, while we do not aim to explicitly localize objects as in WSL studies, we propose a number of WSL-based formulations for addressing the spatial ambiguity in multisource object recognition.

Another important problem is the noise in the type of the label where a scene or an object is labeled as another class instead of the true one. For example, Li et al. (2020b) proposed an error-tolerant deep learning approach for remote sensing image scene classification by iteratively training a set of CNN models that collectively partitioned the input data into a strong data set with all models agreeing on the original labels and a weak data set with the models producing different predictions. Each iteration built new CNN models that used the union of the strong data set with the original labels and the weak data set with the predicted labels as the new training data. Here, we assume that the label itself is correct but an uncertainty exists in its spatial location.

In a more relevant problem caused by misalignment of GIS maps and images used for building extraction, Zhang et al. (2020) added a layer to a segmentation network to model the noise in the labels, and trained the model by calculating the loss using the noisy predictions and the noisy reference labels. Although such an approach can be useful in a task like building extraction, it might not be applicable for problems consisting of small objects like trees where a segmentation-based approach is not feasible due to the size and fine-grained nature of the objects.

Tree classification. In this paper, we illustrate the proposed weakly supervised instance attention model and the multisource fusion schemes using a fine-grained street tree classification problem. Novel applications involving street trees include (Branson et al., 2018) where aerial images and street-view panoramas were jointly used for fine-grained classification. The feature representations computed by deep networks independently trained for the aerial and ground views were concatenated and fed to a linear SVM for classification of 40 tree species. More recently, Laumer et al. (2020) improved the existing street tree inventories where the individual trees were referenced by only street addresses with accurate geographic coordinates that were estimated from multi-view detections in street-view panoramas. The methods proposed in this paper are not specific to tree detection. Thus, a full review on tree species mapping is beyond the scope of this paper. We refer the reader to Fassnacht et al. (2016) that provides a review of such methods in which multispectral, hyperspectral, and LiDAR data sources have been the most widely used modalities.

3. Data set

We conduct our experiments on the same data set as (Sumbul et al., 2019), which is, to our knowledge, the only multisource data set that includes a fine-grained set of classes with an additional challenge of location uncertainty among the data sources due to the sizes of the objects of interest. The data set consists of a total of 48,063 instances of street trees belonging to 40 different classes. The fine-grained nature of the data set is illustrated in Figure 2 where the scientific classification of tree species is presented as a hierarchy in which most species differ only in the lowest levels. The number of samples for each class in this highly imbalanced data set is shown in Table 1.

For each tree sample, there are three images obtained from different sources: an aerial RGB image with 1 foot spatial resolution, an 8-band WorldView-2 MS image with 2 meter spatial resolution, and a LiDAR-based digital surface model with 3 foot spatial resolution. The label and location information for the tree samples were obtained from the point GIS data provided by the Seattle Department of Transportation in Washington State, USA (City of Seattle, Department of Transportation, 2016). Since the GIS data set was constructed as part of a carefully planned field campaign for inventory management, we assumed that the class label for each tree is correct. However, we

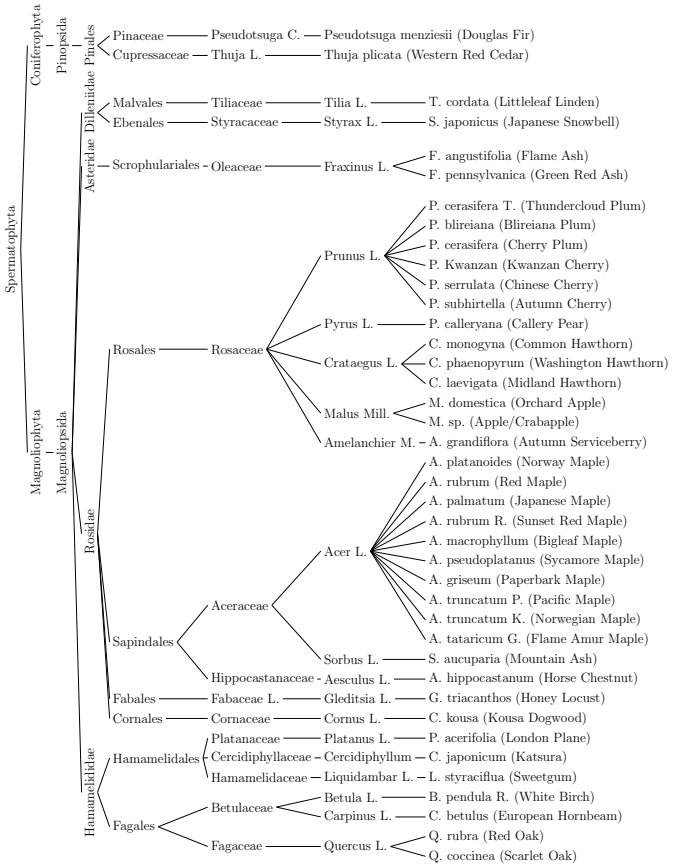

Figure 2: Scientific classification of tree species. The taxonomy starts with the Spermatophyta superdivision and continues with the names of division, class, subclass, order, family, genus, and species in order. At the last level, the common names are given in parentheses next to the scientific names. The classifications are taken from Natural Resources Conservation Service of the United States Department of Agriculture (2016).

used visual interpretation on the aerial RGB image, that corresponds to the data source with the highest spatial resolution, as an additional effort to validate the consistency of the tree locations (Sumbul et al., 2018). Even though it was not possible to visually confirm the tree category from the remotely sensed data, we made sure that the provided coordinate actually coincided with a tree for every single one of the samples. During this process, some samples had to be removed due to mismatches with the aerial data, probably because of temporal differences between ground data collection and aerial data acquisition.

For the confirmed 48,063 samples, RGB images that were centered at the locations provided in the point GIS data were cropped at a size of 25×25 pixels to cover the largest tree in the data set. The validation process for the tree locations, together with the fact that RGB images have higher spatial resolution than MS and LiDAR data, make RGB images suitable to be used as the reference source. This choice can be made with a high confidence if there is additional information regarding the georeferencing process of the data sources and their compatibility with the ground truth object locations.

A tree in a 25×25 pixel RGB image corresponds to 4×4 pixels in MS and 8×8 pixels in LiDAR data. Although each source was previously georeferenced, registration errors can cause significant uncertainties in the locations of small objects such as trees, especially in sources with lower resolution such as MS and LiDAR. To account for the location uncertainties, we use images that cover a larger neighborhood than a single tree. Specifically, we use 12×12 pixel neighborhoods for MS and 24×24 pixel neighborhoods for LiDAR (Sumbul et al., 2019).

Table 1: Class names and number of samples in each class in the data set. The classes follow the same order as in Figure 2.

Class name	Samples	Class name	Samples
Douglas Fir	620	Red Maple	2,790
Western Red Cedar	720	Japanese Maple	1,196
Littleleaf Linden	1,626	Sunset Red Maple	1,086
Japanese Snowbell	460	Bigleaf Maple	885
Flame Ash	679	Sycamore Maple	742
Green Red Ash	660	Paperbark Maple	467
Thundercloud Plum	2,430	Pacific Maple	716
Blireiana Plum	2,464	Norwegian Maple	372
Cherry Plum	2,510	Flame Amur Maple	242
Kwanzan Cherry	2,398	Mountain Ash	672
Chinese Cherry	1,531	Horse Chestnut	818
Autumn Cherry	621	Honey Locust	875
Callery Pear	892	Kousa Dogwood	642
Common Hawthorn	809	London Plane	1,477
Washington Hawthorn	503	Katsura	383
Midland Hawthorn	3,154	Sweetgum	2,435
Orchard Apple	583	White Birch	1,796
Apple/Crabapple	1,624	European Hornbeam	745
Autumn Serviceberry	552	Red Oak	1,429
Norway Maple	2,970	Scarlet Oak	489

4. Methodology

In this section, we first outline the weakly supervised multisource object recognition problem. Then, we explain the single-source weakly supervised instance attention approach. Finally, we present our four formulations to tackle the multisource recognition problem via instance attention.

4.1. Weakly supervised multisource object recognition

The multisource object recognition problem aims to classify an object into one of the C classes by utilizing the images of the object coming from M different sources. This corresponds to learning a classification function that takes the images $x_1, \ldots, x_M$ of an object from M imaging sources and outputs a class prediction $\hat{y} \in \{1, \ldots, C\}$.

To cope with the location uncertainty in the data, we use images that cover a larger area than the objects of interest as the input to the model. More precisely, we assume that each image from the m^{th} source covers an $N_m \times N_m$ pixel neighborhood and contains a smaller object of size $W_m \times W_m$ with an unknown location. In such a setting, the ground truth for an image becomes a weak label in the sense that it does not hold any positional information about the object, which makes the problem a weakly supervised learning problem.

