WEAKLY SUPERVISED APPROACHES FOR IMAGE CLASSIFICATION IN REMOTE SENSING AND MEDICAL IMAGE ANALYSIS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Bulut Aygüneş

December 2020

WEAKLY SUPERVISED APPROACHES FOR IMAGE CLASSIFICATION IN REMOTE SENSING AND MEDICAL IMAGE ANALYSIS

By Bulut Aygüneş

December 2020

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Selim Aksoy(Advisor)

Ramazan Gökberk Cinbiş(Co-Advisor)

Hamdi Dibeklioğlu

Mehmet Erkut Erdem

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

WEAKLY SUPERVISED APPROACHES FOR IMAGE CLASSIFICATION IN REMOTE SENSING AND MEDICAL IMAGE ANALYSIS

Bulut Aygüneş
M.S. in Computer Engineering
Advisor: Selim Aksoy
Co-Advisor: Ramazan Gökberk Cinbiş
December 2020

Weakly supervised learning (WSL) aims to utilize data with imprecise or noisy annotations to solve various learning problems. We study WSL approaches in two different domains: remote sensing and medical image analysis.

For remote sensing, we focus on the multisource fine-grained object recognition problem that aims to classify an object into one of many similar subcategories. The task we work on involves images where an object with a given class label is present in the image without any knowledge of its exact location. We approach this problem from a WSL perspective and propose a method using a single-source deep instance attention model with parallel branches for joint localization and classification of objects. We then extend this model into a multisource setting where a reference source assumed to have no location uncertainty is used to aid the fusion of multiple sources. We show that all four proposed fusion strategies that operate at the probability level, logit level, feature level, and pixel level provide higher accuracies compared to the state-of-the-art. 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 over the default capacity setting.

For medical image analysis, we study breast cancer classification on regions of interest (ROI) of arbitrary shapes and sizes from breast biopsy whole slides. The typical solution to this problem is to aggregate the classification results of fixed-sized patches cropped from ROIs to obtain image-level classification scores. We first propose a generic methodology to incorporate local inter-patch context through a graph convolution network (GCN) that aims to propagate information over neighboring patches in a progressive manner towards classifying the whole ROI. The experiments using a challenging data set for a 3-class ROI-level classification task and comparisons with several baseline approaches show that the proposed model that incorporates the spatial context performs better than commonly used fusion rules. Secondly, we revisit the WSL framework we use in our remote sensing experiments and apply it to a 4-class ROI classification problem. We propose a new training methodology tailored for this WSL task that combines the patches and labels from pairs of ROIs together to exploit the instance attention model's capability to learn from samples with multiple labels, which results in superior performance over several baselines.

Keywords: Weakly supervised learning, multisource classification, fine-grained object recognition, digital pathology, breast histopathology, region of interest classification.

ÖZET

UZAKTAN ALGILAMA VE TIBBİ GÖRÜNTÜ ANALİZİNDE ZAYIF DENETİMLİ GÖRÜNTÜ SINIFLANDIRMA YAKLAŞIMLARI

Bulut Aygüneş
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Selim Aksoy
İkinci Tez Danışmanı: Ramazan Gökberk Cinbiş
Aralık 2020

Zayıf denetimli öğrenme (ZDÖ), çeşitli öğrenme problemlerini çözmek için gürültülü etiketler içeren verilerden faydalanmayı amaçlamaktadır. Biz ZDÖ yaklaşımlarını uzaktan algılama ve tıbbi görüntü analizi olmak üzere iki farklı alanda incelemekteyiz.

Uzaktan algılama alanında, bir nesneyi birçok benzer alt kategoriden birine sınıflandırmayı amaçlayan çok kaynaklı ince taneli nesne tanıma problemine odaklanmaktayız. Bu problemde, verilen sınıf etiketine sahip bir nesne görüntüde bulunmakta, fakat nesnenin görüntü içindeki kesin konumu bilinmemektedir. Bu soruna ZDÖ perspektifinden yaklaşmakta ve iki paralel dal ile nesnelerin aynı anda lokalizasyonunu ve sınıflandırılmasını yapabilen tek kaynaklı bir derin öge dikkat modeli önermekteyiz. Daha sonra bu modeli, konum belirsizliği içermediği varsayılan bir referans kaynağın birden fazla kaynağın birleştirilmesine yardımcı olmak amacıyla kullanıldığı çok kaynaklı bir senaryoya uygun eklemelerle genişletmekteyiz. Önerilen tüm birleştirme stratejilerinin en gelişkin yöntemler ile karşılaştırıldığında daha yüksek doğruluk sağladığını ortaya koymaktayız. Ayrıca, her bir modeli çeşitli parametre karmaşıklık ayarlarında değerlendirerek derinlemesine bir karşılaştırma da yapmakta ve artan model kapasitesinin varsayılan kapasite ayarının daha üzerinde bir performans sergilediğini göstermekteyiz.

Tıbbi görüntü analizi alanında, meme biyopsi tüm slaytlarından gelen değişken şekil ve boyutlardaki ilgi bölgeleri (İB) üzerinde meme kanseri sınıflandırması problemini ele almaktayız. Bu problemle ilgili yaklaşımlar tipik olarak İB'lerden örneklenen sabit boyutlu pencerelerin sınıflandırma sonuçlarının

görüntü düzeyinde sınıflandırma puanları elde etmek için birleştirilmesi şeklinde olmaktadır. İlk olarak, İB'nin tamamının sınıflandırılması doğrultusunda, pencerelerden gelen bilginin komşu pencereler üzerinde kademeli bir şekilde iletilmesini amaçlayan bir çizge evrişimsel ağ (ÇEA) aracılığıyla yerel pencere arası bağlamdan faydalanmayı hedefleyen genel bir yöntem önermekteyiz. Zorlu bir veri kümesinde gerçekleştirdiğimiz 3 sınıflı bir İB sınıflandırma problemi üzerine yaptığımız deneylerde birkaç temel yaklaşımla yaptığımız kıyaslamalar sonucunda, önerilen modelin yaygın olarak kullanılan birleştirme kurallarından daha iyi performans sergilediğini göstermekteyiz. İkinci olarak, uzaktan algılama deneylerimizde kullandığımız ZDÖ sistemini yeniden gözden geçirip 4 sınıflı bir İB sınıflandırma problemine uygulamaktayız. Bu ZDÖ problemi için özel olarak tasarlanmış, İB çiftlerinden gelen pencere ve etiketleri birleştirerek öge dikkat modelinin birden çok etikete sahip örneklerden öğrenme yeteneğinden yararlanmayı amaçlayan yeni bir eğitim yöntemi önermekte ve bu önerilen yöntemin birkaç temel yaklaşıma kıyasla üstün performans sergilediğini gözlemlemekteyiz.

Anahtar sözcükler: Zayıf denetimli öğrenme, çok kaynaklı sınıflandırma, ince taneli nesne tanıma, dijital patoloji, meme histopatolojisi, ilgi bölgesi sınıflandırma.

Acknowledgement

I would like to express my sincere thanks to my advisor Assoc. Prof. Dr. Selim Aksoy for his guidance, support, and patience throughout my master's studies. I found the opportunity to gain valuable experiences and insights that will keep assisting me in my future career, all of which would not have been possible without him. I would like to extend my thanks to my co-advisor Assist. Prof. Dr. Ramazan Gökberk Cinbiş for all the support and advice he provided. His guidance helped me a lot to overcome many of the obstacles I faced throughout this study.

I would also like to thank the committee members of my thesis, Assist. Prof. Dr. Hamdi Dibeklioğlu and Assoc. Prof. Dr. Mehmet Erkut Erdem, for accepting my invitation to my defense and their feedback on this thesis.

I would like to express my gratitude and love to my family for their never-ending support throughout my life and my studies. I would also like to thank my friends Alperen, Argun, Buğra, Cahit, İpek, and Yalçın, who were always there for me during this tough journey.

I would also like to express my thanks to Assoc. Prof. Dr. Kemal Kösemehmetoğlu, Assoc. Prof. Dr. Sevgen Önder, and Prof. Dr. Ayşegül Üner for all the help and expertise they provided during the data collection and annotation process. None of the work I did for medical image analysis would have been possible without their help.

Finally, I would like to thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) for the support they provided with grants 117E172 and 116E445 throughout my master's studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Deep learning models have surpassed the state-of-the-art results of the traditional machine learning algorithms in many different computer vision tasks such as image classification [1], segmentation [2], and object detection [3]. These improvements are enabled by the architectural and computational improvements as well as the availability of large data sets with manual annotations. However, data annotation is generally an expensive, time-consuming, and error-prone process, making it difficult to create large data sets with high-quality labels, especially in fields such as medical imaging, where expert knowledge is required. Consequently, the data labeling process may result in data sets with crude, imprecise, or noisy annotations, which causes traditional supervised methods to be ineffective. The problem of learning from the data with labels of such characteristics is known as weakly supervised learning (WSL), a popular research area in many different fields in computer vision literature [4, 5, 6, 7, 8, 9].

One of the fields that WSL has many potential applications is remote sensing imagery. 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 [10, 11, 12]. 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 co-registered 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 [13]. However, for multisource tasks that involve the classification of relatively small objects such as trees—similar to the problem we focus on—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 distortions [14].

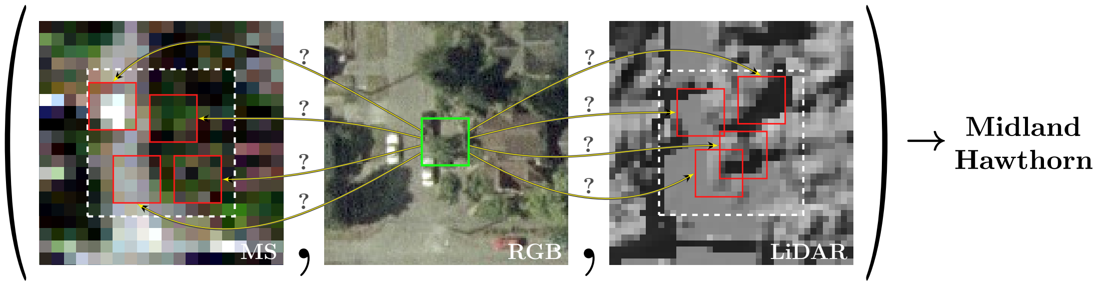

Figure 1.1: Illustration of our multisource fine-grained object recognition problem. The sources are only approximately registered, therefore, it is unclear which pixels in the low-resolution sources (MS and LiDAR) correspond to the object of interest centered in the high-resolution reference source (RGB). Our goal is to implicitly tackle the registration uncertainties through instance attention to correctly predict the object class using information from all sources.

The fine-grained recognition problem we studied in this thesis involves the classification of street trees using RGB, MS, and LiDAR images. 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 images 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 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 1.1.

To the best of our knowledge, the only related work that studied the multisource fine-grained recognition problem is that of [15], 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 thesis, 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 we propose loosely builds upon our preliminary work [16], which has shown that weakly supervised learning objective can be repurposed to improve single-source object recognition when the images contain high location uncertainty. As one of 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.

Another area we study in this thesis is the field of medical image analysis, with our specific focus being multi-class breast cancer classification in breast biopsy images. Breast cancer patients can face a variety of clinical actions such as surgery, radiation, or hormonal therapy depending on the diagnosis made by the pathologists for the biopsy samples. Different types of proliferations in the tissue structures carry different risks of progressing into malignancy; thus, the accuracy of the diagnosis in a fine-grained multi-class setting becomes critical.

Histopathological image analysis aims to serve as an important tool for helping pathologists with the diagnostic process. Deep learning-based approaches, in particular convolutional neural networks (CNN), have been shown to be successful in image classification tasks from various domains including digital pathology [17]. Whole slide images (WSI) that are obtained by digitizing biopsy slides at high magnification often contain many regions of interest (ROI) that can belong to different diagnostic categories and can carry different levels of relevance for the slide-level diagnosis. Furthermore, the pathologists do not have any restrictions on the ROI size when they evaluate the slides, and can select and study the regions at any size and magnification deemed suitable. Therefore, multi-class classification of arbitrarily sized ROIs appears to be an important problem that serves as a necessary step in the diagnostic process of breast cancer.

We have been studying this problem in a WSL perspective where the contributions of the individual patches to the ROI-level diagnosis are not known during training. We proposed a generic feature representation for arbitrarily sized ROIs by using weighted aggregation of the feature representations of fixed-sized patches sampled from these ROIs [18]. Both the patch-level feature representations and the weights were obtained from a convolutional network trained on patches sampled from ROIs in the training data. Similarly, representations like bag-of-words or Fisher encodings are also suitable methods to obtain feature representations for arbitrarily sized ROIs. However, all aforementioned ROI feature representations and aggregation methods based on patch classification results move from the patch level to the ROI level without exploiting the spatial information within the neighborhoods of the patches.

The spatial context formed by individual patches towards their collective contribution to the ROI-level diagnosis remains to be an important detail that has not been studied in earlier work in this domain. As our first contribution to the field of breast histopathology image analysis, we propose to incorporate local context through a graph-based ROI representation over a variable number of patches and their spatial proximity relations. More specifically, we formulate the ROI classification problem as a graph classification problem where vertices denote the patches sampled from a given ROI and edges represent the spatial proximity of those patches. The graph structure, therefore, implicitly encodes the spatial relationships across the patches, which can be used to tackle fine-grained ROI classification in a much more holistic manner compared to mainstream patch-classification based approaches. A preliminary version of this work was presented in [19].

To realize a classification model for the proposed ROI graph representation, we propose a graph convolutional network (GCN) based classification model. Our proposed GCN architecture extracts per-patch representations, propagates information over the neighboring patches in a progressive manner to incorporate the spatial context, and finally, aggregates the resulting patch representations to classify the whole ROI into a diagnostic class. Our experimental results demonstrate the power of the proposed GCN architecture over a number of strong baselines.

Next, we assume an alternative approach to the ROI classification problem regarding the aggregation of patch-level information towards making an ROI-level decision. We first highlight that each patch sampled from an ROI only carries information regarding a very limited context. Given that a patch may carry only a portion of larger structures that should be examined as a whole to come up with a diagnosis, the assignment of a patch to one of the diagnostic classes does not appear to be a plausible approach. Furthermore, since no information other than an ROI-level label is present, there is no way to make safe assumptions about the impact of a patch for the ROI-level decision. A patch can be highly informative, carrying parts of important structures relevant for the diagnosis, or be misleading or confusing, consisting of background or uninformative, common tissue. Considering these issues arising from the lack of information in the patch-level due to the limited context, an ROI-level diagnosis can only serve as a weak label for the patches.

Considering the aggregation of patch-level information as a problem of attending to the patches relevant for the diagnosis and combining the information from these patches to make an ROI-level decision, we propose to apply a similar WSL framework that we use in our remote sensing experiments so that we can utilize ROI-level labels towards learning a model with instance attention behavior. As another contribution of this thesis, we show that such a WSL approach achieves results comparable to or better than several baselines. Furthermore, we propose a novel training strategy inspired by the discussion on the weak labels, consisting of combinations of pairs of ROIs and their reference labels. While the instance attention model without the proposed training methodology can only learn from the patches of an ROI which are relevant to the given ROI label, this proposed technique enables the model to learn from patches which are more relevant to another class, when such patches exist in the ROI, with the help of the combined reference labels. We show that this strategy utilizing the instance attention model's ability to learn from multi-label samples significantly improves the model performance, further strengthening our weak label assumptions.

To sum up our contributions in this thesis, first, we showed that a WSL-based instance attention approach can improve single-source fine-grained object recognition in remote sensing imagery by mitigating the high uncertainty regarding the locations of the objects in the images. Then, we extended this approach to a multisource setting with high uncertainty among the sources by proposing four novel fusion schemes that combine the sources in probability level, logit level, feature level, and pixel level. We also proposed a method to compare different models in a principled way by evaluating each model in different capacity settings, showing that a single capacity setting can hinder making fair comparisons across methods. As our next contribution, we tackled breast cancer classification and proposed a generic graph-based ROI representation encoding the spatial proximity information of patches sampled from the ROI. We proposed a GCN-based classification model that uses this representation to utilize the spatial context formed by the patches, resulting in a better ROI-level classification performance over several baselines. Finally, we reformulated the problem of aggregation of patch-level information from a WSL perspective to benefit from a similar instance attention model we used in our remote sensing experiments and further improved its performance with a novel training scheme designed for this WSL methodology exploiting the noisy nature of patches and multi-label learning capabilities of the instance attention model.