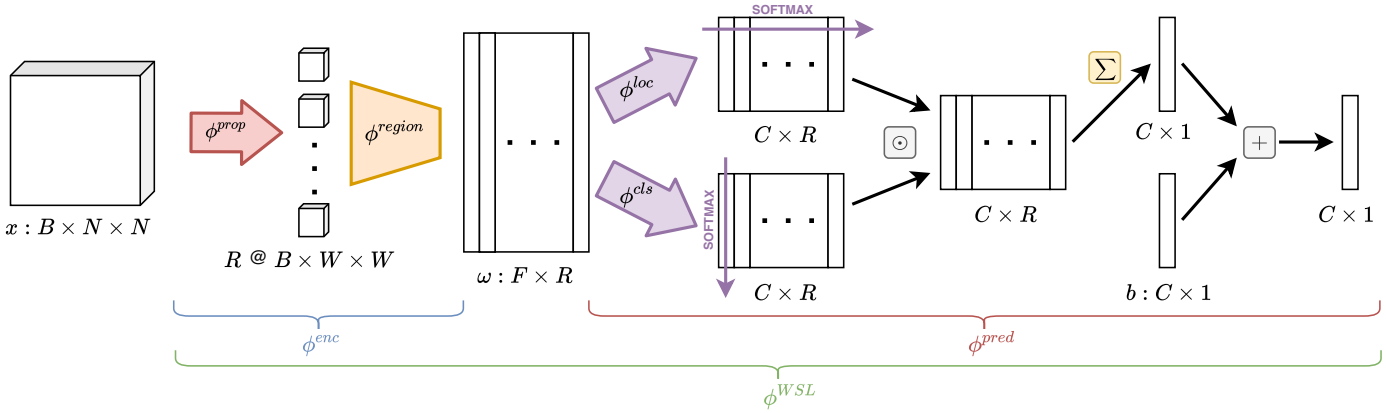

Figure 3: Illustration of the single-source weakly supervised instance attention model. Source index m is omitted from all variables for simplicity. For the m^{th} source, the model takes x_m with size $B_m \times N_m \times N_m$ as the input. Region proposals are extracted using ϕ^{prop} with a sliding window of size $W_m \times W_m$, resulting in $R_m = (N_m - W_m + 1)^2$ candidate regions. Each region is processed by ϕ^{region}, which consists of 3 convolutional and 1 fully-connected layers. For MS, all convolutional layers have 64 kernels of size 3×3 and no pooling is used afterwards. For LiDAR, 64 convolution kernels with size 5×5 for the first two layers and 3×3 for the last layer are used, with max-pooling with kernel size 2×2 and stride 2 after each convolutional layer. The fully-connected layer of ϕ^{region} outputs features of size 128. A fully-connected layer (ϕ^{loc}) and softmax across regions are applied to the resulting matrix of per-region feature vectors ω^m for the localization branch, and another fully-connected layer (ϕ^{cls}) followed by softmax across classes is applied for the classification branch. Hadamard product of the resultant matrices is taken and the result is summed over the regions. Finally, a bias vector b^m is added to obtain the logits. ReLU activation is used throughout the network.

In this work, we focus on RGB, MS, and LiDAR data which are acquired in different conditions (resolution, viewpoint, elevation, time of day, etc.). As a result, different registration uncertainties are present among the data sources, which cause the locations of the same object in the images from different sources to be independent of each other. This becomes one of the major challenges of the weakly supervised multisource object recognition problem.

4.2. *Single-source weakly supervised instance attention*

Location uncertainty in a WSL problem necessitates either explicit or implicit localization of the object to be classified. Successful localization of the object helps to obtain a more reliable representation by eliminating the background clutter, which in turn can improve the classification results. Following this intuition, we construct our instance attention approach by adapting the learning formulation of Weakly Supervised Deep Detection Network (WSDDN) (Bilen and Vedaldi, 2016). WSDDN extracts R candidate regions, some of which are likely to contain the object of interest, from an image x using a region proposal operator, ϕ^{prop}. Each of these regions is transformed into a feature vector of size F using a region encoder network, ϕ^{region}, consisting of three convolutional layers and one fully-connected layer. We refer the reader to the caption in Figure 3 for source-specific architectural details of the region encoder. For input $x \in X$,

$$\phi^{enc} : X \rightarrow \Omega \tag{1}$$

collectively represents candidate region extraction (ϕ^{prop}) and region encoding (ϕ^{region}) operations. Here, the resulting $\omega \in \Omega$ is an $F \times R$ matrix of per-region feature vectors. To simplify the notation, we define the remaining model components as a function of $\omega \in \Omega$.

After the region encoding operation, a localization branch scores candidate regions among themselves using softmax separately for each class, and outputs region localization scores:

$$[\sigma^{loc}(\omega)]_{ci} = \frac{\exp([\phi^{loc}(\omega)]_{ci})}{\sum_{r=1}^{R} \exp([\phi^{loc}(\omega)]_{cr})} \tag{2}$$

where $[\phi^{loc}(\omega)]_{ci}$ is the raw score of the i^{th} candidate region for the class c, obtained by the linear transformation ϕ^{loc}. Similarly, a parallel classification branch assigns region classification scores corresponding to the distribution of class predictions for each region independently:

$$[\sigma^{cls}(\omega)]_{ci} = \frac{\exp([\phi^{cls}(\omega)]_{ci})}{\sum_{k=1}^{C} \exp([\phi^{cls}(\omega)]_{ki})} \tag{3}$$

where $[\phi^{cls}(\omega)]_{ci}$ is the raw score obtained by the linear transformation ϕ^{cls}.

A candidate region that successfully localizes the true object is expected to yield both a higher localization score for the true object class than the other candidates and a higher classification score for the true class than the other classes. This property naturally yields a more instance-centric attention mechanism, compared to mainstream attention formulations that learn to weight candidate regions purely based on the discriminative power of the final attention-driven representation, see e.g. Sumbul et al. (2019). To implement this idea in a differentiable way, region localization and classification scores are element-wise multiplied and summed over all regions to obtain per-class localization scores on the image level:

$$[\phi^{pred}(\omega)]_c = \sum_{i=1}^{R} [\sigma^{loc}(\omega)]_{ci} \odot [\sigma^{cls}(\omega)]_{ci}. \tag{4}$$

The formulation up to this point is quite generic. To implement it in an efficient and effective way for our weakly supervised fine-grained classification task, we use the following

three ideas. First, we obtain candidate regions from the input image in a sliding window fashion with a fixed window of size $W_m \times W_m$, where W_m is experimentally chosen for each source m. Second, we put an additional softmax layer at the end of the network. This additional softmax layer effectively incorporates the prior knowledge that a local image area is likely to contain only a single class of interest. Finally, we add learnable per-class bias parameters to the class detection scores before the softmax operation, and update (4) as follows:

$$[\phi^{pred}(\omega)]_c = \sum_{i=1}^{R} [\sigma^{loc}(\omega)]_{ci} \odot [\sigma^{cls}(\omega)]_{ci} + b_c \tag{5}$$

where b_c is the bias parameter for class c. Combining (1) and (5), we define the whole model as:

$$\phi^{WSL}(x) = \phi^{pred}(\phi^{enc}(x)), \tag{6}$$

and class probabilities as:

$$P(c|x) = [\sigma(\phi^{WSL}(x))]_c \tag{7}$$

where $P(c|x)$ is the probability predicted for the c^{th} class given image x and σ denotes the softmax operation.

4.3. Multisource WSL models

We base our models on the assumption that (at least) one of the m sources does not have a high uncertainty regarding the object location like the other sources. For simplicity, we refer to this source as x_1. This typically corresponds to the high-resolution RGB imagery where georeferencing can be done relatively precisely and the object is located centrally within the image. We aim to use this *reference source* to mitigate the uncertainty in the other sources $(x_2, \ldots, x_M)$, which are referred to as the *additional sources*. The ultimate goal is to increase the overall classification performance by extracting (more) precise information from the additional sources.

To handle this ambiguity in weakly labeled sources, we propose four weakly supervised multisource models with instance attention. These models handle the weakly supervised fusion problem progressively in different levels, as indicated by their names: (i) Probability-Level-Fusion, (ii) Logit-Level-Fusion, (iii) Feature-Level-Fusion, and (iv) Pixel-Level-Fusion. In the following, we define and discuss these model schemes in detail. Figure 4 illustrates the proposed models.

Probability-Level-Fusion. In this model, we propose to combine additional data sources with the reference source by taking an average of the output probabilities of all sources:

$$P(c|x_{1:M}) = \frac{1}{M} \sum_{m=1}^{M} P(c|x_m) \tag{8}$$

where $P(c|x_m)$ is obtained as in (7) using a separate instance attention network (ϕ_m^{WSL}) for each $m \in \{2, \ldots, M\}$ and a simple CNN (ϕ_{ref}^{CNN}) for the reference source x_1. The only difference from (7) is that the logits coming from the additional sources $\phi_m^{WSL}(x_m)$ are divided by a temperature parameter $T_m < 1$ before the softmax operation to sharpen the output distribution, which is much smoother compared to the output of the reference $P(c|x_1)$. A summary of this approach is given in Algorithm 1.

Algorithm 1 Probability-Level-Fusion

Input: $x_1, \ldots, x_M$
Output: $P(\cdot|x_{1:M})$
1: $\mathbf{p}_1 \leftarrow \sigma(\phi_{ref}^{CNN}(x_1))$
2: **for** $m \leftarrow 2$ to M **do**
3: $\quad \mathbf{p}_m \leftarrow \sigma(\phi_m^{WSL}(x_m))$
4: **end for**
5: $P(\cdot|x_{1:M}) \leftarrow \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_m$

Combining the sources at the probability level corresponds to giving equal weights to the outputs of all sources and allowing them to contribute to the final classification evenly. This could cause a source with a more confident prediction to have a higher impact on the final decision, which can be desirable or undesirable depending on the reliability of that particular source. The temperature parameter enables the model to pay more/less attention to some of the sources by adjusting the confidence levels of their predictions.

Logit-Level-Fusion. We propose to combine the sources in the logit level in this model, by taking a weighted sum of the logit vectors obtained from the reference source via reference CNN (ϕ_{ref}^{CNN}) and the additional sources via weakly supervised instance attention networks (ϕ_m^{WSL}) using the following formulation:

$$\phi^{comb}(x_{1:M}) = \alpha_1 \phi_{ref}^{CNN}(x_1) + \sum_{m=2}^{M} \alpha_m S^{-1}(\phi_m^{WSL}(x_m)) \tag{9}$$

$$P(c|x_{1:M}) = [\sigma(\phi^{comb}(x_{1:M}))]_c \tag{10}$$

where S^{-1} is the inverse sigmoid function that maps WSL logits from the interval $[0, 1]$ to $(-\infty, \infty)$ to make them comparable to the logits obtained from the reference network. Weights α_m of the summation in (9) are obtained using softmax over learnable parameters β_m:

$$\alpha_m = \frac{\exp(\beta_m)}{\sum_{i=1}^{M} \exp(\beta_i)}. \tag{11}$$

The Logit-Level-Fusion approach is summarized in Algorithm 2. In this formulation, since the sources with equally confident individual predictions can have different logits, the impact of each source on the final decision can be different. Conversely, even when a source has less confidence in a particular class than some other source, it could contribute more to the score of that class if the corresponding logit is larger. Therefore, combining the sources in the logit-level instead of probability-level aims to add more flexibility to the model in terms of each source's effect on the joint classification result.

Feature-Level-Fusion. For each additional source m, we propose to combine penultimate layer feature vector of the reference network $\phi_{ref}^{enc}(x_1)$ with the candidate region feature representations of each additional source $\phi_m^{enc}(x_m)$. For this purpose, we replicate $\phi_{ref}^{enc}(x_1)$ R_m times, and concatenate with $\phi_m^{enc}(x_m)$ to

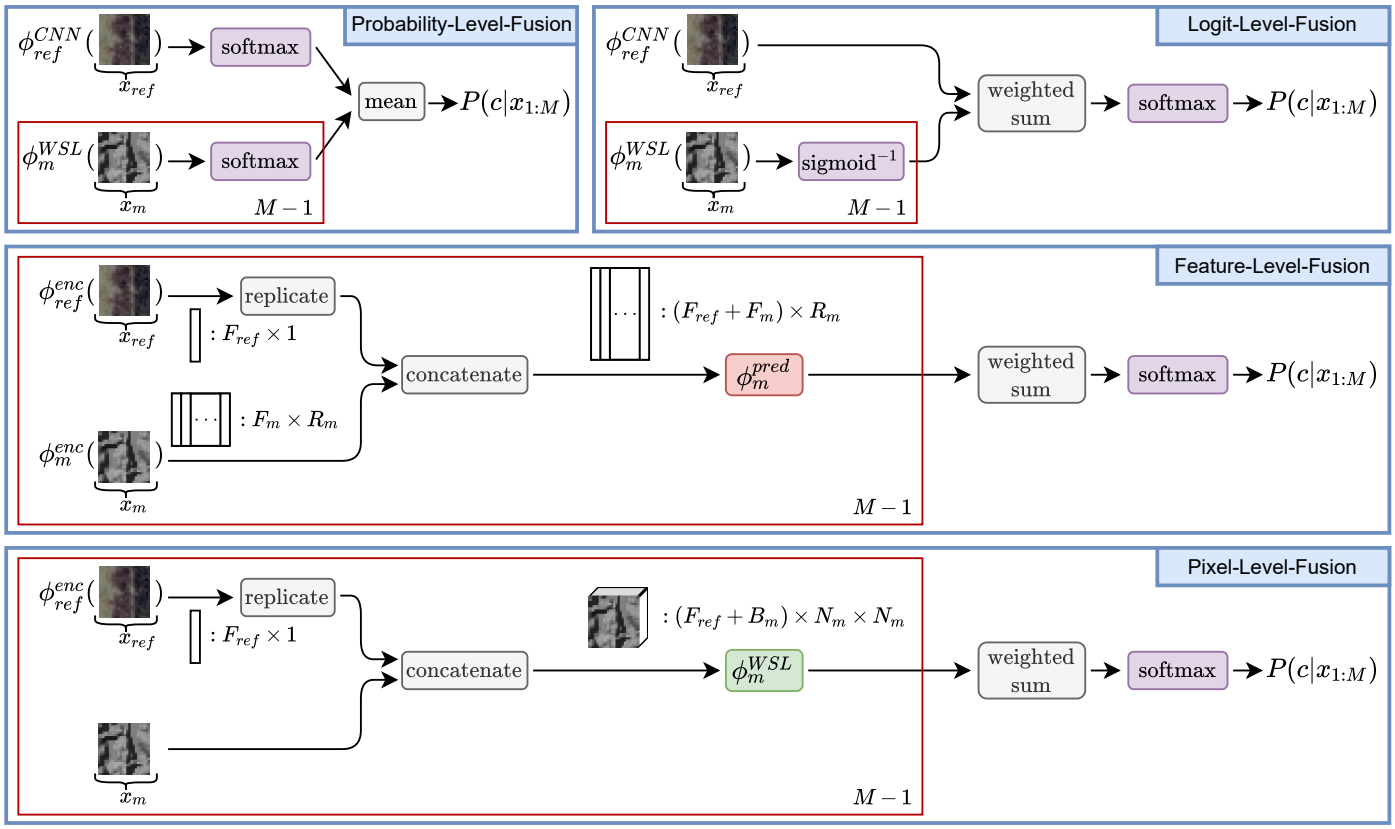

Figure 4: Illustration of the proposed multisource WSL models. The reference source x_1 is represented with subscript *ref*, while subscript $m \in \{2 \dots M\}$ is used for the additional sources. Plate notation is used to represent the repetitions in the model. Variables are described in the text.

Algorithm 2 Logit-Level-Fusion

Input: $x_1, \dots, x_M$
Output: $P(\cdot|x_{1:M})$
1: $\mathbf{y}_1 \leftarrow \phi_{ref}^{CNN}(x_1)$
2: **for** $m \leftarrow 2$ to M **do**
3: $\quad \mathbf{y}_m \leftarrow S^{-1}(\phi_m^{WSL}(x_m))$
4: **end for**
5: $P(\cdot|x_{1:M}) \leftarrow \sigma(\sum_{m=1}^{M} \alpha_m \mathbf{y}_m)$

Algorithm 3 Feature-Level-Fusion

Input: $x_1, \dots, x_M$
Output: $P(\cdot|x_{1:M})$
1: **for** $m \leftarrow 2$ to M **do**
2: $\quad \mathbf{y}_m \leftarrow \phi_m^{pred}(\psi(\phi_{ref}^{enc}(x_1), \phi_m^{enc}(x_m)))$
3: **end for**
4: $P(\cdot|x_{1:M}) \leftarrow \sigma(\sum_{m=2}^{M} \alpha_m \mathbf{y}_m)$

obtain fused feature vectors of size $F_{ref} + F_m$ for each of the R_m candidate regions. The resultant vectors are processed by ϕ_m^{pred} in the same way as the single-source model to obtain a logit vector per additional source. Finally, these logits are combined in the form of a weighted sum:

$$\phi^{comb}(x_{1:M}) = \sum_{m=2}^{M} \alpha_m \phi_m^{pred}(\psi(\phi_{ref}^{enc}(x_1), \phi_m^{enc}(x_m))) \qquad (12)$$

where ψ denotes the aforementioned replication and concatenation operations. Class probabilities are obtained using (10). Instead of an image-level combination, this approach focuses on utilizing the reference source earlier in the candidate region level. The idea behind this is to allow the model to leverage the lower-level information in the reference features and the candidate region features towards better classification and localization of the objects. Algorithm 3 provides a procedural description of Feature-Level-Fusion.

Pixel-Level-Fusion. Finally, we propose another form of a concatenation of the penultimate reference features with the additional sources. This time, instead of concatenating the reference feature vector with the feature vectors of the candidate regions obtained via ϕ_m^{enc}, we replicate and concatenate them directly to the pixels of the images of the additional sources (x_m), similar to the fusion technique in (Sumbul et al., 2019). The fused input for the m^{th} source is then processed by ϕ_m^{WSL} to obtain per-source logits. Finally, we take a weighted sum of the logits to obtain the combined logit vector:

$$\phi^{comb}(x_{1:M}) = \sum_{m=2}^{M} \alpha_m \phi_m^{WSL}(\psi(\phi_{ref}^{enc}(x_1), x_m)), \qquad (13)$$

which is followed by (10) to obtain class probabilities. In this scheme, the motivation behind combining reference features with the input pixels is that a higher-level descriptor of the tar-

Algorithm 4 Pixel-Level-Fusion

Input: $x_1, \ldots, x_M$
Output: $P(\cdot|x_{1:M})$
1: **for** $m \leftarrow 2$ to M **do**
2: $\quad \mathbf{y}_m \leftarrow \phi_m^{WSL}(\psi(\phi_{ref}^{enc}(x_1), x_m))$
3: **end for**
4: $P(\cdot|x_{1:M}) \leftarrow \sigma(\sum_{m=2}^{M} \alpha_m \mathbf{y}_m)$

get object coming from the reference source could be useful in the pixel-level to guide the network towards a better localization, and therefore a better classification, of the object. The Pixel-Level-Fusion approach is summarized in Algorithm 4.