Chapter 2

Related Literature

2.1 Related Work in Remote Sensing

2.1.1 Multisource Image Analysis

There are many studies in the remote sensing literature that focus on multisource image analysis [20, 21], which has also received the attention of data fusion contests [22, 23, 24, 25]. The research includes statistical learning methods such as dependence trees [26], kernel-based methods [27], copula-based multivariate model [28], and active learning [29]. Another well-studied problem is manifold alignment [30, 31, 32] 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 thesis, 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 images preprocessed to the same resolution, followed by a convolutional neural network (CNN) [33], while in feature-level fusion, hand-crafted [34] or CNN-based [35, 36, 37, 38] features, obtained 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.

2.1.2 Weakly Supervised Remote Sensing

Similar WSL problems have been studied in the remote sensing literature. For example, Han *et al.* [39] presented a binary object detection method where positive instances for the object of interest were sampled based on saliency scores while trying to keep inter-class separability and intra-class compactness high. New positive training samples were selected using a classifier trained on the current training set, and a new classifier was trained on the updated training set in an iterative fashion. Zhang *et al.* [40] followed a similar iterative instance mining procedure by updating the negative instance set and the classifier in each iteration.

These WSL approaches differ from ours in the sense that they use iterative instance mining to improve the classifier which is then used to detect objects in a sliding window fashion. Our method, instead, focuses on directly training a model that detects and classifies the object of interest given the whole input image, by using end-to-end learning from weakly labeled training instances. Sumbul *et al.* [41] presented an alternative approach in the same problem setting by using an attention mechanism that learns a weighted combination of features extracted from fixed-sized regions obtained at each possible position in the image for classification.

Some WSL approaches in remote sensing have utilized class activation maps for object localization. For example, Ji *et al.* [5] 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.* [42]

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.* [43] localized objects by using a combination of two different convolutional layers. Zhang *et al.* [44] suggested using gradients of network layers to obtain saliency maps for background and foreground classes. Similarly, Ma *et al.* [45] obtained saliency maps by utilizing gradients with respect to the input pixels to localize residential areas in aerial images. Ali *et al.* [46] 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. However, all of these approaches focus on a binary classification scenario. For the multi-class setting, Li *et al.* [47] 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.* [48] 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.

In a more relevant problem caused by misalignment of GIS maps and images used for building extraction, Zhang *et al.* [6] 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.

2.1.3 Tree Classification

In this thesis, 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 [11] 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.* [49] 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 thesis are not specific to tree detection. Thus, a full review on tree species mapping is beyond the scope of this work. We refer the reader to [50] that provides a review of such methods.

2.2 Related Work in Breast Histopathology

As the mainstream CNN architectures for image classification typically require fixed-sized inputs, their common use in the digital pathology domain has also been in the classification of fixed-sized biopsy image patches. For example, commonly used data sets [51, 52] include fixed-sized images that are manually selected from biopsy samples with the goal of preparing benchmarks in image classification competitions. The generally studied setting has been to aggregate the classification results of fixed-sized patches cropped from these images to obtain an image-level classification score. Locations of these fixed-sized patches are usually chosen from a regular grid [53, 54], although some studies focus on a nuclei guided selection [55] or random sampling from a foreground mask instead [56]. Aggregation methods typically include fixed fusion rules such as averaging class probabilities or majority voting [53, 57, 58, 59].

GCN models have been previously shown to be effective in the utilization of the spatial context in visual inputs [60, 61]. However, although more traditional graph-based approaches using Delaunay triangulations, Voronoi diagrams, or similar structures are common in the digital histopathology [62, 63], the application of graph neural networks is limited, with the available work utilizing GCNs to process graphs constructed based on nuclei detected from the input [64, 65, 66, 9].

Various WSL approaches are proposed in the medical imaging field to deal with problems such as mitosis [67], lung cancer [68], and prostate cancer detection [69, 9]. Among the WSL studies focusing specifically on breast cancer, Shen *et al.* utilized an attention mechanism for weakly supervised localization problem on breast mammography images [70]. In a similar setting with breast mammograms, Liang *et al.* used class activation maps for weakly supervised localization of breast cancer [71]. Lu *et al.* also employed an attention mechanism in a WSL setting on the whole slide level for various classification problems, including breast cancer classification [8]. Other examples of usage of attention mechanisms in breast histopathology are present in the literature as well [72, 73], even though the problems they approach are not WSL problems.

Another problem closely related to WSL is multiple instance learning (MIL), which is the problem of learning from a bag-of-instances carrying a single label. Campanella *et al.* [69] is one of the works that studied different cancer classification problems from a WSL and MIL perspective on the slide level by learning to rank patches cropped from the slide to disregard the ones with lower ranks. Aggregation of the remaining patches to obtain a slide-level prediction was carried out using a recurrent neural network (RNN). Ilse *et al.* utilized attention-based pooling of patches cropped from fixed-sized ROIs by modeling the ROI as a bag-of-patches in an MIL perspective similar to our approach [74]. Li *et al.* used a similar attention-based bag-of-patches approach in a multi-stage fashion for both localization and grading of cancer in the slide level [75].

Chapter 3

Data Sets

3.1 Data Set for Remote Sensing Experiments

We conduct our experiments on the same data set as [15], 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 number of samples for each class in this highly imbalanced data set is shown in Table 3.1. The fine-grained nature of the data set is illustrated in Figure 3.1 where the scientific classification of tree species is presented as a hierarchy in which most species differ only in the lowest levels.

For each tree sample, there are three images obtained from different sources: an aerial RGB image, an 8-band WorldView-2 MS image, and a LiDAR-based digital surface model with 1 foot, 2 meter, and 3 foot spatial resolution, respectively. The RGB images are centered at the locations provided in the point GIS data [77], and are cropped at a size of 25×25 pixels to cover the largest tree in the data set. Furthermore, each RGB image is visually inspected to eliminate erroneous samples that result from temporal differences between ground data collection and

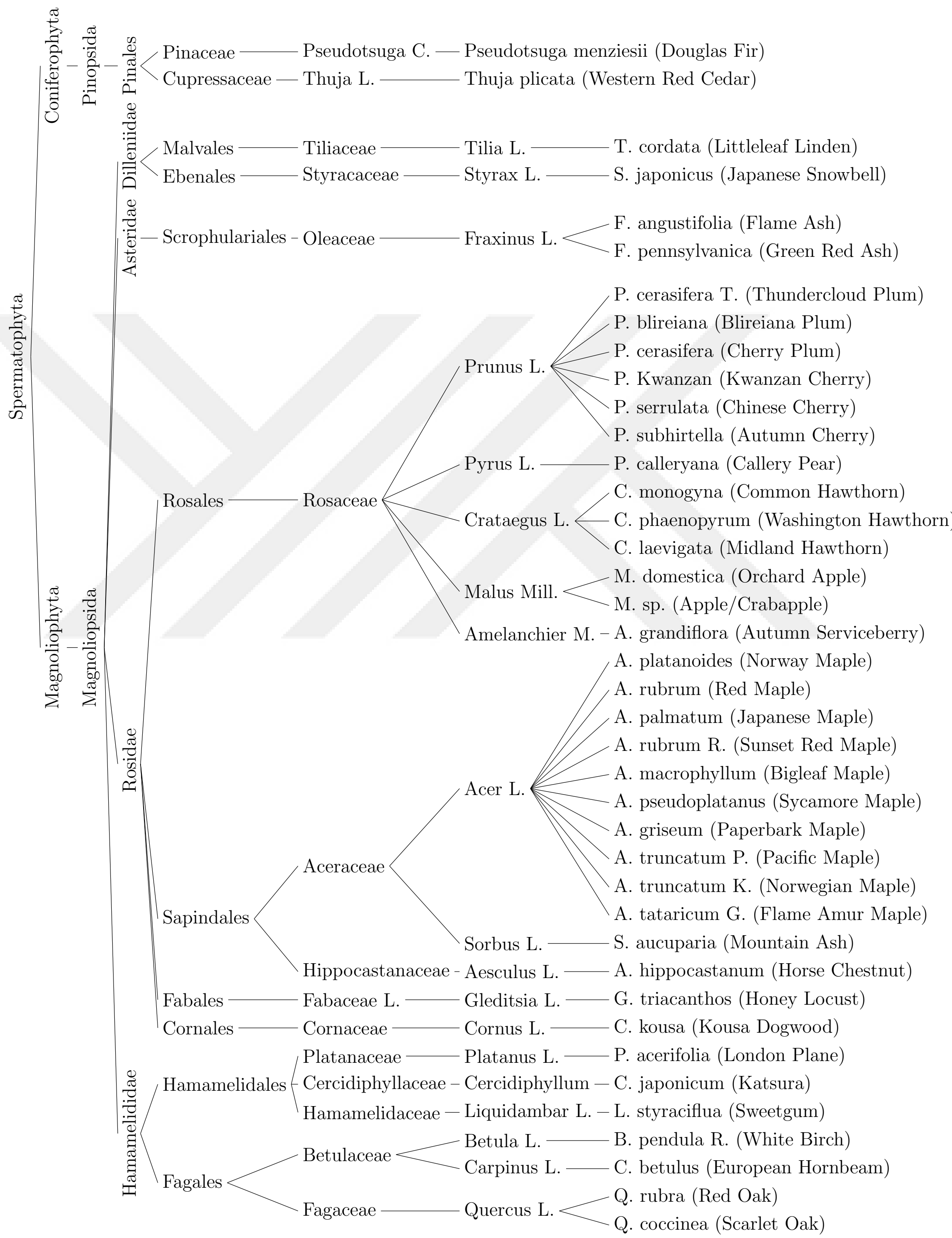

Figure 3.1: Scientific classification of tree species. The taxonomy starts with the Spermatophyta superdivision and continues with the names of division, class, subclass, order, family, genus, and species in order. At the last level, the common names are given in parentheses next to the scientific names. The classifications are taken from [76].

Table 3.1: Class names and number of samples in each class in the data set.

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

aerial data acquisition [12]. This, together with the fact that RGB images have higher spatial resolution than MS and LiDAR images, make RGB images suitable to be used as the reference source.

A tree in a 25×25 pixel RGB image corresponds to 4×4 pixels in MS and 8×8 pixels in LiDAR images. Although each source is previously georeferenced, registration errors can cause significant uncertainties in the locations of small objects such as trees, especially in images 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 images for MS and 24×24 pixel images for LiDAR [15].

3.2 Data Set for Breast Histopathology Experiments

3.2.1 Data Set Description

We constructed a new data set that currently contains 1,376 ROIs annotated within 121 WSIs that were digitized from haematoxylin and eosin stained specimens belonging to 99 different patients. The specimens were selected from the archives of the Department of Pathology at Hacettepe University based on their slide-level diagnoses. The WSIs were acquired at 40× magnification by using an Olympus slide scanner, resulting in an average image size of 170,000 × 132,000 pixels. The ROIs were annotated by experienced pathologists in free form with no restriction in the sizes and shapes of the image masks. The resulting annotations were collected into 4 diagnostic classes: *benign* (including samples containing non-proliferative changes, apocrine metaplasia, usual ductal hyperplasia, columnar cell hyperplasia, flat epithelial hyperplasia, and intraductal papilloma without atypia), *atypia* (including samples containing atypical ductal hyperplasia, atypical lobular hyperplasia, and intraductal papilloma with atypia), *in situ* carcinoma (including both ductal carcinoma in situ and lobular carcinoma in situ), and *invasive* carcinoma. The experiments using this data set are conducted at 10× magnification in Section 5.2 and at 5× magnification in Section 5.3. The class-specific ROI size statistics for 10× magnification in Table 3.2 show a high variation for the samples in the data set.

Table 3.2: ROI size statistics per diagnostic class in number of pixels at 10× magnification. Rows show the average ROI size, the standard deviation of ROI sizes, and the ratio of the largest ROI size to the smallest one, respectively.

	Benign	Atypia	In Situ	Invasive
Average	$1966K$	$473K$	$4227K$	$13528K$
Standard deviation	$4207K$	$687K$	$8325K$	$19263K$
Max-min ratio	2216.0	704.5	1989.1	762.5

3.2.2 Stain Normalization

Since the specimens were prepared at different times, they have a high variation in their staining. Thus, we performed stain normalization by matching the histograms of the hematoxylin and eosin channels of each slide to the hematoxylin and eosin histograms of a target slide chosen from the data set [78]. To obtain hematoxylin and eosin histograms, we applied color deconvolution [79] to each slide using a unique stain matrix estimated for that slide. Hematoxylin stain vector estimation was carried out by computing the median of the pixels inside the nucleus mask of the slide in the optical density space, separately for red, green, and blue channels. For eosin, the median was computed over a mask obtained by eliminating the nuclei and high luminosity regions. We estimated the nucleus masks using a pre-trained convolutional network.

3.2.3 Data Set Partitioning with Genetic Algorithm

Finally, we partitioned the data set into four folds by using ROI-level diagnostic labels by making sure that each fold has slides (and ROIs) corresponding to independent patients. To achieve both slide-level and ROI-level diversity among the folds, we employed a genetic algorithm which rewards splitting the data set into similar numbers of ROIs and slides per class within each fold. Two randomly chosen folds are combined as the training set, whereas validation and test sets are randomly selected among the remaining folds. The ROI-level and slide-level class distributions of the three sets are given in Table 3.3. Since each slide has a different number of ROIs, that may also have a set of labels different from that of the slide, the resulting data set has a heavy class imbalance.

Table 3.3: Class distribution of slides and ROIs in training, validation, and test sets. Note that a slide can contain multiple ROIs corresponding to different diagnostic labels, resulting in a multi-label setting for each slide. Thus, the numbers of slides for each diagnostic class in the table do not sum up to the total number of slides for a given set. We focus on ROI-level classification in this thesis.

		Benign	Atypia	In Situ	Invasive	Total
Slide	Training Set	42	20	22	15	53
	Validation Set	18	11	10	7	23
	Test Set	20	10	11	7	26
	Total	80	41	43	29	102
ROI	Training Set	291	73	192	118	674
	Validation Set	147	38	128	59	372
	Test Set	144	35	95	56	330
	Total	582	146	415	233	1376

Chapter 4

Weakly Supervised Fine-grained Object Recognition in Remote Sensing

4.1 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) [80]. 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 4.1 for source-specific architectural details of the region encoder. For input $x \in X$,

$$\phi^{enc} : X \rightarrow \Omega \tag{4.1}$$

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:

$$[\sigma^{loc}(\omega)]_{ci} = \frac{\exp\big([\phi^{loc}(\omega)]_{ci}\big)}{\sum_{r=1}^{R} \exp\big([\phi^{loc}(\omega)]_{cr}\big)} \tag{4.2}$$

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:

$$[\sigma^{cls}(\omega)]_{ci} = \frac{\exp\big([\phi^{cls}(\omega)]_{ci}\big)}{\sum_{k=1}^{C} \exp\big([\phi^{cls}(\omega)]_{ki}\big)} \tag{4.3}$$

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. [15]. 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:

$$[\phi^{pred}(\omega)]_c = \sum_{i=1}^{R} [\sigma^{loc}(\omega)]_{ci} \odot [\sigma^{cls}(\omega)]_{ci}. \tag{4.4}$$

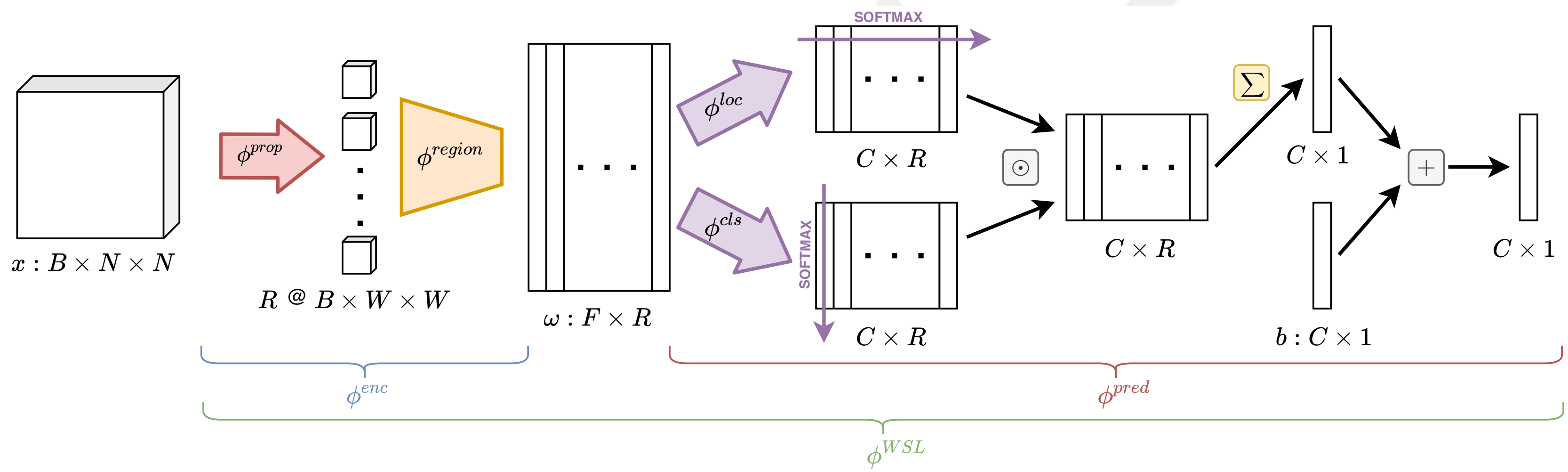

Figure 4.1: Illustration of the single-source weakly supervised instance attention model. Source index m is omitted from all variables for simplicity. For the m^{th} source, the model takes x_m with size $B_m \times N_m \times N_m$ as the input. Region proposals are extracted using ϕ^{prop} with a sliding 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.

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 (4.4) as follows:

$$[\phi^{pred}(\omega)]_c = \sum_{i=1}^{R} [\sigma^{loc}(\omega)]_{ci} \odot [\sigma^{cls}(\omega)]_{ci} + b_c \tag{4.5}$$

where b_c is the bias parameter for class c. Combining (4.1) and (4.5), we define the whole model as:

$$\phi^{WSL}(x) = \phi^{pred}\big(\phi^{enc}(x)\big), \tag{4.6}$$

and class probabilities as:

$$P(c|x) = \big[\sigma\big(\phi^{WSL}(x)\big)\big]_c \tag{4.7}$$

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

4.2 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 is $N_m \times N_m$ pixels 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.

In this thesis, we focus on RGB, MS, and LiDAR images which are acquired in different imaging 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.

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 4.2 illustrates the proposed models.

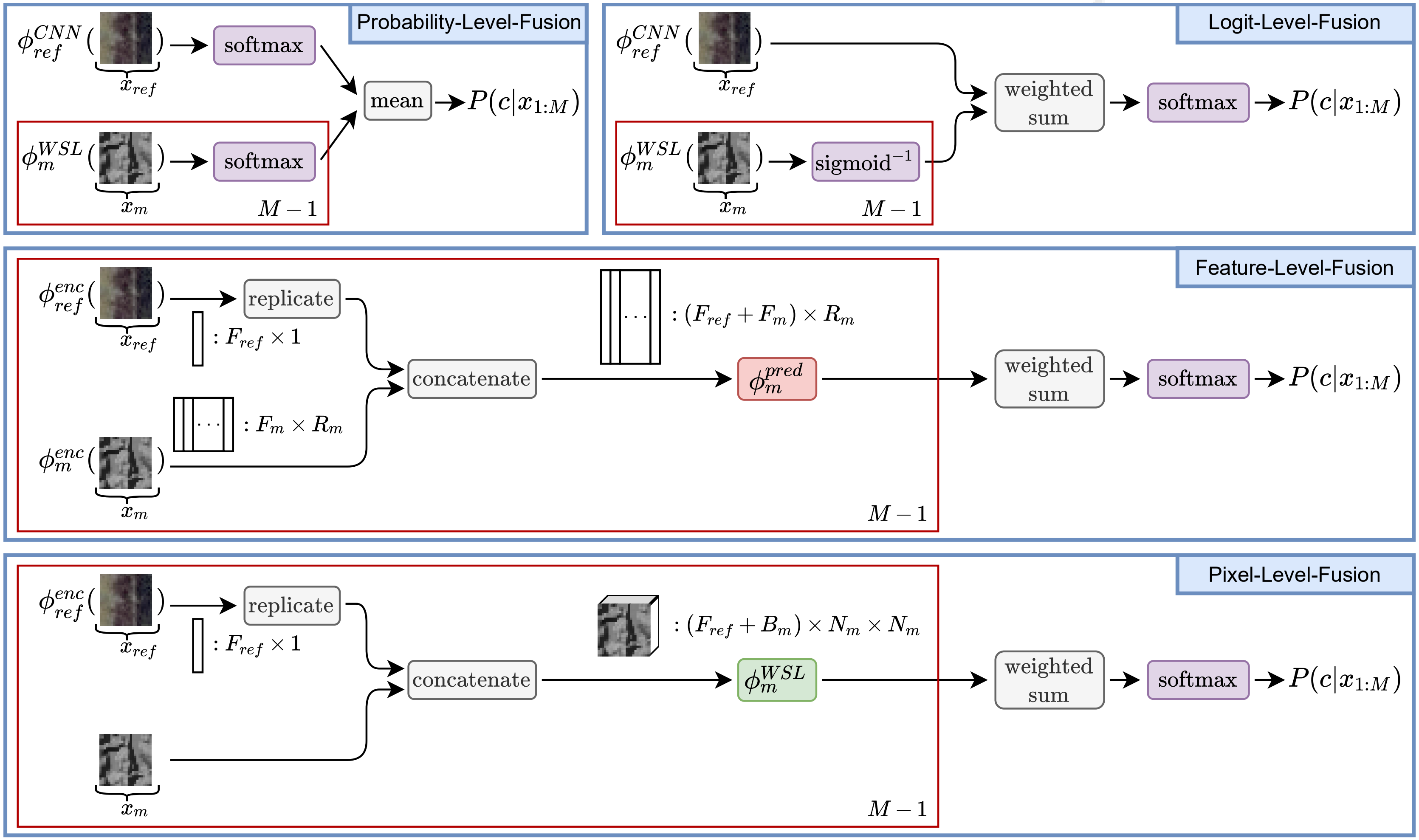

Figure 4.2: Illustration of the proposed multisource WSL models. The reference source x_1 is represented with subscript ref, while subscript $m \in \{2 \dots M\}$ is used for the additional sources. Plate notation is used to represent the repetitions in the model. Variables are described in the text.

4.2.1 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:

$$P(c|x_{1:M}) = \frac{1}{M}\sum_{m=1}^{M} P(c|x_m) \tag{4.8}$$

where $P(c|x_m)$ is obtained as in (4.7) 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 (4.7) 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 $P(c|x_1)$.

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.

4.2.2 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:

$$\phi^{comb}(x_{1:M}) = \alpha_1 \phi_{ref}^{CNN}(x_1) + \sum_{m=2}^{M} \alpha_m S^{-1}\big(\phi_m^{WSL}(x_m)\big) \tag{4.9}$$

$$P(c|x_{1:M}) = \big[\sigma\big(\phi^{comb}(x_{1:M})\big)\big]_c \tag{4.10}$$

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 (4.9) are obtained using softmax over learnable parameters β_m:

$$\alpha_m = \frac{\exp(\beta_m)}{\sum_{i=1}^{M}\exp(\beta_i)}. \tag{4.11}$$

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.

4.2.3 Feature-Level-Fusion

For each additional source m, we propose to combine penultimate layer feature vector of the reference network $\phi^{enc}_{ref}(x_1)$ with the candidate region feature representations of each additional source $\phi^{enc}_m(x_m)$. For this purpose, we replicate $\phi^{enc}_{ref}(x_1)$ R_m times, and concatenate with $\phi^{enc}_m(x_m)$ to 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^{pred}_m 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:

$$\phi^{comb}(x_{1:M}) = \sum_{m=2}^{M} \alpha_m \phi^{pred}_m\big(\psi(\phi^{enc}_{ref}(x_1), \phi^{enc}_m(x_m))\big) \tag{4.12}$$

where ψ denotes the aforementioned replication and concatenation operations. Class probabilities are obtained using (4.10). 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.

4.2.4 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 [15]. 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:

$$\phi^{comb}(x_{1:M}) = \sum_{m=2}^{M} \alpha_m \phi_m^{WSL}\big(\psi(\phi_{ref}^{enc}(x_1), x_m)\big), \tag{4.13}$$

which is followed by (4.10) 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 target 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.

4.3 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.

4.3.1 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 hyperparameters.

For all the 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 [15]. 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.

4.3.2 Implementation Details

4.3.2.1 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 [15]. 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.

4.3.2.2 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 4.3.

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 (4.7) becomes:

$$P(c|x) = \left[\sigma\left(\phi^{WSL}(x)/T\right)\right]_c. \tag{4.14}$$

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

4.3.2.3 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

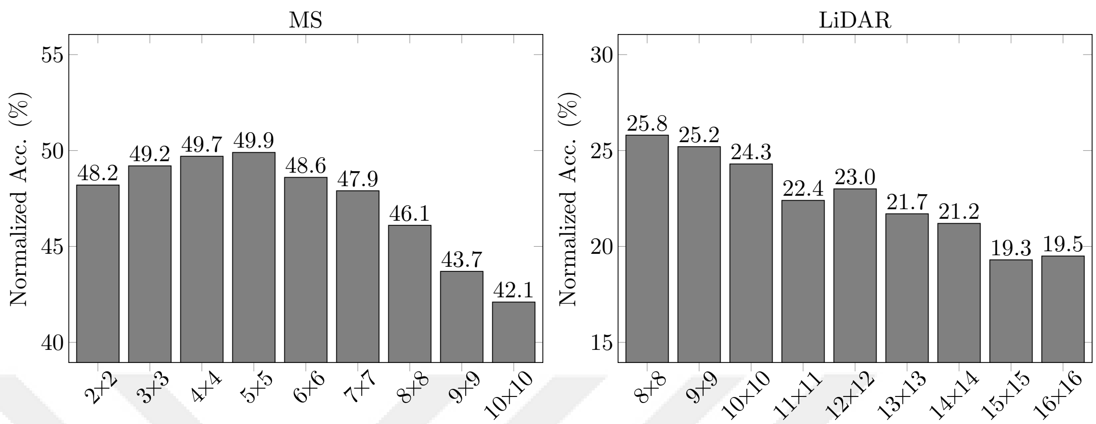

Figure 4.3: Impact of region proposal size (W) on normalized validation accuracy. Proposals are extracted within 12×12 pixel images for MS and 24×24 pixel images for LiDAR as described in Section 3.1

(T_m) on the validation set via grid search, resulting in 1/48 for MS and 1/18 for LiDAR for the fused model.

4.3.2.4 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 (4.11) 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.

4.3.2.5 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 (4.12) 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 (4.11) 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.

4.3.2.6 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.

4.3.3 Results

4.3.3.1 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 baseline models. We note that RGB is the reference high-resolution source and contains centered object instances, therefore, instance attention is not applicable to RGB inputs. 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.

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 performs significantly better

Table 4.1: 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
Instance attention, ϕ^{cls} only (MS)	47.7
Instance attention (MS)	48.3
Single-source baseline (LiDAR)	21.2
Instance attention, ϕ^{cls} only (LiDAR)	24.3
Instance attention (LiDAR)	25.3

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. Single-source results are summarized in Table 4.1.

4.3.3.2 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

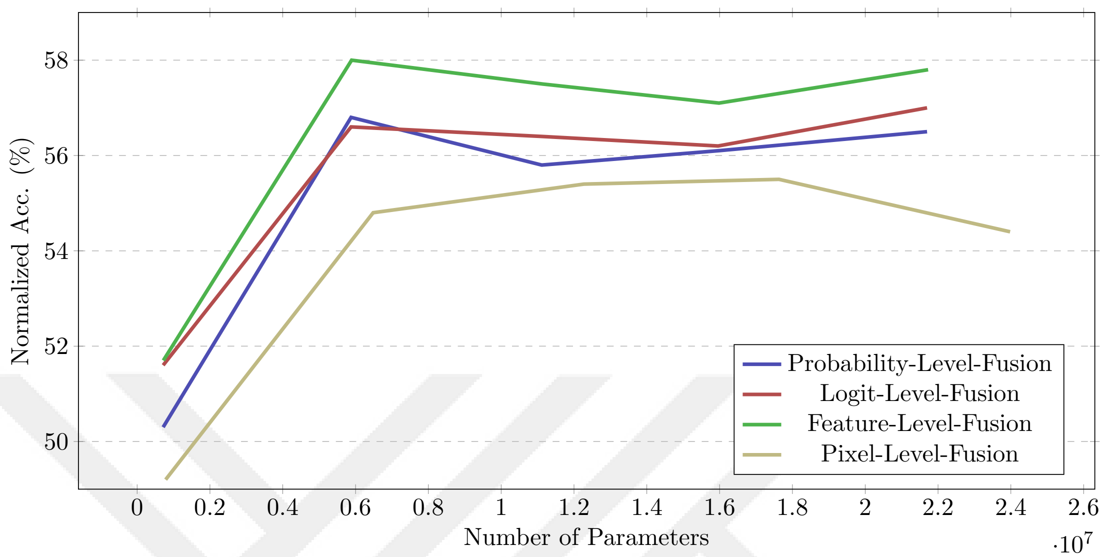

Figure 4.4: Effect of width-wise increasing the number of parameters, hence the model capacity, on normalized test accuracy for RGB & MS. This analysis aims to make a fair comparison among the multi-source instance attention formulations.

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 [15].

Figure 4.4 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 models are compared in the literature. First, evaluating a model in a single capacity setting might yield sub-optimal

Table 4.2: Multi-source 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 [15]	39.1
Recurrent attention model [81]	41.6
MRAN [15]	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 [15]	41.4
Recurrent attention model [81]	42.6
MRAN [15]	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

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 4.2, 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.

4.3.3.3 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 [15]. The second method is the *recurrent attention model* [81]. 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 *Multi-source Region Attention Network* (MRAN) [15], which has been shown to be an effective method for multisource fine-grained object recognition. MRAN extracts candidate regions from an MS and/or LiDAR image 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 4.2 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 4.2 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 4.4 and both RGB & MS and RGB, MS & LiDAR settings in Table 4.2 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.

4.3.3.4 Class-Specific Results

Figure 4.5 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

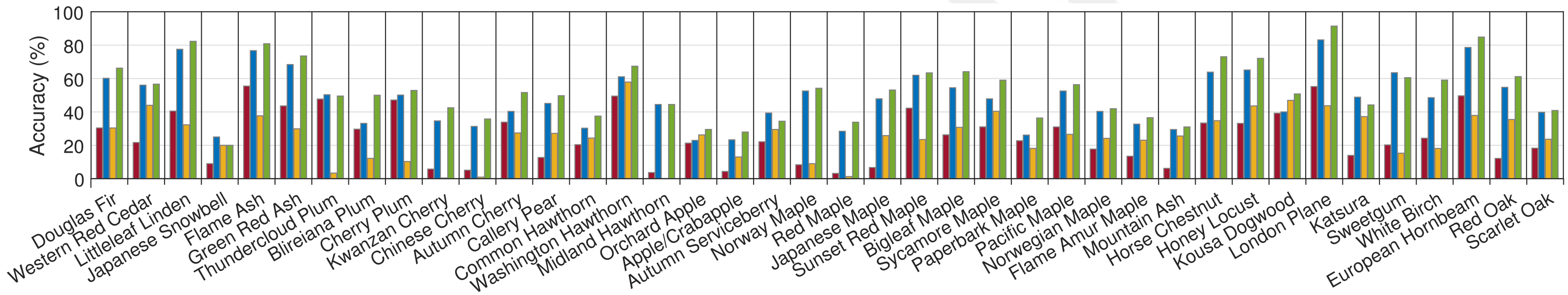

Figure 4.5: Example results for the class-specific performances of the proposed methods. From left to right: single-source RGB (red), single-source MS with instance attention (blue), single-source LiDAR with instance attention (yellow), and Feature-Level-Fusion (RGB, MS & LiDAR) (green). Best viewed in color.