5. Experiments

In this section, we first describe our experimental setup and implementation details for all methods. Then, we present our multisource results and compare them with other multisource methods as well as our single-source results.

5.1. Experimental setup

We conduct all experiments using two different multisource settings: (i) RGB & MS, and (ii) RGB, MS & LiDAR. The exact training procedure of each model differs from each other, especially in how they are pre-trained, which we observed to be very important on the final performance of the model. Here, we first outline the common aspects of the overall training procedure which is shared among all models. Then, we give the model-specific details about certain changes in the training procedure and hyper-parameters.

For all experiments, we randomly split the data set into training (60%), validation (20%), and test (20%) sets. All of the models are trained on the training set using Adam optimizer with learning rate 10^{-3}. ℓ_2-regularization with weight 10^{-5} is applied to all trainable parameters. These settings are same as in (Sumbul et al., 2019). We use batches of size 100 in each iteration. Each batch is drawn from an oversampled version of the training set to cope with the class imbalance. Oversampling rate for each class is proportional to the inverse frequency of that class. We augment the training set by shifting each image in both spatial dimensions with the amount of shift in each dimension randomly chosen between 0 and 20% of the width/height. We adopt early-stopping with a patience value of 200 to schedule the learning rate and terminate the training. If the validation accuracy does not improve for 200 consecutive epochs, we first load the checkpoint with the highest accuracy and decrease the learning rate by a factor of 10. If no improvement is observed for another 200 epochs, we stop the training and choose the checkpoint with the highest validation accuracy for testing. We use normalized accuracy as the performance metric where the per-class accuracy ratios are averaged to avoid biases towards classes with more examples.

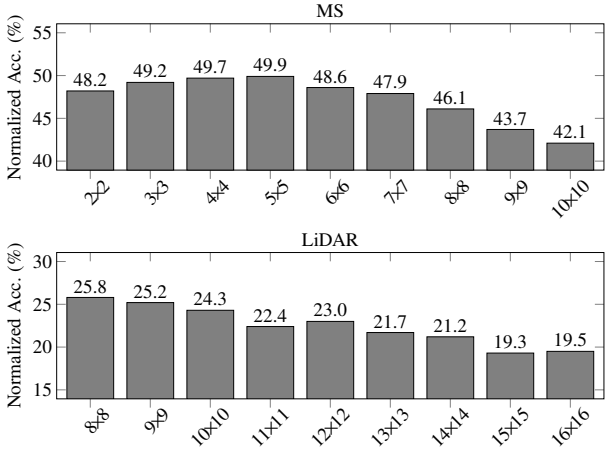

Figure 5: Impact of region proposal size (W) in terms of normalized validation accuracy for the single-source instance attention models. Proposals are extracted within 12×12 pixel neighborhoods for MS and 24×24 pixel neighborhoods for LiDAR as described in Section 3. The smallest proposal sizes for MS and LiDAR are 2×2 and 8×8, respectively, because no pooling is used in the feature encoding network of the former whereas three pooling operations are included in the convolutional layers for the latter as described in Figure 3.

5.2. Implementation details

Single-source baseline classification networks. We train three separate single-source classification networks for RGB, MS, and LiDAR. The basic network architectures (CNN) are taken from (Sumbul et al., 2019). Dropout regularization is applied with a drop probability of 0.25 in the convolutional layers and 0.5 in the first fully-connected layer. We use the pre-trained RGB network to initialize the reference branch of all proposed multisource models and the pre-trained MS/LiDAR networks to initialize the MS/LiDAR branches. Such a pre-training strategy increases the validation score, which in turn improves the performance of the multisource models that are fine-tuned after being initialized.

Single-source instance attention models. Fully-connected layers of classification and localization branches of weakly supervised instance attention networks are initialized randomly while convolutional layers are initialized from the corresponding pre-trained baseline single-source classification networks. Similar to the basic classification network, we apply dropout with 0.25 drop probability in the convolutional layers and 0.5 drop probability in the first fully-connected layer. We choose the region size parameter W as 5 pixels for MS and 8 pixels for LiDAR, which yield the highest validation accuracies in the experiments summarized in Figure 5.

Due to the multiplication of softmax outputs of classification and localization branches, output logits lie in the interval $[0, 1]$ when bias is not taken into account. Applying the final softmax operation before loss calculation with such logits results in smooth class distributions. Our experiments confirm that sharpening these distributions by introducing a temperature parameter (T) improves the performance of the model. With the addition of temperature, the final softmax in (7) becomes:

$$P(c|x) = [\sigma(\phi^{WSL}(x)/T)]_c. \quad (14)$$

Using our preliminary results on the validation set, we fix T to $1/60$ for both MS and LiDAR.

Probability-Level-Fusion. We observe that fine-tuning the network consisting of a pre-trained basic RGB network and pre-trained instance attention models combined as Probability-Level-Fusion does not improve the validation score. Furthermore, random initialization instead of pre-training worsens the network performance. Upon this observation, although it is possible to train/fine-tune the whole model end-to-end, we decide not to apply any fine-tuning. We choose temperature parameters (T_m) on the validation set via grid search, resulting in $1/48$ for MS and $1/18$ for LiDAR for the fused model.

Logit-Level-Fusion. We initialize the RGB network and instance attention models from pre-trained models as in Probability-Level-Fusion. β_m parameters in (11) are chosen as 1 for RGB and 2.5 for MS branch in the RGB & MS setting; 1 for RGB, 2.5 for MS, and 1.5 for LiDAR branch in the RGB, MS & LiDAR setting using the validation set. We also observe the temperature parameter to be useful in this case as well, and set it to 0.25 for both MS and LiDAR branches. The whole network is trained in an end-to-end fashion using dropout with a drop probability of 0.25 in the convolutional and 0.5 in the first fully-connected layers of all branches.

Feature-Level-Fusion. Even though it is possible to train both MS and LiDAR branches of the model jointly in all-sources setting, we obtain a higher validation accuracy when we combine separately trained RGB & MS and RGB & LiDAR models. After individual training of the MS and LiDAR branches, we choose the logit combination weights α_m in (12) on the validation set as 0.74 for MS and 0.26 for LiDAR to obtain the combined RGB, MS & LiDAR classification results. As an alternative, we have tried incorporating logit combination similar to (11) but it performed worse.

For the training of RGB & MS and RGB & LiDAR models, we initialize the whole RGB network from the pre-trained basic CNN model following the same approach as the previous models. For MS and LiDAR branches, convolutional layers are initialized from pre-trained instance attention models while fully-connected layers are initialized randomly, since the sizes of the fully-connected layers in classification and localization branches change due to feature concatenation. Furthermore, we observe that freezing all pre-trained parameters and training the rest of the models yields better validation accuracies. Although we freeze some of the network parameters, we find that leaving the dropout regularization on for the frozen layers improves the performance. For the RGB & MS setting, we use a 0.5 and 0.1 drop probability for the convolutional and penultimate fully-connected layers, respectively, and T is tuned to 0.05. For the RGB & LiDAR setting, we use a 0.1 and 0.5 drop probability for the convolutional and penultimate fully-connected layers, respectively, and T is tuned to 0.025.

Pixel-Level-Fusion. We make the same observation in this model as in Feature-Level-Fusion that combining separately trained RGB & MS and RGB & LiDAR models results in better validation performance than training both branches jointly. Similarly, we obtain higher validation accuracy for logit combination weights α_m chosen as 0.76 for MS and 0.24 for LiDAR using a grid search.

For the training of RGB & MS and RGB & LiDAR branches, the basic RGB network is initialized using the pre-trained model. Since the size of the first convolutional layer of the instance attention model is different from its single-source version, the first layer is initialized randomly. We also observe that random initialization of the classification and localization branches results in higher scores. Other layers are initialized from the pre-trained instance attention model and the whole network is fine-tuned end-to-end. The drop probability of dropout is chosen as 0.25 for the convolutional layers and 0.5 for the fully-connected layers. The temperature parameter is set to $1/60$ and kept constant as in the other models.

5.3. Results

Single-source results and ablation study. We first evaluate the effectiveness of the instance attention framework in the case of single-source object recognition. For this purpose, we compare the MS-only and LiDAR-only instance attention models to the corresponding single-source baselines described in Section 5.2, as well as a single-source spatial transformer network (STN) based model. For the latter, we adapt the methodology of He and Chen (2019) to our case and use an STN to select a candidate region from each input image. In the STN baseline, selected candidate regions are scaled to the same size as the input images and are classified by a CNN with the same architecture as the single-source baseline models. We restrict STN to use only translation and scaling transformations and use the same scale for both spatial dimensions, which results in three parameters to estimate. We estimate these parameters using a separate CNN with the same architecture as the single-source baseline model, but replace the final layer with a 3-dimensional fully-connected layer. We note that RGB is the reference high-resolution source and contains centered object instances, therefore, instance attention and STN are not applicable to RGB inputs.

Single-source results are presented in Table 2. From the results we can see that instance attention significantly improves both the MS-only results (from 40.6% to 48.3%) and the LiDAR-only results (21.2% to 25.3%). The essential reason for the large performance gap is the fact that single-source baselines aim to model the images holistically. This can be interpreted as separately modeling each potential instance location of each class when applied to a larger area, which is clearly very ineffective. In contrast, instance attention models rely on local recognition of image regions and attention-driven accumulation of local recognition results, which is much more resilient to positional ambiguity. We also observe that instance attention yields consistently better results in comparison to STN on both MS and LiDAR inputs.