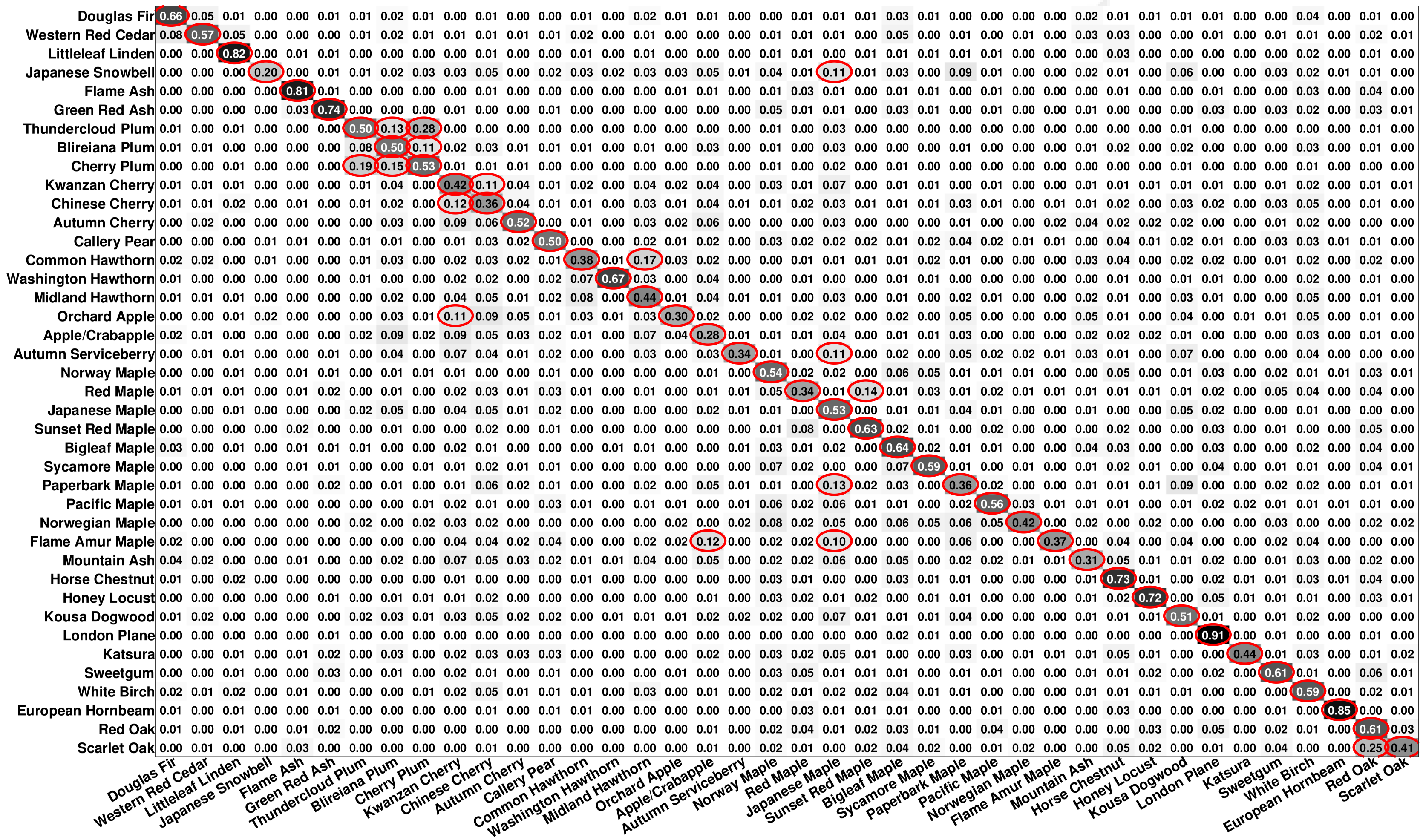

Figure 4.6: Confusion matrix for Feature-Level-Fusion (RGB, MS & LiDAR). Values represent percentages where each row sums to 1. The numbers greater than 0.10 are marked with red ellipses. The ones on the diagonal show correct classification, whereas off-diagonal entries indicate notable confusions.

in the bottom-most row of Table 4.2 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 4.6 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 3.1. 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 [12] 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.

4.3.3.5 Qualitative Results

Figure 4.7 illustrates the region scores, normalized to the $[0, 1]$ range, obtained by multiplying per-region classification and localization scores in the Hadamard product in (4.4) for the predicted class. These results show that all proposed methods are capable of localizing the target objects. However, the region scores for MS tend to have a smoother distribution with mostly a single local maximum, while LiDAR scores appear 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.

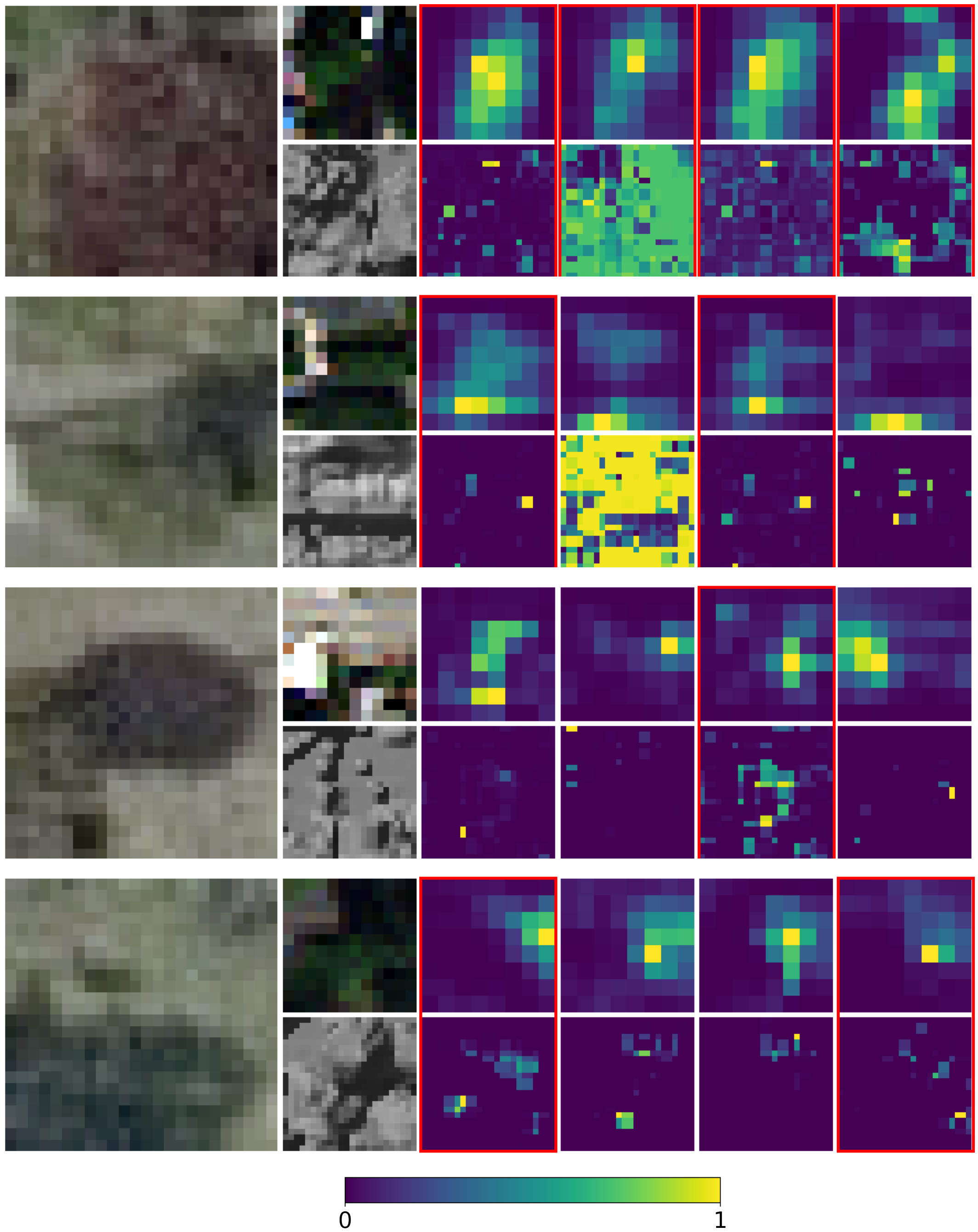

Figure 4.7: Region scores for sample test images. RGB images are shown in the first column, MS (top) and LiDAR (bottom) images shown in the second. Remaining columns show instance attention results respectively for Probability-Level-Fusion, Logit-Level-Fusion, Feature-Level-Fusion, and Pixel-Level-Fusion. Results for correct class predictions are denoted with 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.

Chapter 5

Weakly Supervised ROI Classification in Whole Slide Breast Histopathology Images

5.1 Problem Definition

The ultimate goal is to classify a given ROI image of arbitrary size into one of the diagnostic classes. During training, we have access to example ROI images and their ROI-level class annotations. Each training ROI sample is associated with one particular class label; however, not all patches inside an ROI homogeneously belong to the same class. Furthermore, diagnosing an ROI may require examining a context bigger than a patch to fully utilize the structural information in the ROI which can cover large areas. Thus, a patch can be uninformative, containing clutter or diagnostically insignificant structures or misleading or confusing, containing structures common in more than one diagnostic classes. One interpretation of this is that it is not always possible and meaningful to assign a patch to one of the diagnostic categories. However, it is also possible for a patch to be highly informative, covering the most diagnostically relevant portions of the ROI.

Therefore, the goal is to learn the classification model in a weakly-supervised manner over the noisy patches and ROI-level annotations without having access to patch-level labels, with each ROI being a mixture of both relevant and irrelevant patches.

5.2 Graph Convolutional Networks

5.2.1 Methodology

In this approach, we tackle the ROI classification problem as a graph classification problem, where vertices represent patches and edges represent spatial relations across the patches. In this manner, we aim to aggregate information from patches and admit arbitrary-sized ROIs in a principled manner. For this purpose, we construct an ROI graph, by first regularly sampling fixed-sized patches from the ROI, and associating each vertex with the corresponding image patch. Then, we add a binary edge between each pair of patches that are within a pre-defined proximity threshold ϵ. This leads to a sparse binary ROI graph. Example graphs for several ROIs are shown in Figure 5.1.

We propose a graph convolutional network (GCN) for the ROI graph classification problem. The first part of the network employs a ResNet-50 [1] convolutional sub-network that extracts a fixed-length (2048-dimensional) feature vector for each vertex (*i.e.,* patch). In order to propagate information across the patches and incorporate local contextual information, we apply two consecutive GCN layers. Here, we incorporate the GCN layer definition introduced by Kipf and Welling [82]. According to this definition, the GCN layer first calculates the weighted sum of the feature vectors of a vertex and the neighboring vertices of that vertex. Here, the GCN aggregation weights are induced by the symmetrically normalized adjacency matrix [82]. Then, a linear transformation followed by a nonlinear activation function is applied to these aggregated feature vectors to obtain per-vertex output vectors of the GCN layer. While the first GCN layer

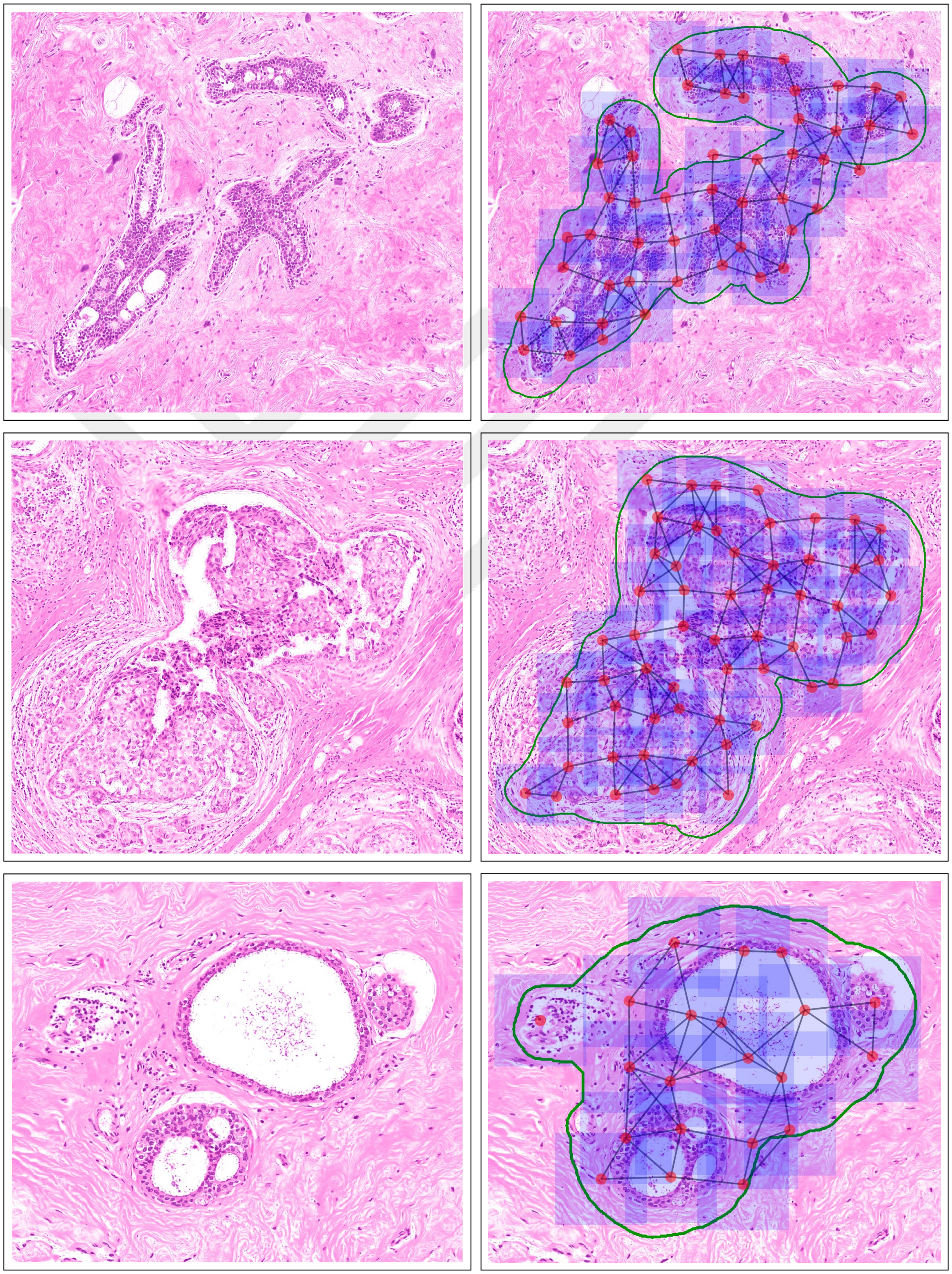

Figure 5.1: Sample ROIs (left) and the constructed ROI graphs (right). ROI boundaries drawn by the pathologists are shown in green and the sampled patches are displayed in blue. ROI graphs constructed with the proximity threshold ϵ chosen as 200 pixels are overlaid with vertices in red and edges in black.

transforms each vertex into a 256-dimensional vector, the second one results in 4-dimensional vectors, which can be interpreted as internal per-vertex classification scores. Finally, we apply global average pooling over all resulting vertex vectors to obtain a fixed-length representation for the whole graph, and compute a soft-max to obtain the final ROI classification probabilities. The model is summarized in Figure 5.2.

Our training pipeline consists of two separate training steps: first, training a patch classifier network to be used as a feature extractor; second, training the GCN layers for ROI classification. For the first step, we label each patch extracted from an ROI with the annotation of that ROI. Using these labels, we fine-tune the ResNet sub-network which is initialized via a classification model pre-trained on the ImageNet data set. In the second step, we train GCN layers with randomly initialized parameters using the features extracted by the ResNet while the ResNet parameters are kept frozen. The training is carried out in a weakly-supervised fashion, purely based on ROI class labels, with no patch-level annotations. For this purpose, we minimize the negative log-likelihood of the true class label of each ROI example through stochastic gradient-descent.

5.2.2 Experiments

In this section, we present two sets of experimental settings and results. The first setting is based on the experiments we conducted in one of our published works [19], where we tackled the 3-class ROI classification problem with a preliminary version of our data set. In the second set of experiments, we return to the 4-class setting with the current version of the data set and experiment with additional graph-based architectures. For each setting, we provide the training details of the patch-level feature extraction network and the GCN-based contextual classification network, define the baseline methods for empirical comparison, and present our experimental results.

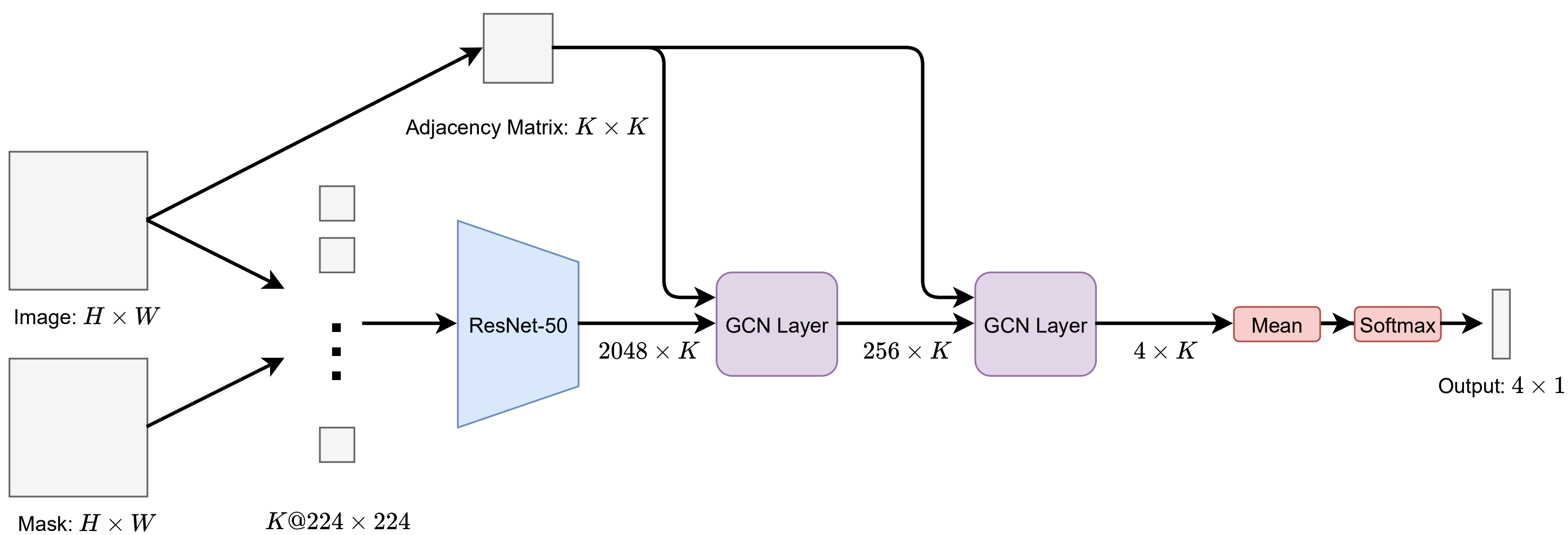

Figure 5.2: The proposed ROI graph classification architecture. K represents the number of patches (*i.e.*, vertices) inside the ROI, which is represented by the input graph. 4 corresponds to the number of diagnostic classes. Input mask represents the ROI area within the input image.

Table 5.1: Class distribution of ROIs in training, validation, and test sets for the 3-class setting.

	Benign	In Situ	Invasive	Total
Training Set	226	154	102	482
Validation Set	109	56	50	215
Test Set	105	69	49	223
Total	440	279	201	920

5.2.2.1 Three-Class ROI Classification Experiments

The results reported in this section are based on the experiments in one of our published works [19], which were conducted on a previous version of the current data set, with less samples from each of the classes. In our preliminary experiments with the data set in [19], we observed that consistently all methods performed poorly on the atypia class. This is likely to be a result of the fact that there were relatively few ROIs belonging to this class in our data set that was acquired from whole slides sampled from routine clinical practice. In addition, distinguishing atypia from in situ or benign cases appears to be a more challenging problem [83], which is also emphasized in other related work that do not include samples from the atypia class [52, 63]. Following these observations, we excluded the ROIs belonging to the atypia class, and conducted our in-detail experiments using the three remaining diagnostic classes: benign, in situ carcinoma, and invasive carcinoma. Training, validation and test distributions of the resultant data set is given in Table 5.1.

Experimental setup. We use two versions of the model described in Section 5.2.1. The version denoted as $\mathrm{GCN}_{\epsilon=0}$ uses the identity matrix as the adjacency matrix. This corresponds to the GCN layers acting as fully-connected classification layers without any spatial context information. The loss function that is used during the training of the model is still the same as the one described in Section 5.2.1. The version denoted as $\mathrm{GCN}_{\epsilon=200}$ uses a threshold of 200 pixels in construction of the adjacency matrix.

We additionally compare the performance of the proposed model with the following methods:

- Base-Penultimate: A patch-level feature representation is extracted directly from the penultimate layer activations of the fine-tuned convolutional network (ResNet) [56]. Then, the feature representations of the individual patches inside an ROI are aggregated by average pooling to obtain the feature representation of the ROI. Finally, a multi-layer perceptron (MLP) classifier is trained on the ROI-level feature representations and the labels in the training set.
- Majority-Voting: The fine-tuned convolutional network (ResNet) is used to assign a class label to each patch individually. Then, the class label for an ROI is obtained through majority voting of the class labels of its corresponding patches.
- Learned-Fusion: Similar to Majority-Voting, each patch is assigned a score for each class by the patch classification network (ResNet) [84]. Then, a class histogram for each ROI is constructed by summing over the class scores of its corresponding patches. Finally, the extracted histograms are used as feature vectors to be classified through an MLP.

In our comparisons, we use *normalized accuracy*, which is obtained by averaging per-class accuracy scores, to avoid biases towards classes with higher number of samples.

Implementation details. For training the ResNet-based patch-level feature extraction model, we sample patches of size 224×224 with an overlap of 74 pixels across consecutive patches at $10\times$ magnification. We over-sample patches from classes with fewer examples to reduce the effect of class imbalance. We apply random horizontal/vertical flips, random rotations of 90 degrees, and random hue jitter for data augmentation. We use the Adam optimizer with a learning rate of 5×10^{-7}, apply weight decay to all ResNet parameters with weight 10^{-3}, and use dropout [85] before the classification layer with 0.7 drop probability. Each batch contains 128 patches. After training the ResNet model in this manner, we keep its parameters frozen during the following GCN training stage.

For training the proposed GCN-based contextual ROI classification model, we keep the ResNet model parameters frozen and use pre-extracted patch descriptors in order to simplify training over ROIs with a variety of bounding aspect ratios and sizes. To obtain patch samples from an ROI, we use the same patch sampling and data augmentation techniques as in our aforementioned ResNet training approach with two exceptions: (i) we turn off random hue jittering (just to simplify the training pipeline implementation) and (ii) we sample additional patches by jittering the center coordinates of the original patch samples in both horizontal and vertical directions by random amounts. We obtain the random coordinate jitter values by sampling uniformly from the interval $[-45, 45]$. The size of the first GCN layer is chosen as 256 and the second one as the number of classes. We again use the Adam optimizer with a learning rate of 1×10^{-3}, apply weight decay to all GCN parameters with weight 10, and use dropout in the first GCN layer with 0.6 drop probability. We apply batch normalization after the first GCN layer. Each batch consists of 1024 ROIs chosen from the augmented training set.

A detail that deserves attention is the way to use dropout before a GCN layer: while applying dropout, we jointly process the patches within a single ROI and drop the same dimensions from their feature vectors instead of applying dropout to their descriptors independently. Otherwise (when applied independently), the effect of dropout is diminished due to local feature averaging in the following GCN layer. This is akin to using channel-wise *spatial dropout* [86] in convolutional layers to keep dropout effective without getting diluted due to the correlations across neighboring pixels.

Results. The experimental results are summarized in Tables 5.2 and 5.3. These results show that the proposed model using the graph convolutional layers in the non-degenerate graph setting ($\text{GCN}_{\epsilon=200}$) achieves a considerably higher accuracy than the no-edge case ($\text{GCN}_{\epsilon=0}$), with an improvement of 1.5 points. This result suggests that the proposed GCN network effectively leverages the local contextual relations across the patches. We also observe that it outperforms the Majority-Voting, Learned-Fusion, and Base-Penultimate baselines, which suggests that our

Table 5.2: Experimental results of 3-class ROI classfication experiments (in normalized accuracy).

	Normalized Acc. (%)
Majority-Voting	66.13
Learned-Fusion	67.03
Base-Penultimate	71.85
$\text{GCN}_{\epsilon=0}$	76.90
$\text{GCN}_{\epsilon=200}$	78.56

Table 5.3: Confusion matrix of $\text{GCN}_{\epsilon=200}$.

		Predicted		
		Benign	In Situ	Invasive
Reference	Benign	89	15	1
	In Situ	22	45	2
	Invasive	0	7	42

weakly supervised learning scheme is an effective approach for training with ROI-level labels.

Figure 5.3 shows example patch classification results for the ResNet-based patch classifier (left column), $\text{GCN}_{\epsilon=0}$ model (middle column), and $\text{GCN}_{\epsilon=200}$ model (right column). For the GCN models, the patch classification results are obtained by using the final unnormalized scores computed before the global average pooling layer. These results shown in the figure highlight that the final $\text{GCN}_{\epsilon=200}$ model yields significantly fewer local misclassifications and spatially smoother predictions, compared to both the patch classifier and the $\text{GCN}_{\epsilon=0}$. We note that both GCN models are trained to minimize the ROI classification loss, not per-patch classification losses. Therefore, overall, these results also support the hypothesis that the proposed GCN model can learn to leverage the spatial context information encoded in the graph structure.

5.2.2.2 Four-Class ROI Classification Experiments

The experiments we conduct in this section are based on the most up-to-date version of the data set described in Section 3, which is curated by gathering more

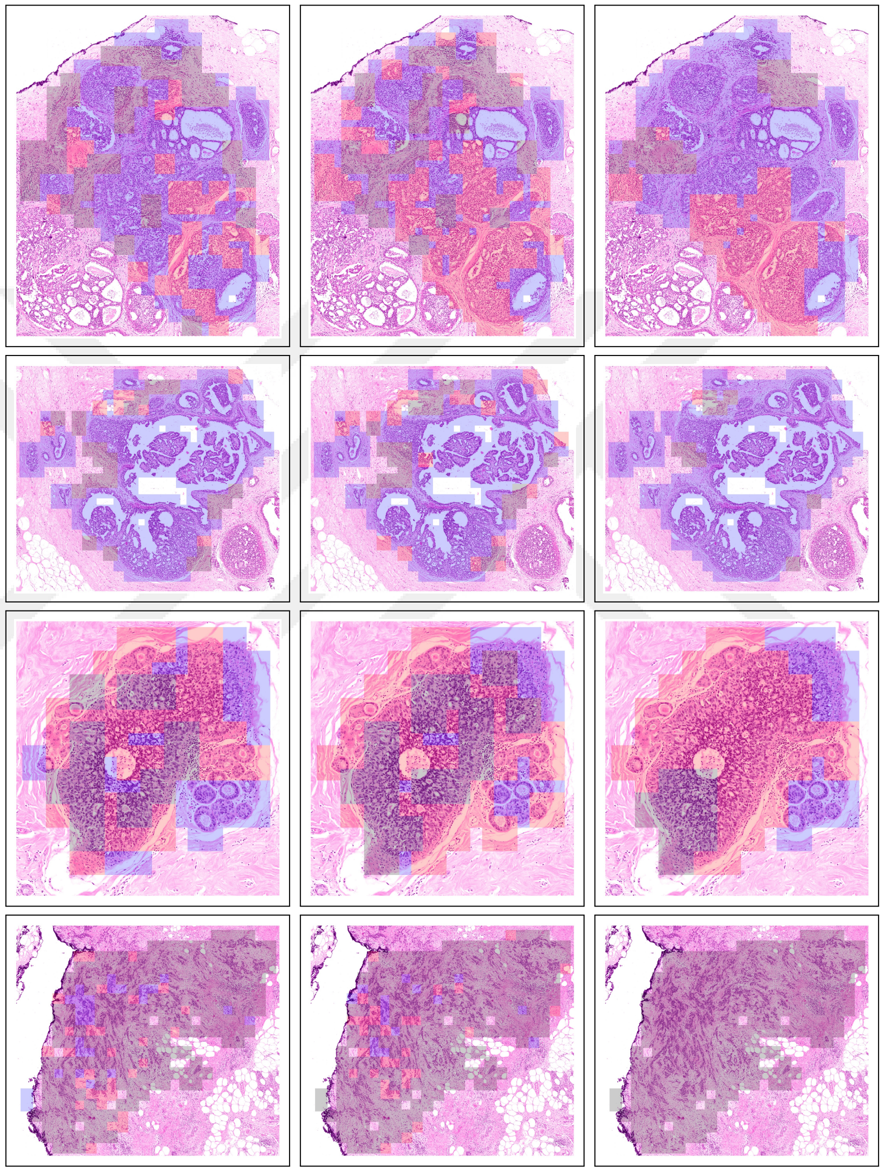

Figure 5.3: Predictions of the individual patches for sample test ROIs (rows), obtained from the fine-tuned patch classifier network (left), GCN$_{\epsilon=0}$ (middle), and GCN$_{\epsilon=200}$ (right). Patches predicted as benign are shown in blue, in situ carcinoma in red, and invasive carcinoma in black. Predictions for the individual patches are obtained just before the global mean-pooling layer for the GCN models. The first two ROIs from the top are diagnosed as benign, the next one as in situ, and the last as invasive. The local averaging effect of GCN$_{\epsilon=200}$ is visible at the right-most row in the form of smoother patch prediction distributions.

samples from each class, enabling us to experiment with samples from atypia class as well.

Experimental setup. For the experiments in this section, we preserve the overall architecture described in Figure 5.2 but replace ResNet-50, which is used as the feature extraction network, with the efficientnet-b0 architecture [87], which results in patch feature representations of size 1280 instead of 2048. Next, we experiment with the following two graph convolutional layer architectures in addition to the GCN layer we describe in Section 5.2.1:

- Graph Attention Network (GAT): After applying a linear transformation to the vertex features, this graph convolutional layer employs a multi-head attention mechanism consisting of a linear layer and a nonlinearity inside the first-order neighborhoods of each vertex to combine these transformed features into per-vertex aggregated features [88]. Then, a nonlinearity is applied to the aggregated features to obtain the final output of the layer.
- Attention-based Graph Neural Network (AGNN): The attention mechanism of AGNN obtains attention weights inside the first-order neighborhood of each vertex by calculating the cosine similarity between the feature vector of the vertex and neighboring vertices, which are then multiplied with a learnable parameter β and normalized over the neighborhood using softmax [89].

The main metric we use for hyperparameter selection and evaluation of the proposed methods and the baselines is *g-mean*, which is calculated as the geometric mean of class recalls. G-mean favors a more balanced evaluation of per-class performances by penalizing the cases where per-class accuracies differ a lot even though the mean accuracy is high. However, we also provide per-class and macro averaged precision, recall, and F-measure scores. To calculate macro averaged F-measure, we first calculate per-class F-measure scores by taking the harmonic mean of per-class precision and recalls. Then, we average these scores to obtain the macro averaged F-measure.

Table 5.4: Class distribution of ROIs in training, validation, and test sets after eliminating small ROIs.

	Benign	Atypia	In Situ	Invasive	Total
Training Set	238	46	173	117	574
Validation Set	104	26	103	58	291
Test Set	123	22	77	56	278
Total	465	94	353	231	1143

We conduct the first experiments regarding the baseline and graph-based methods in 10$\times$ magnification, similar to the Section 5.2.2.1. Then, we evaluate the graph-based methods in 5$\times$ magnification to assess the effect of the magnification change on the classification performance.

Implementation details. The training pipeline consists of two stages: training of efficientnet-based patch-level feature extractor and graph-based ROI classification model, similar to Section 5.2.2.1. The first operation common to these steps is the patch sampling, which is done using the same strategy as in Section 5.2.2.1 by sampling patches of size 224×224 with a 74-pixel overlap between the consecutive patches. After patch sampling, we remove all the ROIs with the number of sampled patches smaller than 5 from the data set. The distribution of the training, validation, and test sets after this operation is given in Table 5.4. Additionally, we apply random vertical/horizontal flips and random rotations to the patches and jitter the patch sampling coordinates by a number of pixels randomly chosen from the interval $[-45, 45]$ in both horizontal and vertical directions to augment the training set during both stages of the training.