As an ablative experiment, we additionally evaluate the importance of the localization branch (i.e., the ϕ^{loc} component) in instance attention models. We observe that the localization branch improves the MS-only result from 47.7% to 48.3% and the LiDAR-only result from 24.3% to 25.3%. These results show that the model with only the classification branch already

Table 2: Single-source baseline networks, single-source instance attention models, and ablation study results in terms of normalized test accuracy (%).

Model	Accuracy
Single-source baseline (RGB)	25.3
Single-source baseline (MS)	40.6
Single-source STN (He and Chen, 2019) (MS)	41.1
Instance attention, ϕ^{cls} only (MS)	47.7
Instance attention (MS)	48.3
Single-source baseline (LiDAR)	21.2
Single-source STN (He and Chen, 2019) (LiDAR)	20.8
Instance attention, ϕ^{cls} only (LiDAR)	24.3
Instance attention (LiDAR)	25.3

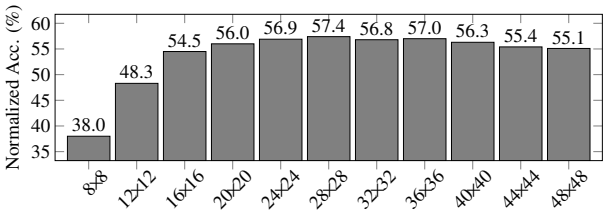

Figure 6: Effect of neighborhood size (N) in terms of normalized test accuracy for the MS-only instance attention model. 5×5 pixel proposals are extracted within these image neighborhoods.

performs significantly better than single-source baseline models thanks to handling object recognition locally. Incorporation of the localization branch further improves the results thanks to better handling of positional ambiguity.

Finally, we compare the MS-only and LiDAR-only instance attention models against the RGB-only single-source baseline. We observe that all MS models significantly outperform the RGB-only result, highlighting the value of detailed spectral information. We also observe that LiDAR is much less informative compared to MS, and, only the full single-source instance attention model for LiDAR is able to match the results of the RGB-only baseline model.

Effect of neighborhood size. The proposed instance attention model uses $W \times W$ pixel windows as region proposals within $N \times N$ neighborhoods as shown in Figure 3. We have used 12×12 pixel neighborhoods for MS and 24×24 pixel neighborhoods for LiDAR to be consistent with our previous work (Sumbul et al., 2019). To study the effect of different neighborhood sizes, we perform additional experiments by fixing the region size parameter W to the best performing value of 5 pixels as presented in Figure 5 and by varying the neighborhood size parameter N for the MS data. The results are presented in Figure 6. We observe that the accuracy increases beyond the previously used setting of 12×12 pixels until the neighborhood size reaches 28×28 and starts to slightly decrease afterwards. Even though this increase seems to imply the necessity of using larger neighborhoods, it is important to note that there are other factors that affect this performance. An analysis of the tree locations in the GIS data shows that the average distance between neighboring trees in the MS data is slightly above 6 pixels. This means that increasing the window size too much also increases

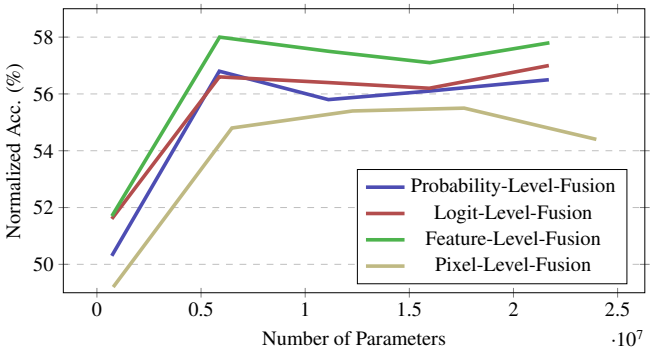

Figure 7: Effect of width-wise increasing the number of parameters, hence the model capacity, in terms of normalized test accuracy for the RGB & MS setting. This analysis aims to make a fair comparison among the multisource instance attention formulations.

the risk of including highly overlapping regions across training and test samples. In addition, these relative accuracy improvements can also be a superficial result of the exploitation of background patterns, which is more likely to happen when the provided background contexts are large enough to be informative for class discrimination. In the rest of the experiments, therefore, we continue to use 12×12 pixels for MS and 24×24 pixels for LiDAR for the neighborhood size as in Sumbul et al. (2019).

Performance versus model capacity. We now examine how the proposed models perform with different model capacities for the RGB & MS setting. We believe that it is immensely important to compare formulations with similar model complexities, and evaluate how their performances vary as a function of model complexity, to reach accurate conclusions. For that purpose, we use the number of parameters as a proxy for the model capacity, and train all models by keeping the network depth (i.e., number of layers) constant while increasing the network width (i.e., number of filters for the convolutional layers and number of output units for the fully-connected layers). We run these experiments in five different settings, in each of which the width is increased by a constant factor, starting from the default model capacity setting where the number of parameters of each method is comparable to the model in (Sumbul et al., 2019).

Figure 7 shows the number of parameters of each model in each of these settings and their corresponding test scores. According to this, although Logit-Level-Fusion and Feature-Level-Fusion produce very similar results in the default setting with fewer parameters, the gap between Feature-Level-Fusion and the other methods increases as the model capacity increases, which points out that Feature-Level-Fusion is superior to other methods. Furthermore, all models' scores tend to increase up to some point before plateauing, except for Pixel-Level-Fusion, which starts to drop as the number of parameters increases, due to the 8-channel input size of MS being constant while the number of RGB features concatenated to them increasing to the point that they dominate the MS input.

The results of the model capacity experiments highlight two important points which are often overlooked when different mod-

Table 3: Multisource instance attention results and comparison to state-of-the-art in terms of normalized test accuracy (%).

Model	Accuracy
Two sources (RGB & MS)	
Basic multisource model (Sumbul et al., 2019)	39.1
Recurrent attention model (Fu et al., 2017)	41.6
MRAN (Sumbul et al., 2019)	46.6
Instance attention - Probability-Level-Fusion	50.3
Instance attention - Logit-Level-Fusion	51.6
Instance attention - Feature-Level-Fusion	**51.7**
Instance attention - Pixel-Level-Fusion	49.2
Three sources (RGB, MS & LiDAR)	
Basic multisource model (Sumbul et al., 2019)	41.4
Recurrent attention model (Fu et al., 2017)	42.6
MRAN (Sumbul et al., 2019)	47.3
Instance attention - Probability-Level-Fusion	51.9
Instance attention - Logit-Level-Fusion	50.9
Instance attention - Feature-Level-Fusion	**53.0**
Instance attention - Pixel-Level-Fusion	51.6
Instance attention - Feature-Level-Fusion *(inc. capacity)*	58.0

Table 4: Impact of data augmentation via random horizontal and vertical shifts in terms of normalized test accuracy (%) for the RGB & MS setting.

Model	Accuracy	
	w/ aug.	w/o aug.
Instance att. - Probability-Level-Fusion	50.3	49.7
Instance att. - Logit-Level-Fusion	51.6	50.1
Instance att. - Feature-Level-Fusion	51.7	50.4
Instance att. - Pixel-Level-Fusion	49.2	47.5

els are compared in the literature. First, evaluating a model in a single capacity setting might yield sub-optimal results and prevent us from observing the full potential of the model. As an example, while Feature-Level-Fusion, the best performing model according to Table 3, achieves a test score of 51.7% in the default setting, it shows a significantly higher performance of 58.0% test accuracy with an increase in the model capacity. Second, comparing different methods in a single capacity setting might be an unreliable way of assessing the superiority of one method to another. For instance, the small difference of 0.1% between the Logit-Level-Fusion and Feature-Level-Fusion scores in the default setting hinders us to reach a clear conclusion between the two methods. However, observation of a 1.4% difference with a higher capacity enables us to verify Feature-Level-Fusion's superiority. Furthermore, the performance difference between Logit-Level-Fusion and Probability-Level-Fusion closes or becomes reversed at different points as we increase the number of parameters.

Comparison to the state-of-the-art. We compare the four proposed models against three state-of-the-art methods. The first method is named the *basic multisource model* that implements the commonly used scheme of extracting features independently from individual sources and concatenating them as the multisource representation that is used as input to fully-connected layers for the final classification. We use the end-to-end trained implementation in (Sumbul et al., 2019). The second method is the *recurrent attention model* (Fu et al., 2017). This model processes a given image at different scales using a number of classification networks. An attention proposal network is used to select regions to attend in a progressive manner. Classification networks are trained with intra-scale classification loss while inter-scale ranking loss, which enforces the next scale classification network to perform better than the previous scale, is used to train the attention proposal networks. The third state-of-the-art method is the *Multisource Region Attention Network* (MRAN) (Sumbul et al., 2019), which has been shown to be an effective method for multisource fine-grained object recognition. MRAN extracts candidate regions from MS and/or LiDAR data in a sliding window fashion and extracts features from these candidates by processing them with a CNN. The features are pooled in a weighted manner to obtain an attention-based representation for the corresponding source. Attention weights are obtained through a separate network that takes pixel-wise concatenation of RGB features, coming from the same basic single-source network architecture that we use, to the candidate regions as the input. The final multisource representation is obtained by the concatenation of RGB, MS and/or LiDAR representations, which is used for classification.

Table 3 lists the normalized test accuracies for the default model capacity setting (except for the bottom-most row), where the number of parameters is comparable to MRAN to enable comparisons with the state-of-the-art. Looking at these results, we see that all proposed methods outperform MRAN as well as the basic multisource model and the recurrent attention model. An interpretation for this could be that the instance attention is better suited to the classification task, arguably thanks to stronger emphasis on particular candidate regions. The last row of Table 3 shows the performance of Feature-Level-Fusion for RGB & MS in a higher model capacity setting with an 11.4% improvement over MRAN for RGB & MS, indicating that the model we propose can be scaled-up for increased performance. Feature-Level-Fusion consistently outperforming Probability-Level-Fusion and Logit-Level-Fusion for all capacity settings in Figure 7 and both RGB & MS and RGB, MS & LiDAR settings in Table 3 indicates that combining the reference source with the additional sources earlier helps the network to better locate and classify the object of interest by making use of the additional information in the reference features which is not present in the logit level. The drop on the performance of Pixel-Level-Fusion, on the other hand, shows that fusing high-level reference features with low-level pixel values is not as effective as using reference features just before the classification and localization branches.

Effect of data augmentation. As previously explained in Section 5.1, we use random shift based spatial data augmentation during training. In this part, we analyze the effect of this data augmentation policy on the recognition rates. For this purpose, we train and evaluate the multisource instance attention models for the RGB & MS setting with and without data augmentation. The results presented in Table 4 show that data augmentation

Table 5: Stability analysis of the multisource instance attention models in terms of normalized test accuracy (%). Mean and standard deviation of the scores of five different runs are shown.

Model	Accuracy
Two sources (RGB & MS)	
Instance attention - Probability-Level-Fusion	50.5 ± 0.8
Instance attention - Logit-Level-Fusion	52.0 ± 0.5
Instance attention - Feature-Level-Fusion	51.5 ± 0.8
Instance attention - Pixel-Level-Fusion	49.5 ± 0.3
Three sources (RGB, MS & LiDAR)	
Instance attention - Probability-Level-Fusion	51.9 ± 0.4
Instance attention - Logit-Level-Fusion	51.4 ± 0.6
Instance attention - Feature-Level-Fusion	53.1 ± 0.6
Instance attention - Pixel-Level-Fusion	51.1 ± 0.3

consistently improves each model by amounts varying from 0.6 to 1.7 points. We also observe that relative performances of the multisource models remain the same with and without data augmentation.

Stability analysis. The previous experiments use a single split of the data set into training, validation, and test sets. In this part, we evaluate the stability of the multisource instance attention models under different partitionings of the data set. For this purpose, we use a random split of the data set into five folds. (All of the previous experiments correspond to the combination of the first three folds (60%) as the training set, with the fourth fold (20%) being the validation and the fifth fold (20%) being the test sets.) For the stability analysis, we train and evaluate all models five times and report the mean and standard deviation of the test scores of these five runs. In each run, we use a unique combination of three folds as the training set, one of the remaining folds as the validation set, and the other fold as the test set. As a result, each of the five folds appears as an independent test set in the five runs. The results presented in Table 5 show that the performance variation across the folds is relatively small, which is reassuring about the stability of the models.

Class-specific results. Figure 8 presents example results for the class-specific performances of the proposed methods. We observe that the classes receive different levels of contributions from different sources. When we consider the models for the individual sources, the MS network performs significantly better than the RGB network where all classes have an improvement between 1% and 44% in accuracy. On the other hand, half of the classes have better performance under the RGB network compared to the other half that perform better with the LiDAR network. When we compare the effect of the instance attention mechanism to the baseline single-source MS network, we observe that every one of the 40 classes enjoys an improvement in the range from 1% to 19%. Similarly, for the use of the attention mechanism in the single-source LiDAR network, 27 of the classes receive higher scores with a maximum of 21%. When we consider the best performing fusion model (Feature-Level-Fusion) under the RGB & MS versus RGB, MS & LiDAR settings, we observe that 30 of the classes have improvements up to 7% with the latter. Most of the classes that do not improve are among the ones with the least number of samples in the data set. Finally, when the increased capacity network in the bottom-most row of Table 3 is compared to the default capacity one, the maximum improvement for the individual classes increases to 25%. Overall, although the highest scoring model contains both MS and LiDAR sources, the contribution of the LiDAR data to the performance seems to be less significant compared to the MS data. This indicates that the richer spectral information in the MS images provides more useful information than the LiDAR data for the fine-grained classification task. In addition, the proposed weakly supervised instance attention mechanism benefits the source (MS) with the smallest expected object size (maximum of 4×4 pixels) the most.

Figure 9 shows the confusion matrix resulting from the Feature-Level-Fusion (RGB, MS & LiDAR) model. We observe that most confusions are among the tree classes that belong to the same families in the scientific taxonomy shown in Figure 2. For example, 28% of the thundercloud plum samples are wrongly predicted as cherry plum and 13% are wrongly predicted as blireiana plum, whereas 19% of the cherry plum samples are wrongly predicted as thundercloud plum and 15% are wrongly predicted as blireiana plum. Similarly, Kwanzan cherry and Chinese cherry have the highest confusion with each other, with 11% and 12% misclassification, respectively, 17% of common hawthorn are confused with midland hawthorn, and 25% of scarlet oak are misclassified as red oak. As the largest family of trees, maples also have confusions among each other, with notable errors for red maple, paperback maple, and flame amur maple. In particular, flame amur maple has some of the highest confusions as being the class with the fewest number of samples. As other examples for the cases with the highest confusion, 11% of the Japanese snowbell samples and 11% of autumn serviceberry are wrongly predicted as Japanese maple. All of these three types of trees have moderate crown density and have a spread in the $15-25$ feet range. Furthermore, autumn serviceberry and Japanese maple both have heights in the $15-20$ feet range (see (Sumbul et al., 2018) for a description of the attributes for the tree categories in the data set). As a final example, 11% of orchard apple samples are wrongly predicted as Kwanzan cherry, with both species having moderate crown density, medium texture, and spread in the $15-25$ feet range. Similar behaviors are observed for all other models. Since most of these types of trees are only distinguished with respect to their sub-species level in the taxonomy and have almost the same visual appearance, their differentiation using regions of few pixels from an aerial view is a highly challenging problem. We think that the overall normalized accuracy of 53% shows a significant performance for the fine-grained classification of 40 different tree categories.

Qualitative results. Figure 10 illustrates the region scores, normalized to the $[0, 1]$ range, obtained by multiplying per-region classification and localization scores in the Hadamard product in (4) for the predicted class. Our first observation is that the region scores for MS tend to have a smoother distribution with mostly a single local maximum, while LiDAR scores appear

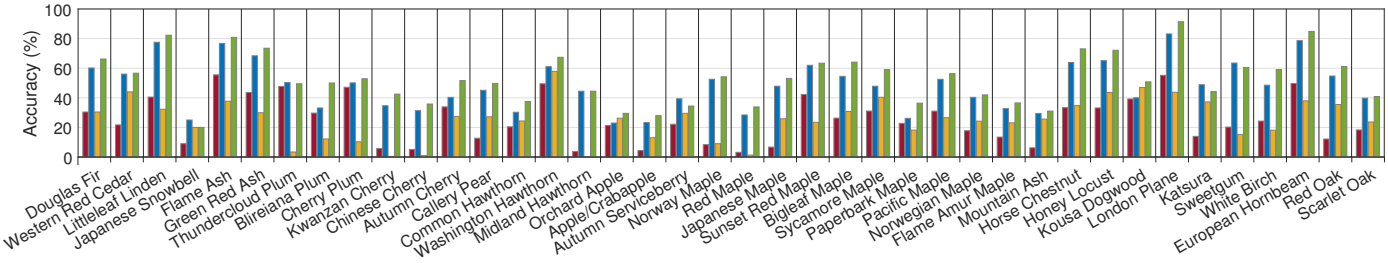

Figure 8: Example results for the class-specific performances of the proposed methods. From left to right: single-source RGB (red), single-source MS with instance attention (blue), single-source LiDAR with instance attention (yellow), and Feature-Level-Fusion (RGB, MS & LiDAR) (green). Best viewed in color.

to be much noisier. This is in line with our previous observation that the information provided by the MS data appears to be more useful for the localization of the target object, which could explain its significantly higher contribution to the multisource classification results compared to LiDAR.

Figure 10(a) supports this observation, where all methods successfully localize the objects in the MS image and classify the input correctly, even though Logit-Level-Fusion and Pixel-Level-Fusion highlight different regions than the other methods for LiDAR. However, it is also possible to observe some cases as Figure 10(b), where strong predictions by the LiDAR branch affect the final classification even when similar localization results are achieved by different methods for the MS data. In this particular case, the differences in the results of the LiDAR branches of Logit-Level-Fusion and Pixel-Level-Fusion have an impact on the misclassification of the input.

Next, we examine the localization results of the misclassified samples to better understand the effect of localization for the failure cases. For Figure 10(c), only Feature-Level-Fusion, which localizes the object differently than the other methods for both the MS and LiDAR data, is able to achieve a correct classification result. Looking at the corresponding MS image, we observe that the localization result indeed corresponds to a tree for Feature-Level-Fusion, which points out that the misclassification of the other methods could be due to wrong localization. A similar case is seen in Figure 10(d) as well, where only Pixel-Level-Fusion succeeds at correctly classifying the input with a different localization than the others. However, even though MS and LiDAR inputs seem to coincide up to a certain degree for this particular example, the position of the localized object seems to differ a lot between the MS and LiDAR data, which highlights the possibility that even in the case of correct classification, the models can attend to contextual cues rather than the object itself.