The second common operation we do in both stages of the training is oversampling of the minority classes to cope with the class imbalance. Our oversampling strategy consists of two steps. First, we oversample the minority ROIs in a way that approximately the same number of patches are sampled from each ROI class. Then, we further oversample ROIs belonging to the atypia and in situ classes by rates of 3 and 2, respectively, which are chosen using the validation set.

For training of patch-level feature extraction network, we use an efficientnet-b0 model pre-trained on ImageNet and fine-tune its weights using ROI-level labels as the patch-level labels in a patch classification task. As for regularization techniques, we use dropout in the penultimate layer of efficientnet-b0 with drop probability of 0.1 and stochastic depth [90] with drop connect ratio 0.25. We use weight decay regularization with a decay rate of 10^{-4} on all parameters except the batch normalization and bias parameters, following the recommendation in [91]. We use stochastic gradient descent with a constant learning rate of $3 \cdot 10^{-2}$ and a momentum of 0.75 to optimize the parameters. All hyperparameters are chosen with random search to optimize the g-mean score on the validation set for the Majority-Voting baseline. We train the patch classifier network for 40 epochs, where each epoch consists of 500 batches with 512 patches. We use the checkpoint with the highest validation score for the patch-level feature extractor and proceed with the ROI classification models.

The same patch extraction, oversampling and augmentation strategies are used during the training of graph-based ROI classification models. For regularization, we use dropout with drop probability of 0.5 in the first graph convolutional layer and use weight decay with a decay rate of 10^{-4} in the weight parameters of the graph convolutional layers. Similar to the patch-level feature extractor training, we use stochastic gradient descent with a momentum of 0.75, but lower the learning rate to $3 \cdot 10^{-3}$. For GCN and AGNN, we choose the number of hidden units as 128. For GAT, we use 4 attention heads with each head consisting of 128 units, resulting in 512-dimensional vertex feature representations. The proximity threshold parameter ϵ is chosen as 200 pixels while constructing the adjacency matrices for all graph-based models. All models are trained for 40 epochs, where each epoch consists of 500 batches with 64 ROIs. We use the checkpoint from the epoch with the highest validation score to report the test performance.

Results. The experimental results are summarized in Table 5.2.2.2. Learned-Fusion is not included in the comparison since it failed to obtain a g-mean score higher than 0 in any of the various hyperparameter settings we tried for the 4-class classification setting. The other baselines Majority-Voting and Base-Penultimate are surpassed by all graph-based methods for all the metrics listed in Table 5.2.2.2, which further emphasizes the observation made in Section 5.2.2.1 that graph-based approaches are more effective over methods that do not utilize the spatial context in ROI classification problem.

Among all the graph-based methods, we observe that GAT has the highest g-mean (0.724) and average F-measure (0.672) scores, which highlights the benefit of using an attention mechanism in the graph layers. AGNN's lower performance than GAT shows the superiority of the attention formulation of GAT over AGNN's cosine similarity-based formulation, which could be attributed to the higher capacity of the attention layer of GAT consisting of more parameters and multiple attention heads.

Looking at the last three rows of Table 5.2.2.2 where we evaluate the graph-based methods in 5× magnification, we see a drop in g-mean scores of all methods compared to 10× magnification. The most apparent benefit of using 5× magnification is the inclusion of a wider context inside each patch. However, given our experimental results which demonstrate that graph-based methods already utilize the spatial context effectively in 10×, it is possible that switching to a lower magnification does more harm than good due to loss of fine-grained information present in 10× magnification.

5.3 Weakly Supervised Instance Attention

5.3.1 Methodology

Although our results in the previous section show that utilization of the spatial context formed by patches can lead to better classification performance, the problem of individual contributions of patches to the ROI-level diagnosis remains unexplored. Based on the discussion in Section 5.1, where we argue that not all patches yield equally useful information towards ROI-level classification, with

Table 5.5: Experimental results of 4-class ROI classification experiments. Per-class and macro averaged precision, recall, and F-measure scores are listed in addition to the g-mean scores. For per-class scores, the abbreviations B, A, IS, and INV depict benign, atypia, in situ, and invasive classes, respectively.

	G-mean	Precision					Recall					F-measure				
		B	A	IS	INV	Avg.	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.
Majority-Voting	0.508	0.820	0.098	0.639	0.817	0.593	0.582	0.265	0.495	0.875	0.554	0.680	0.143	0.558	0.845	0.556
Base-Penultimate	0.687	0.852	0.204	0.782	0.939	0.694	0.561	0.864	0.558	0.821	0.701	0.676	0.330	0.652	0.876	0.634
GCN (10×)	0.694	0.859	0.210	0.774	0.927	0.692	0.691	0.591	0.623	0.911	0.704	0.766	0.310	0.691	0.919	0.671
GAT (10×)	0.724	0.895	0.233	0.769	0.907	0.701	0.626	0.773	0.649	0.875	0.731	0.737	0.358	0.704	0.891	0.672
AGNN (10×)	0.711	0.847	0.226	0.837	0.941	0.713	0.585	0.955	0.532	0.857	0.732	0.692	0.365	0.651	0.897	0.651
GCN (5×)	0.682	0.907	0.190	0.723	0.962	0.696	0.715	0.545	0.610	0.911	0.696	0.800	0.282	0.662	0.936	0.670
GAT (5×)	0.715	0.902	0.227	0.759	0.925	0.703	0.675	0.773	0.571	0.875	0.723	0.772	0.351	0.652	0.899	0.668
AGNN (5×)	0.691	0.855	0.211	0.812	0.923	0.700	0.577	0.909	0.506	0.857	0.712	0.689	0.342	0.624	0.889	0.636

some patches being informative and some being misleading or confusing, we revisit the WSL framework we use in Chapter 4 and propose a similar instance attention mechanism for ROI classification problem in this section. Then, we further improve the performance of the proposed instance attention model with a new training scheme.

5.3.1.1 Instance-Attention

In this method, we represent each ROI as a bag-of-patches by using the same sampling strategy as in Section 5.2 before the graph construction, which may result in a mixture of patches with some patches carrying critical information for the classification of the ROI and some patches being uninformative, confusing, or misleading, as discussed in Section 5.1. Given that each ROI is annotated by a single diagnostic label which carries no information about the relevance of any of the patches, a weakly supervised instance attention approach similar to the one described in Section 4.1 can be utilized to attend to the relevant patches while disregarding the irrelevant ones.

To approach the ROI classification problem from a WSL perspective, we use the formulation in Equation (4.7) which incorporates ϕ^{enc} operation consisting of candidate region extraction (ϕ^{prop}) and region encoding (ϕ^{region}), as well as ϕ^{pred} operation which combines the localization (ϕ^{loc}) and classification (ϕ^{cls}) branch outputs, with a few minor changes. The first change is to replace the sliding window candidate region extractor with the patch sampling approach mentioned in the previous paragraph, which is more suitable for inputs with irregular shapes such as ROIs. The second change is switching to the original ϕ^{pred} definition in Equation (4.4) by removing the bias and softmax from ϕ^{pred} since we now assume that some patches sampled from an ROI are likely to be similar to the patches sampled from ROIs belonging to other diagnostic classes. With these changes, Equation (4.10) is updated as:

$$P(c|x) = \left[\phi^{WSL}(x)\right]_c \tag{5.1}$$

where $P(c|x)$ is the probability that a structure belonging to the diagnostic category c exists in the ROI x. Note that $P(c|x)$ is already in the interval [0-1] for all classes due to the previous softmax operations taking place in the localization and classification branches, which allows us to use them directly as probability scores and calculate the loss as:

$$\mathcal{L}(x,y) = \sum_{c=1}^{C} [y]_c log(P(c|x)) + (1 - [y]_c)(1 - log(P(c|x))) \tag{5.2}$$

which is sum of C binary cross-entropy terms where C is the number of classes and y is the one-hot encoding of the class label of x. We call this approach simply as Instance-Attention for the rest of this thesis.

5.3.1.2 Instance-Attention-Mixup

Following the discussion on the ROI-level labels and the noisy nature of the sampled patches, we propose a novel addition to the weakly supervised training pipeline of the instance attention approach mentioned above, which we call Instance-Attention-Mixup since it is loosely inspired from the mixup augmentation [92] where the training data is augmented with convex combinations of randomly chosen input pairs and their labels. In Instance-Attention-Mixup, however, we propose to combine pairs of ROIs from different classes by putting all patches sampled from both ROIs into a larger bag-of-patches and combine their one-hot encodings into a *two-hot* encoding by summation. More specifically, we define $\phi^{\textit{WSL-mixup}}$ as:

$$\phi^{\textit{WSL-mixup}}(x_1, x_2) = \phi^{pred}(\psi(\phi^{enc}(x_1), \phi^{enc}(x_2))) \tag{5.3}$$

where ψ denotes the concatenation operation and x_1 and x_2 denote the first and the second ROI, respectively. Concatenation of the results of ϕ^{enc} achieves the desired construction of the bigger bag-of-patches since ϕ^{enc} encodes each patch independently after sampling them. With this change, Equation (5.1) is updated as:

$$P(c|x_1, x_2) = \left[\phi^{\textit{WSL-mixup}}(x_1, x_2)\right]_c \tag{5.4}$$

The updated loss function for Instance-Attention-Mixup is given as:

$$\begin{aligned}\mathcal{L}(x_1, x_2, y_1, y_2) = &\sum_{c=1}^{C}([y_1]_c + [y_2]_c) log(P(c|x_1, x_2)) \\ &+(1 - ([y_1]_c + [y_2]_c))(1 - log(P(c|x_1, x_2)))\end{aligned} \tag{5.5}$$

where y_1 and y_2 are the one-hot encodings of the labels of the first and the second ROI, respectively. Note that the summation term $[y_1]_c + [y_2]_c$ is either 0 or 1, since the labels of the ROIs are chosen to be different.

The overall approach is summarized in Figure 5.4. The motivation for such an approach is to utilize the instance attention model's capability to learn from inputs with multiple labels. Given that each ROI has a single label, the localization and classification branches of Instance-Attention without the proposed modification can only learn how to attend to and classify patches relevant to the current ROI class, even if the ROI contains patches that are more likely to belong to other classes. For localization and classification branches to benefit from the presence of patches in an ROI more relevant to a class different from the given ROI class, label encoding vector y of the ROI should signal the presence of both classes. Specifically, given the reference class index as c_{ref} and the other class present in the ROI as c_{other}, label encoding of the ROI should be:

$$[y]_c = \begin{cases} 1 & \text{if } c \in \{c_{ref}, c_{other}\} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

to be able to learn from both classes. However, the presence of c_{other} is not guaranteed for all ROIs, and we would not have any information regarding what c_{other} could be even if it were guaranteed to exist. By combining two ROIs with different reference labels into a larger bag-of-patches, we can safely construct the encoding y in Equation (5.6) since c_{other} is not only guaranteed to exist but also known thanks to the contribution of the second ROI.

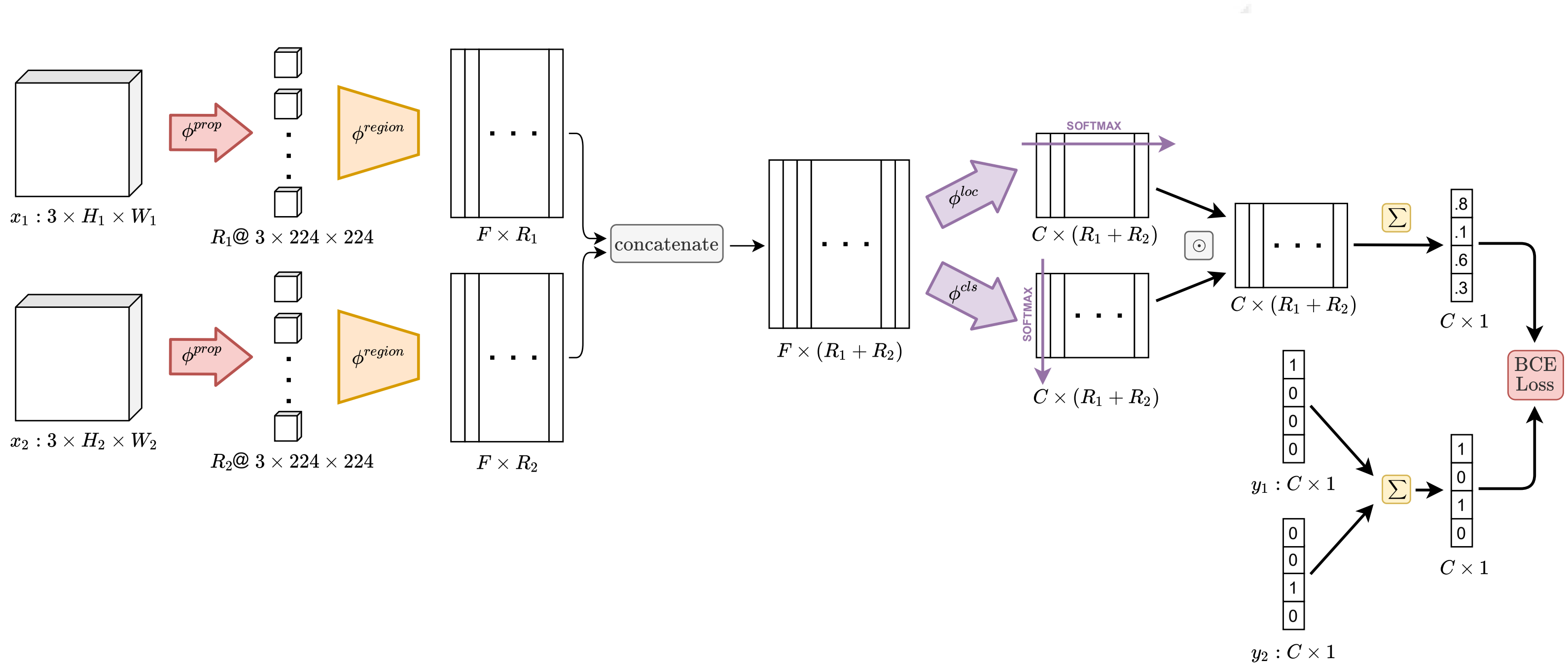

Figure 5.4: The proposed Instance-Attention-Mixup training strategy. x_1 and x_2 denotes the ROIs to be mixed, with their respective one-hot label vectors being y_1 and y_2. Patches sampled from each ROI are processed separately by using the same patch encoder network ϕ^{region}, which results in matrices of size $F \times R_1$ and $F \times R_2$, where F is the number of dimensions of the patch feature representations, and R_1 and R_2 are the number of patches sampled from the first and the second ROI, respectively. These are then concatenated into a larger matrix of size $F \times (R_1 + R_2)$ to obtain a larger bag-of-patches representation and processed by the localization (ϕ^{loc}) and classification (ϕ^{cls}) branches. Hadamard product of the outputs of ϕ^{loc} and ϕ^{cls} is taken, and resulting scores are summed up in the patch dimension to obtain a C-dimensional score vector where C is the number of classes, with each score indicating the probability of that class being present inside the bag. This vector is used to calculate a sum of C binary cross-entropy terms as the training loss together with the two-hot reference label vector obtained by summing y_1 and y_2 up.

5.3.2 Experiments

5.3.2.1 Experimental Setup

For the performance comparison of the proposed models in this section, we use the following methods as baselines:

- Patch-Classifier-Majority: A patch classifier network consisting of a CNN encoder and a fully-connected classification layer is trained using the ROI labels as the labels for each patch of the ROI. After the patch classifier is trained, the inference is made by classifying each patch of an ROI independently and then taking the patch predictions' majority vote as the ROI-level prediction.
- Patch-Classifier-Mean-Prob: The same steps as the Patch-Classifier-Majority are followed except for the inference phase, where the patch probability scores are averaged to obtain an ROI-level probability score vector.
- ROI-Classifier-Penultimate-Mean: This baseline uses the same architecture as Patch-Classifier-Majority, where a CNN encodes each patch into a feature representation. After that, the feature vectors of the patches of an ROI are averaged to obtain an ROI-level representation, which is then processed by a fully-connected classification layer to come up with an ROI-level prediction.
- mMIL: This baseline is the generalization of the max-pooling aggregation rule to the multi-class setting as described in [8], where a class score vector is obtained for each patch using a CNN and a classification layer. Then, the vector having the highest class score among all patches and classes is selected as the ROI-level score vector to be used in the loss calculation.
- MIL-Attention: Similar to ROI-Classifier-Penultimate-Mean, each patch is encoded into a feature vector via an encoder CNN. These vectors are aggregated using a weighted summation to obtain an ROI-level feature representation, which is fed into a classification layer to get an ROI-level prediction. Weights of this summation operation are obtained via an attention layer

with the following formulation:

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_k^\top)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_j^\top)\}} \tag{5.7}$$

where a_k are the attention weights, $\mathbf{h}_k$ are patch feature vectors, $\tanh(\cdot)$ is the hyperbolic tangent activation funciton and $\mathbf{w} \in \mathbb{R}^{M\times 1}$ and $\mathbf{V} \in \mathbb{R}^{M\times F}$ are learnable parameters [74]. Here, F is the number of dimensions of the patch feature representation vectors, and M is a hyperparameter denoting the number of hidden units in the attention layer.

- MIL-Attention-Gated: This baseline applies the gated variant of the attention mechanism used in MIL-Attention, which has the following form:

$$a_k = \frac{\exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_j^\top))\}} \tag{5.8}$$

where $\mathbf{U} \in \mathbb{R}^{M\times F}$ represent additional learnable weigts and $\mathrm{sigm}(\cdot)$ is the sigmoid activation function [74].

- MIL-Per-Class-Attention-Gated: This method is the extended version of MIL-Attention-Gated which uses per-class attention branches as described in [8], which has the following form:

$$a_{kc} = \frac{\exp\{\mathbf{w_c}^\top (\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^{K} \exp\{\mathbf{w_c}^\top (\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_j^\top))\}} \tag{5.9}$$

where c is the class index and a_{kc} is the attention weight of the k^{th} patch for class c. The difference between Equation (5.8) and Equation (5.9) is that each attention branch applies a separate linear transformation in the form of $\mathbf{w_c}$ in Equation (5.9), after a shared attention backbone. A different weighted combination of patch feature representations is obtained for each branch, which results in a distinct ROI-level feature vector per branch. A separate linear transformation is applied to each per-branch ROI feature vector to transform the vector into a single number that represents the logit value for that particular class.

- CLAM: Clustering-constrained Attention Multiple instance learning (CLAM) is a combination of MIL-Per-Class-Attention-Gated trained using standard cross-entropy loss and per-class binary patch clustering layers trained using multi-class SVM loss, which is designed help the model to better separate the positive and negative patches for each class from each other [8]. These two losses are combined as:

$$\mathcal{L}_{total} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{SVM}\mathcal{L}_{SVM} \tag{5.10}$$

where $\mathcal{L}_{CE}$ and $\mathcal{L}_{SVM}$ are the aforementioned cross-entropy and multi-class SVM losses and λ_{CE} and λ_{SVM} are hyperparameters that correspond to the weights of these loss terms, respectively.

Although the authors study the slide-level weakly supervised classification problem in the original implementation, we adapt their methodology to the ROI-level weakly supervised classification setting. While adapting the methodology to the problem we focus on, we make several changes to the original implementation. First, we use the penultimate layer features of an efficientnet-b0 pre-trained on ImageNet as patch-level features and fine-tune its weights, while the original CLAM implementation utilizes intermediate layer features of an ImageNet pre-trained ResNet-50 without fine-tuning. We also remove the fully-connected layer which is applied to ResNet-50 features before the attention layers in the original implementation to make the method comparable to the other methods we experiment with. Finally, instead of choosing a fixed number of patches with the highest attention score and the same number of patches with the lowest attention score, we choose a ratio (b) of the highest-scoring patches and the same ratio of the lowest-scoring patches to calculate the multi-class SVM loss.

- CLAM-Smooth: This version of CLAM uses the smooth top1 SVM loss instead of the multi-class SVM loss used in CLAM [8].
- Instance-Attention-Label-Smoothing: In this method, label smoothing is applied to each ROI so that the label encodings of the ROIs can slightly signal the presence of every class other than the reference label. The formulation for label smoothing can be described as:

$$[y]_c = \begin{cases} 1-\epsilon & \text{if } c = c_{ref} \\ \epsilon/(C-1) & \text{otherwise} \end{cases} \tag{5.11}$$

where ϵ is the hyperparameter controlling the smoothing strength and C is the number of classes.

- Instance-Attention-Random-Label-Add: This baseline consists of converting the one-hot label encoding vector of an ROI into two-hot with a probability p_{flip} by flipping one of the zeros in the vector chosen randomly.

We use g-mean as the main metric regarding the hyperparameter selection and comparison across the methods. We also report per-class and macro averaged precision, recall, and F-measure scores, calculated in the same way we describe in Section 5.2.2.2.

5.3.2.2 Implementation details

For the training of the ROI classification models, we use the training pipeline we describe for the patch classifier training in Section 5.2.2.2 with the same augmentation, regularization, and oversampling strategies, except that we carry out the training end-to-end. One of the challenges in the end-to-end training of the proposed methods is the large memory requirement for batch processing of the ROIs, which can cover very large areas in 10× magnification (see Table 3.2). We switch to 5× magnification to reduce the area per-ROI and utilize mixed precision training to further reduce the memory requirements to overcome this issue. Another challenge is that the number of patches in a batch consisting of a fixed number of ROIs can differ significantly between the iterations due to the size variation between the ROIs. This can introduce inefficiency to the training by causing the computational resources to be underutilized if the batch contains small ROIs or memory issues when the batch consists of large ROIs. To overcome these problems, we use batches of 512 patches instead of a fixed number of ROIs to utilize the computational resources in all training iterations fully. Secondly, we limit the number of patches coming from each ROI to 30 by randomly dropping patches if their number exceeds this value. This serves as a way to overcome the memory problems with large ROIs as well as a form of regularization, akin to the multiple instance augmentation technique proposed in [93]. Additionally, we use efficientnet-b0 architecture as the region encoder network ϕ^{region}, for its efficiency and low memory requirements [87], which results in feature representations of size 1280 for each patch. We use an efficientnet-b0 model pre-trained on ImageNet and fine-tune its weights in all experiments.

We introduce an additional hyperparameter for Instance-Attention-Mixup training called mixup probability (p_{mixup}), which denotes the probability to mix the patches of a given ROI with another ROI chosen randomly among the ROIs inside the same batch belonging to other classes. No mixup is applied if there are no ROIs inside the batch with a different label than the current ROI. When there is at least one ROI with a different label, we first choose a label randomly among the existing labels inside the batch, excluding the current label. Then, we choose the ROI with the number of patches closest to the number of patches of the current ROI among the ROIs with the chosen label. The reason for this is to construct as balanced bags-of-patches as possible to prevent one ROI from dominating the other. We use a mixup probability of 0.25 in our experiments.

For Instance-Attention-Label-Smoothing, we choose the smoothing parameter ϵ as 0.2. For the random label addition probability p_{flip} in Instance-Attention-Random-Label-Add, we use the same value as p_{mixup} for a fair comparison with Instance-Attention-Mixup.

For MIL-Attention, MIL-Attention-Gated, MIL-Per-Class-Attention-Gated, CLAM, and CLAM-Smooth, we choose the number of hidden units in the attention layers (M) as 128. Among the hyperparameters specific to CLAM and CLAM-Smooth, we set both the margin and the temperature parameters of the multi-class SVM loss in CLAM and the smooth top1 SVM loss in CLAM-Smooth to 1.0, similar to the original implementation. Cross-entropy and SVM loss weights λ_{CE} and λ_{SVM} are set to 0.7 and 0.3, respectively. The last hyperparameter b, the ratio of the highest-scoring patches and the lowest-scoring patches to be selected from each ROI for the SVM loss calculation, is selected as 0.1.

We train all models for 40 epochs, with each epoch consisting of 500 batches sampled with the oversampling strategy mentioned in Section 5.2.2.2. In all of our experiments regarding both the proposed models and the baselines, we first apply this oversampling methodology to obtain an oversampled version of the training data set before proceeding with the training pipeline. We check the g-mean scores on the validation set after each epoch and use the model checkpoint with the highest validation score to report the test results.

5.3.2.3 Results

Results of our experiments are shown in Table 5.6. The first observation we make on these results is the superior performance of the patch-classifier-based approaches over ROI-Classifier-Penultimate-Mean. One of the key differences between patch classification and ROI classification approaches is the content of the batches used in training. In patch classifier training, each batch is constructed by randomly selecting patches among the oversampled ROIs, which may result in a batch of patches all sampled from different ROIs in the most extreme case. In the ROI classifier training, however, we construct batches by first sampling an ROI, then putting all patches of the ROI inside the batch with an upper limit on the number of patches from the same ROI as described in Section 5.3.2.2. This usually results in a smaller number of ROIs inside a batch than the batches used in the patch classifier, even though the batch size is the same in terms of patches. Considering this, one reason for the superior performance of the Patch-Classifier-Majority and Patch-Classifier-Mean-Prob over ROI-Classifier-Penultimate-Mean could be the usage of batches with more diversity in terms of ROIs. One other difference between these approaches is the mean-pooling in the ROI classifier applied to the patch features before the classification layer. Since the classification loss is computed on the ROI-level logits obtained after the pooling operation, the loss propagated through each patch becomes smaller for the ROIs with many patches. This results in more weight given to the patches of smaller ROIs during the training, which might not have a meaningful basis.

Regarding the MIL and attention-based approaches, we first observe that mMIL is the worst performing method among all baselines, which points out that max-pooling-based MIL formulation is inferior compared to the mean-pooling or attention-based formulations. We also observe that MIL-Attention yields a low g-mean score, mostly due to its low recall on atypia, even though its gated variant

Table 5.6: Experimental results for baselines and proposed methods trained end-to-end. Per-class and macro averaged precision, recall, and F-measure scores are listed in addition to the g-mean scores. For per-class scores, the abbreviations B, A, IS, and INV depict benign, atypia, in situ, and invasive classes, respectively.

	G-mean	Precision					Recall					F-measure				
		B	A	IS	INV	Avg.	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.
Patch-Classifier-Majority	0.672	0.920	0.172	0.773	0.869	0.683	0.650	0.500	0.662	0.946	0.690	0.762	0.256	0.713	0.906	0.659
Patch-Classifier-Mean-Prob	0.655	0.899	0.159	0.750	0.911	0.680	0.650	0.500	0.623	0.911	0.671	0.755	0.242	0.681	0.911	0.647
ROI-Classifier-Penultimate-Mean	0.636	0.865	0.148	0.824	0.903	0.685	0.732	0.409	0.545	1.000	0.672	0.793	0.217	0.656	0.949	0.654
mMIL	0.545	0.843	0.099	0.688	0.867	0.624	0.610	0.364	0.429	0.929	0.583	0.708	0.155	0.528	0.897	0.572
MIL-Attention	0.619	0.846	0.163	0.857	0.889	0.689	0.846	0.318	0.545	1.000	0.677	0.846	0.215	0.667	0.941	0.667
MIL-Attention-Gated	0.640	0.890	0.145	0.870	0.873	0.694	0.724	0.455	0.519	0.982	0.670	0.798	0.220	0.650	0.924	0.648
MIL-Per-Class-Attention-Gated	0.656	0.876	0.159	0.857	0.918	0.703	0.748	0.455	0.545	1.000	0.687	0.807	0.235	0.667	0.957	0.667
CLAM	0.660	0.876	0.179	0.870	0.889	0.703	0.805	0.455	0.519	1.000	0.695	0.839	0.256	0.650	0.941	0.672
CLAM-Smooth	0.663	0.873	0.185	0.867	0.918	0.711	0.837	0.455	0.506	1.000	0.700	0.855	0.263	0.639	0.957	0.679
Instance-Attention	0.662	0.907	0.171	0.681	0.962	0.680	0.634	0.545	0.610	0.911	0.675	0.746	0.261	0.644	0.936	0.647
Instance-Attention-Label-Smoothing	0.609	0.829	0.133	0.830	0.981	0.693	0.789	0.364	0.506	0.946	0.651	0.808	0.195	0.629	0.964	0.649
Instance-Attention-Random-Label-Add	0.654	0.860	0.175	0.812	0.949	0.699	0.797	0.455	0.506	1.000	0.689	0.827	0.253	0.624	0.974	0.670
Instance-Attention-Mixup	0.689	0.885	0.200	0.760	0.981	0.707	0.691	0.727	0.494	0.911	0.706	0.776	0.314	0.598	0.944	0.658

MIL-Attention-Gated achieves comparable results to other methods, showing the potential benefits of the gating mechanism in attention layers. A per-class attention formulation combined with the gating mechanism boosts the results further, as observed in MIL-Per-Class-Attention-Gated, CLAM, and CLAM-Smooth results. Among these methods, we observe CLAM and CLAM-Smooth to yield superior performances, with a small difference of 0.003 between their g-mean scores.

The proposed Instance-Attention approach shows performance comparable to CLAM, CLAM-Smooth, MIL-Per-Class-Attention-Gated, and Patch-Classifier-Mean-Prob, although it falls short of Patch-Classifier-Majority, which surpasses all MIL and attention-based baselines as well. Although this result does not seem in favor of Instance-Attention at first, it might be hinting at the possibility of obtaining better results with Instance-Attention if more diverse batches were used, given that the same discussion at the beginning of this section applies to Instance-Attention as well.

Next, we observe that Instance-Attention-Mixup surpasses all the baselines with a g-mean score of 0.689. Looking at the large performance difference between Instance-Attention-Label-Smoothing and Instance-Attention-Mixup, we can infer that the proposed ROI mixing strategy has a more meaningful impact on training than a simple label smoothing in such a WSL setting with a bag of noisy instances. Furthermore, we observe a huge performance drop in Instance-Attention with the addition of label smoothing, which points out that simple label smoothing can have a detrimental effect on the performance instead. Similarly, we observe a performance drop with Instance-Attention-Random-Label-Add as well, implying that the proposed mixup methodology is more beneficial than introducing random perturbations to the labels.

To compare the two best performing models, Patch-Classifier-Majority and Instance-Attention-Mixup, we look at the confusion matrices of the corresponding models obtained on the test set, given in Tables 5.7 and 5.8, respectively. Although both methods show similar performances for benign and invasive ROIs, their performances differ a lot for atypia and in situ classes. Specifically, we

Table 5.7: Confusion matrix of Patch-Classifier-Majority.

		Predicted			
		Benign	Atypia	In Situ	Invasive
Reference	Benign	80	35	6	2
	Atypia	2	11	6	3
	In Situ	5	18	51	3
	Invasive	0	0	3	53

Table 5.8: Confusion matrix of Instance-Attention-Mixup.

		Predicted			
		Benign	Atypia	In Situ	Invasive
Reference	Benign	85	29	8	1
	Atypia	3	16	3	0
	In Situ	6	33	38	0
	Invasive	2	2	1	51

observe that Patch-Classifier-Majority performs better on in situ while Instance-Attention-Mixup obtains a higher score on atypia. One possible reason for this behavior is that the oversampling applied to the atypia class causes Instance-Attention-Mixup to mix more atypical ROIs with the other ones, resulting in a model more biased towards atypia, which results in decreased performance in other classes. Its lower performance on in situ can be explained from a similar perspective as well. Since in situ is the second most oversampled class after atypia, most of the mixture of ROIs are expected to be mixtures of atypia and in situ. This might introduce an additional bias towards atypia between these two classes, given that atypia is the most oversampled class.

To assess the significance of the performance improvement obtained by Instance-Attention-Mixup, we apply a statistical significance analysis between each of the baselines and Instance-Attention-Mixup, using the predictions obtained on the test set. McNemar's test [94] is recommended for cases where there is a hold-out evaluation set with a limited amount of data [95]. However, due to McNemar's test being applicable only to 2×2 contingency tables, we use Bhapkar's test [96] which can be used with tables of any size, hence allowing us to use 4×4 contingency tables where each column/row corresponds to one of the four classes we have. The highest p-value is obtained for mMIL with a value of 0.2465, and the second highest p-value obtained for Patch-Classifier-Mean-Prob with a value of 0.0187, while all remaining p-values are observed to be smaller than 0.0025. These results indicate a significant difference between the performance of Instance-Attention-Mixup and the other baselines, except for mMIL where we observe a high p-value. However, given that the performance of mMIL is considerably low compared to all other methods we evaluate according to Table 5.6, we do not accredit much importance to the lack of statistical significance between Instance-Attention-Mixup and mMIL results implied by the high p-value.