Even though localization has a substantial impact on the performance, we also observe failure cases for some samples such as Figure 10(g), where the models output incorrect predictions even though the localization is successful. This result shows that we could achieve higher scores with the proposed approaches by improving their fine-grained classification performances in addition to their localization capabilities.

6. Conclusions

We studied the multisource fine-grained object recognition problem where the objects of interest in the input images have a high location uncertainty due to the registration errors and small sizes of the objects. We approached the location uncertainty problem from a weakly supervised instance attention perspective by cropping input images at a larger neighborhood around the ground truth location to make sure that an object with a given class label is present in the neighborhood even though the exact location is unknown. Using such a setting, we formulated the problem as the joint localization and classification of the relevant regions inside this larger neighborhood. We first outlined our weakly supervised instance attention model for the single-source setting. Then we provided four novel fusion schemes to extend this idea into a multisource scenario, where a reference source, assumed to contain no location uncertainty, can be used to help the additional sources with uncertainty to better localize and classify the objects.

Using normalized accuracy as the performance measure, we observed that all of the proposed multisource methods achieve higher classification scores than the state-of-the-art baselines with the best performing method (Feature-Level-Fusion) showing a 5.1% improvement over the best performing baseline using RGB & MS data, and a 5.7% improvement using RGB, MS & LiDAR data. Additionally, we provided an in-depth comparison of the proposed methods with a novel evaluation scheme studying the effect of increased model capacity on the model performance. As a result of this experiment, we confirmed that Feature-Level-Fusion is indeed the most promising approach among all proposed methods, with an accuracy of 58.0% using RGB & MS data, which is a 6.3% improvement compared to the default capacity setting. Future work directions include the use of additional multisource fine-grained data sets for illustrating the generalizability of the proposed method, the use of additional measures such as the Kappa coefficient for performance evaluation, and the extension of the proposed model to handle other types of uncertainties such as temporal changes in addition to the spatial uncertainties studied in this paper.

Conflict of interest

We confirm that there are no known conflicts of interest associated with this work.

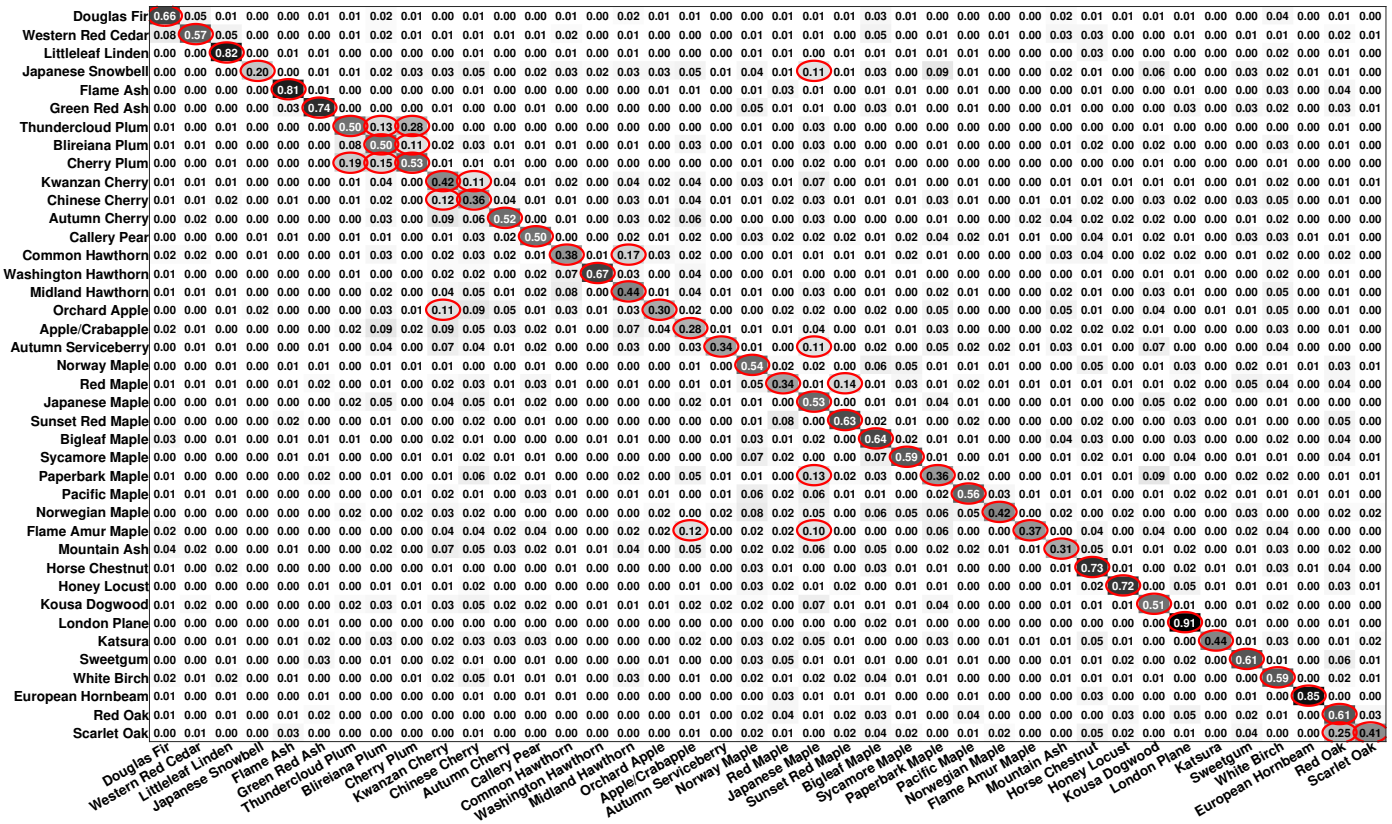

Figure 9: Confusion matrix for Feature-Level-Fusion (RGB, MS & LiDAR). Values represent percentages where each row sums to 1. The numbers greater than 0.10 are marked with red ellipses. The ones on the diagonal show correct classification, whereas off-diagonal entries indicate notable confusions. The Kappa coefficient for this matrix is obtained as 0.5116.

Acknowledgment

This work was supported in part by the TUBITAK Grant 116E445 and in part by the BAGEP Award of the Science Academy.

References

Ali, M.U., Sultani, W., Ali, M., 2020. Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery. ISPRS J. Photogram. Remote Sens. 162, 115–124.

Aygüneş, B., Aksoy, S., Cinbiş, R.G., 2019. Weakly supervised deep convolutional networks for fine-grained object recognition in multispectral images, in: IEEE Intl. Geosci. Remote Sens. Symp., pp. 1478–1481.

Bilen, H., Vedaldi, A., 2016. Weakly supervised deep detection networks, in: IEEE Conf. Comput. Vis. Pattern Recog., pp. 2846–2854.

Branson, S., Wegner, J.D., Hall, D., Lang, N., Schindler, K., Perona, P., 2018. From Google Maps to a fine-grained catalog of street trees. ISPRS J. Photogram. Remote Sens. 135, 13–30.

Campos-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Saux, B.L., Beaupere, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., Randrianarivo, H., Ferecatu, M., Shimoni, M., Moser, G., Tuia, D., 2016. 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. Top. Appl. Earth Obs. Remote Sens. 9, 5547–5559.

Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Rojo-Alvarez, J.L., Martinez-Ramon, M., 2008. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. IEEE Trans. Geosci. Remote Sens. 46, 1822–1835.

Chen, B., Huang, B., Xu, B., 2017. Multi-source remotely sensed data fusion for improving land cover classification. ISPRS J. Photogram. Remote Sens. 124, 27–39.

City of Seattle, Department of Transportation, 2016. Seattle street trees. URL: `http://web6.seattle.gov/SDOT/StreetTrees/`.

Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., Benediktsson, J.A., 2015. Challenges and opportunities of multimodality and data fusion in remote sensing. Proc. IEEE 103, 1585–1601.

Datcu, M., Melgani, F., Piardi, A., Serpico, S.B., 2002. Multisource data classification with dependence trees. IEEE Trans. Geosci. Remote Sens. 40, 609–617.

Debes, C., Merentitis, A., Heremans, R., Hahn, J., Frangiadakis, N., van Kasteren, T., Liao, W., Bellens, R., Pizurica, A., Gautama, S., Philips, W., Prasad, S., Du, Q., Pacifici, F., 2014. Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 2405–2418.

Fassnacht, F.E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L.T., Straub, C., Ghosh, A., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sensing of Environment 186, 64–87.

Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., pp. 4476–4484.

Gao, G., Gu, Y., 2019. Tensorized principal component alignment: A unified framework for multimodal high-resolution images classification. IEEE Trans. Geosci. Remote Sens. 57, 46–61.

Ghamisi, P., Hofle, B., Zhu, X.X., 2017. Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 3011–3024.

Gomez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G., 2015. Multimodal classification of remote sensing images: A review and future directions. Proc. IEEE 103, 1560–1584.

Han, Y., Bovolo, F., Bruzzone, L., 2016. Edge-based registration-noise estimation in VHR multitemporal and multisensor images. IEEE Geosci. Remote Sens. Lett. 13, 1231–1235.

He, X., Chen, Y., 2019. Optimized input for CNN-based hyperspectral image

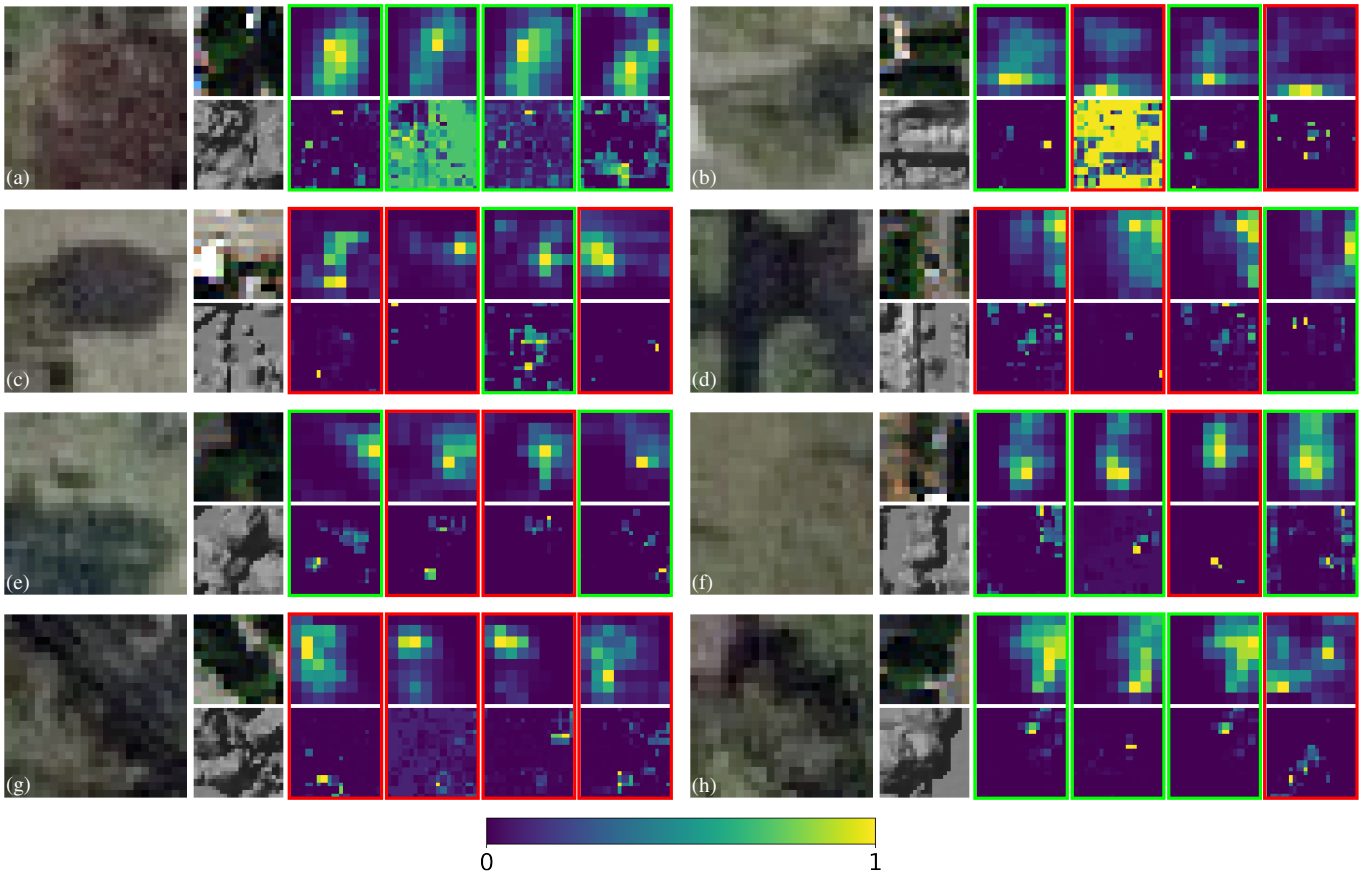

Figure 10: Region scores for sample test images. RGB images are shown in the first column, MS (top) and LiDAR (bottom) neighborhoods shown in the second. Remaining columns show instance attention results respectively for Probability-Level-Fusion, Logit-Level-Fusion, Feature-Level-Fusion, and Pixel-Level-Fusion. Results for correct class predictions are denoted with green boxes and those with wrong predictions are shown with red boxes. Region scores are obtained as the multiplication of per-region classification and localization scores corresponding to the predicted class. Best viewed in color.

classification using spatial transformer network. IEEE Geosci. Remote Sens. Lett. 16, 1884–1888.

Hong, D., Yokoya, N., Ge, N., Chanussot, J., Zhu, X.X., 2019. Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification. ISPRS J. Photogram. Remote Sens. 147, 193–205.

Hu, J., Mou, L., Schmitt, A., Zhu, X.X., 2017. Fusionet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data, in: Joint Urban Remote Sens. Event.

Hua, Y., Mou, L., Zhu, X.X., 2019. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. ISPRS J. Photogram. Remote Sens. 149, 188–199.

Ienco, D., Interdonato, R., Gaetano, R., Ho Tong Minh, D., 2019. Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture. ISPRS J. Photogram. Remote Sens. 158, 11–22.

Ji, J., Zhang, T., Yang, Z., Jiang, L., Zhong, W., Xiong, H., 2019. Aircraft detection from remote sensing image based on a weakly supervised attention model, in: IEEE Intl. Geosci. Remote Sens. Symp., pp. 322–325.

Laumer, D., Lang, N., van Doorn, N., Mac Aodha, O., Perona, P., Wegner, J.D., 2020. Geocoding of trees from street addresses and street-level images. ISPRS J. Photogram. Remote Sens. 162, 125–136.

Li, Y., Chen, W., Zhang, Y., Tao, C., Xiao, R., Tan, Y., 2020a. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. Remote Sensing of Environment 250, 112045.

Li, Y., Zhang, Y., Huang, X., Yuille, A.L., 2018. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. ISPRS J. Photogram. Remote Sens. 146, 182–196.

Li, Y., Zhang, Y., Zhu, Z., 2020b. Error-tolerant deep learning for remote sensing image scene classification. IEEE Transactions on Cybernetics , 1–13.

Liao, W., Huang, X., Coillie, F.V., Gautama, S., Pizurica, A., Philips, W., Liu, H., Zhu, T., Shimoni, M., Moser, G., Tuia, D., 2015. Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS Data Fusion Contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 2984–2996.

Ma, J., Zhang, L., Sun, Y., 2020. ROI extraction based on multiview learning and attention mechanism for unbalanced remote sensing data set. IEEE Trans. Geosci. Remote Sens. .

Morchhale, S., Pauca, V.P., Plemmons, R.J., Torgersen, T.C., 2016. Classification of pixel-level fused hyperspectral and lidar data using deep convolutional neural networks, in: 8th Workshop on Hyperspectral Image and Signal Processing, pp. 1–5.

Natural Resources Conservation Service of the United States Department of Agriculture, 2016. USDA Plants. URL: `https://plants.usda.gov/java/`.

Oliveau, Q., Sahbi, H., 2017. Learning attribute representations for remote sensing ship category classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 2830–2840.

Pibre, L., Chaumont, M., Subsol, G., Ienco, D., Derras, M., 2017. How to deal with multi-source data for tree detection based on deep learning, in: IEEE Global Conf. Signal Inf. Process.

Sumbul, G., Cinbis, R.G., Aksoy, S., 2018. Fine-grained object recognition and zero-shot learning in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 56, 770–779.

Sumbul, G., Cinbis, R.G., Aksoy, S., 2019. Multisource region attention network for fine-grained object recognition in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 57, 4929–4937.

Tuia, D., Volpi, M., Trolliet, M., Camps-Valls, G., 2014. Semisupervised manifold alignment of multimodal remote sensing images. IEEE Trans. Geosci. Remote Sens. 52, 7708–7720.

Voisin, A., Krylov, V.A., Moser, G., Serpico, S.B., Zerubia, J., 2014. Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach. IEEE Trans. Geosci. Remote Sens. 52, 3346–3358.

Wang, S., Chen, W., Xie, S.M., Azzari, G., Lobell, D.B., 2020. Weakly supervised deep learning for segmentation of remote sensing imagery. Remote Sensing 12, 207.

Xu, J., Wan, S., Jin, P., Tian, Q., 2019. An active region corrected method for weakly supervised aircraft detection in remote sensing images, in: 11th International Conference on Digital Image Processing.

Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., Zhang, B., 2018. Multisource remote sensing data classification based on convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56, 937–949.

Yokoya, N., Ghamisi, P., Xia, J., Sukhanov, S., Heremans, R., Tankoyeu, I., Bechtel, B., Le Saux, B., Moser, G., Tuia, D., 2018. Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 1363–1377.

Zhang, L., Ma, J., Lv, X., Chen, D., 2019. Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images. IEEE Geosci. Remote Sens. Lett. 17, 117–121.

Zhang, Y., Yang, H.L., Prasad, S., Pasolli, E., Jung, J., Crawford, M., 2015. Ensemble multiple kernel active learning for classification of multisource remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 845–858.

Zhang, Z., Guo, W., Li, M., Yu, W., 2020. GIS-supervised building extraction with label noise-adaptive fully convolutional neural network. IEEE Geosci. Remote Sens. Lett. .