Given the superior performances of the Instance-Attention-Mixup as discussed in this section and GAT as discussed in Section 5.2.2.2 over several strong baselines, a natural question to be asked is whether a combination of these two methods can achieve a better classification performance. To investigate this question, we experiment with the combination of GAT with both Instance-Attention and Instance-Attention-Mixup in 5× and 10× magnification settings. To make this combination possible, we switch from the end-to-end training scheme described in Section 5.3.2.2 to the two-stage pipeline we use for graph network experiments in Section 5.2.2.2. The main reason for this change is that in order to construct a meaningful graph-based ROI representation to leverage the graph layers' capability of exploiting the spatial context, we need to have as many patches from an ROI inside the batch as possible. However, this is not possible in the end-to-end setting with batches of multiple ROIs, especially when larger ROIs are present inside the batch, due to memory limitations. We carry out the combination of two methods by replacing the fully-connected layers in the localization (ϕ^{loc}) and classification (ϕ^{cls}) branches with two GAT layers with a ReLU nonlinearity in between. We use 4 attention heads with size 128 in the first GAT layer similar to Section 5.2.2.2.

The results of Instance-Attention-Mixup and GAT combination is presented in Table 5.9. Note that Instance-Attention and Instance-Attention-Mixup results at 5× magnification in Table 5.9 differ from the results in Table 5.6, due to the differences in the training pipeline discussed above. Given that the performances of these methods are lower than the previous results, we can argue that the end-to-end training scheme is more beneficial for Instance-Attention and

Table 5.9: Effect of magnification and GAT layer on Instance-Attention and Instance-Attention-Mixup in two-stage training. For per-class scores, the abbreviations B, A, IS, and INV depict benign, atypia, in situ, and invasive classes, respectively.

	Mag.	G-mean	Precision					Recall					F-measure				
			B	A	IS	INV	Avg.	B	A	IS	INV	Avg.	B	A	IS	INV	Avg.
Instance-Attention	5×	0.649	0.902	0.185	0.649	1.000	0.684	0.748	0.455	0.649	0.804	0.664	0.818	0.263	0.649	0.891	0.655
Instance-Attention-Mixup		0.679	0.897	0.208	0.691	1.000	0.699	0.780	0.500	0.610	0.893	0.696	0.835	0.293	0.648	0.943	0.680
Instance-Attention + GAT		0.720	0.888	0.267	0.778	0.907	0.710	0.772	0.545	0.727	0.875	0.730	0.826	0.358	0.752	0.891	0.707
Instance-Attention-Mixup + GAT		0.728	0.905	0.239	0.783	0.936	0.716	0.699	0.727	0.701	0.786	0.728	0.789	0.360	0.740	0.854	0.686
Instance-Attention	10×	0.514	0.824	0.116	0.612	0.824	0.594	0.724	0.364	0.532	0.500	0.530	0.771	0.176	0.569	0.622	0.535
Instance-Attention-Mixup		0.545	0.861	0.095	0.697	0.976	0.657	0.756	0.273	0.597	0.714	0.585	0.805	0.141	0.643	0.825	0.604
Instance-Attention + GAT		0.682	0.875	0.195	0.725	0.923	0.679	0.569	0.682	0.649	0.857	0.689	0.690	0.303	0.685	0.889	0.642
Instance-Attention-Mixup + GAT		0.684	0.903	0.200	0.743	0.900	0.687	0.683	0.591	0.675	0.804	0.688	0.778	0.299	0.707	0.849	0.658

Instance-Attention-Mixup. Our second observation is that even though GAT performs better at 10× magnification than 5× on its own (see Table 5.2.2.2), its combination with Instance-Attention and Instance-Attention-Mixup performs significantly higher at 5× than 10×. Overall, we conclude that the best classification performance among all our experimental results is achieved by the combination of Instance-Attention-Mixup and GAT at 5× magnification, with a g-mean score of 0.728.

Chapter 6

Conclusion

We studied weakly supervised learning based approaches in image classification problems from two different domains: remote sensing and medical image analysis. In remote sensing, we tackled 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. 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.

For the medical image analysis domain, we focused on breast cancer classification in breast biopsy images with two different approaches. More specifically, we tackled the problem of classifying ROIs of arbitrary shape and size in breast histopathology images. We based our first approach on the observation that the mainstream approach of first classifying individual patches and then combining the patch classification results fails to leverage the rich spatial context in complex tissue structures. To address this limitation without resizing ROIs into predefined fixed sizes, we proposed a graph-based ROI representation and a graph-neural-network-based architecture that operate on these graph structures. Our experimental results indicate significant improvements over a number of strong baselines and suggest that the proposed approach is capable of leveraging the spatial context information in ROIs.

In our second approach, we modeled the ROIs as bags-of-patches, with the ROI-level label serving as a weak label for each patch, which sees only a limited part of the ROI. We hypothesized that patches carry different levels of information, with some patches being relevant to the diagnosis of the ROI, while others being uninformative or even misleading. To obtain an ROI-level decision from such patches, we revisited the WSL formulation we applied for our remote sensing experiments. Utilizing a similar instance attention approach, we managed to obtain results comparable to several baselines. We further extended this approach by introducing a novel WSL training methodology based on our assumption of noisy patches. We showed that the proposed methodology surpasses all the baselines, which confirms the validity of our assumptions about the noisy nature of the patches sampled from ROIs. Finally, we combined the instance attention approach with the graph-based methodology and showed that it is possible to obtain an even higher classification performance by jointly benefiting from the spatial context formed by the patches and focusing on the information coming from the relevant patches.

One future direction that could follow the research done in this thesis regarding the remote sensing experiments is to consider the fusion of multiple sources using the feature representations obtained from the intermediate layers of the networks. For breast histopathology, a similar WSL framework can also be applied at the whole slide-level using slide-level weak labels. Although this problem could be more challenging due to the presence of even weaker labels, it would make the data labeling process more manageable, enabling more whole slide data to be collected for a richer data set. Finally, a multi-resolution approach can be incorporated into the WSL pipeline to further strenghten the attention and classification capabilities by leveraging the structural information emerging in different magnification levels.

Bibliography

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

[3] R. Girshick, "Fast R-CNN," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1440–1448, 2015.

[4] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[5] J. Ji, T. Zhang, Z. Yang, L. Jiang, W. Zhong, and H. Xiong, "Aircraft detection from remote sensing image based on a weakly supervised attention model," in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 322–325, 2019.

[6] Z. Zhang, W. Guo, M. Li, and W. Yu, "GIS-supervised building extraction with label noise-adaptive fully convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[7] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[8] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, “Data efficient and weakly supervised computational pathology on whole slide images,” *arXiv preprint arXiv:2004.09666*, 2020.

[9] J. Wang, R. J. Chen, M. Y. Lu, A. Baras, and F. Mahmood, “Weakly supervised prostate TMA classification via graph convolutional networks,” in *IEEE International Symposium on Biomedical Imaging*, pp. 239–243, IEEE, 2020.

[10] Q. Oliveau and H. Sahbi, “Learning attribute representations for remote sensing ship category classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, pp. 2830–2840, June 2017.

[11] 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 Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 13–30, January 2018.

[12] G. Sumbul, R. G. Cinbis, and S. Aksoy, “Fine-grained object recognition and zero-shot learning in remote sensing imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 770–779, February 2018.

[13] B. Chen, B. Huang, and B. Xu, “Multi-source remotely sensed data fusion for improving land cover classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 124, pp. 27–39, February 2017.

[14] Y. Han, F. Bovolo, and L. Bruzzone, “Edge-based registration-noise estimation in VHR multitemporal and multisensor images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, pp. 1231–1235, September 2016.

[15] G. Sumbul, R. G. Cinbis, and S. Aksoy, “Multisource region attention network for fine-grained object recognition in remote sensing imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4929–4937, 2019.

[16] B. Aygunes, S. Aksoy, and R. G. Cinbis, "Weakly supervised deep convolutional networks for fine-grained object recognition in multispectral images," in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 1478–1481, 2019.

[17] Litjens, G., et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, December 2017.

[18] C. Mercan, B. Aygunes, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Deep feature representations for variable-sized regions of interest in breast histopathology," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[19] B. Aygunes, S. Aksoy, R. G. Cinbis, K. Kosemehmetoglu, S. Onder, and A. Uner, "Graph convolutional networks for region of interest classification in breast histopathology," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, vol. 11320, pp. 134 – 141, 2020.

[20] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, pp. 1560–1584, September 2015.

[21] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proceedings of the IEEE*, vol. 103, pp. 1585–1601, September 2015.

[22] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pizurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, pp. 2405–2418, June 2014.

[23] W. Liao, X. Huang, F. V. Coillie, S. Gautama, A. Pizurica, W. Philips, H. Liu, T. Zhu, M. Shimoni, G. Moser, and D. Tuia, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the

2014 IEEE GRSS Data Fusion Contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, pp. 2984–2996, June 2015.

[24] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. L. Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest — part A: 2-D contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, pp. 5547–5559, December 2016.

[25] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, I. Tankoyeu, B. Bechtel, B. Le Saux, G. Moser, and D. Tuia, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 1363–1377, May 2018.

[26] M. Datcu, F. Melgani, A. Piardi, and S. B. Serpico, "Multisource data classification with dependence trees," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, pp. 609–617, March 2002.

[27] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 1822–1835, June 2008.

[28] 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 Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 3346–3358, June 2014.

[29] Y. Zhang, H. L. Yang, S. Prasad, E. Pasolli, J. Jung, and M. Crawford, "Ensemble multiple kernel active learning for classification of multisource remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, pp. 845–858, February 2015.

[30] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 7708–7720, December 2014.

[31] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 193–205, January 2019.

[32] G. Gao and Y. Gu, "Tensorized principal component alignment: A unified framework for multimodal high-resolution images classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 46–61, January 2019.

[33] 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 *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pp. 1–5, August 2016.

[34] P. Ghamisi, B. Hofle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 3011–3024, 2017.

[35] L. Pibre, M. Chaumont, G. Subsol, D. Ienco, and M. Derras, "How to deal with multi-source data for tree detection based on deep learning," in *IEEE Global Conference on Signal and Information Processing*, 2017.

[36] 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 *Joint Urban Remote Sensing Event*, 2017.

[37] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 937–949, February 2018.

[38] D. Ienco, R. Interdonato, R. Gaetano, and D. Ho Tong Minh, "Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 11–22, December 2019.

[39] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2015.

[40] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5553–5563, 2016.

[41] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Multisource region attention network for fine-grained object recognition in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4929–4937, 2019.

[42] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sensing*, vol. 12, no. 2, p. 207, 2020.

[43] J. Xu, S. Wan, P. Jin, and Q. Tian, "An active region corrected method for weakly supervised aircraft detection in remote sensing images," in *International Conference on Digital Image Processing*, 2019.

[44] L. Zhang, J. Ma, X. Lv, and D. Chen, "Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 1, pp. 117–121, 2019.

[45] J. Ma, L. Zhang, and Y. Sun, "ROI extraction based on multiview learning and attention mechanism for unbalanced remote sensing data set," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[46] M. U. Ali, W. Sultani, and M. Ali, "Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 115–124, April 2020.

[47] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 182–196, December 2018.

[48] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 188–199, March 2019.

[49] D. Laumer, N. Lang, N. van Doorn, O. Mac Aodha, P. Perona, and J. D. Wegner, "Geocoding of trees from street addresses and street-level images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 125–136, April 2020.

[50] F. E. Fassnacht, H. Latifi, K. Sterenczak, A. Modzelewska, M. Lefsky, L. T. Waser, C. Straub, and A. Ghosh, "Review of studies on tree species classification from remotely sensed data," *Remote Sensing of Environment*, vol. 186, pp. 64–87, 2016.

[51] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[52] G. Aresta, et al., "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019.

[53] T. Araujo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polonia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS ONE*, vol. 12, no. 6, 2017.

[54] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification,"

in *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1868–1873, 2017.

[55] A. Golatkar, D. Anand, and A. Sethi, "Classification of breast cancer histology using deep learning," in *International Conference Image Analysis and Recognition*, pp. 837–844, Springer, 2018.

[56] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "From patch-level to ROI-level deep feature representations for breast histopathology classification," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, (San Diego, California), 2019.

[57] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri, "Patch-based system for classification of breast histology images using deep learning," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 90–103, 2019.

[58] K. Das, S. P. K. Karri, A. G. Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in *IEEE International Symposium on Biomedical Imaging*, pp. 1024–1027, 2017.

[59] Y. Feng, L. Zhang, and J. Mo, "Deep manifold preserving autoencoder for classifying breast cancer histopathological images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[60] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2018.

[61] C. C. de Amorim, D. Macêdo, and C. Zanchettin, "Spatial-temporal graph convolutional networks for sign language recognition," in *International Conference on Artificial Neural Networks*, pp. 646–657, 2019.

[62] H. Sharma, N. Zerbe, S. Lohmann, K. Kayser, O. Hellwich, and P. Hufnagl, "A review of graph-based methods for image analysis in digital histopathology," *Diagnostic Pathology*, vol. 1, no. 1, 2015.

[63] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. Van Der Laak, "Automated detection of DCIS in whole-slide H&E stained breast histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2141–2150, 2016.

[64] Y. Zhou, S. Graham, N. Alemi Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot, "CGC-Net: Cell graph convolutional network for grading of colorectal cancer histology images," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[65] P. Pati, G. Jaume, L. A. Fernandes, A. Foncubierta-Rodríguez, F. Feroce, A. M. Anniciello, G. Scognamiglio, N. Brancati, D. Riccio, M. Di Bonito, G. De Pietro, G. Botti, O. Goksel, J.-P. Thiran, M. Frucci, and M. Gabrani, "Hact-net: A hierarchical cell-to-tissue graph neural network for histopathological image classification," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, (Cham), pp. 208–219, Springer International Publishing, 2020.

[66] D. Anand, S. Gadiya, and A. Sethi, "Histographs: graphs in histopathology," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, vol. 11320, pp. 150 – 155, 2020.

[67] C. Li, X. Wang, W. Liu, L. J. Latecki, B. Wang, and J. Huang, "Weakly supervised mitosis detection in breast histopathology images using concentric loss," *Medical Image Analysis*, vol. 53, pp. 165–178, 2019.

[68] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE Transactions on Cybernetics*, 2019.

[69] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[70] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. Kim, L. Moy, K. Cho, *et al.*, "An interpretable classifier for high-resolution breast

cancer screening images utilizing weakly supervised localization," *arXiv preprint arXiv:2002.07613*, 2020.

[71] G. Liang, X. Wang, Y. Zhang, and N. Jacobs, "Weakly-supervised self-training for breast cancer localization," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1124–1127, IEEE, 2020.

[72] B. Xu, J. Liu, X. Hou, B. Liu, J. Garibaldi, I. O. Ellis, A. Green, L. Shen, and G. Qiu, "Attention by selection: A deep selective attention approach to breast cancer classification," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1930–1941, 2019.

[73] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1306–1315, 2019.

[74] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[75] J. Li, W. Li, A. Sisk, H. Ye, W. D. Wallace, W. Speier, and C. W. Arnold, "A multi-resolution model for histopathology image classification and localization with multiple instance learning," *arXiv preprint arXiv:2011.02679*, 2020.

[76] Natural Resources Conservation Service of the United States Department of Agriculture, "USDA Plants," October 2016.

[77] City of Seattle, Department of Transportation, "Seattle street trees," October 2016.

[78] A. Basavanhally and A. Madabhushi, "EM-based segmentation-driven color standardization of digitized histopathology," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, vol. 8676, 2013.

[79] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[80] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2846–2854, 2016.

[81] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4476–4484, 2017.

[82] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[83] J. G. Elmore, G. M. Longton, M. S. Pepe, P. A. Carney, H. D. Nelson, K. H. Allison, B. M. Geller, T. Onega, A. N. A. Tosteson, E. Mercan, L. G. Shapiro, T. T. Brunye, T. R. Morgan, and D. L. Weaver, "A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis," *Journal of Pathology Informatics*, vol. 8, no. 1, pp. 1–12, 2017.

[84] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2424–2433, 2016.

[85] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[86] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 648–656, 2015.

[87] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, 2019.

[88] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[89] K. K. Thekumparampil, C. Wang, S. Oh, and L.-J. Li, "Attention-based graph neural network for semi-supervised learning," *arXiv preprint arXiv:1803.03735*, 2018.

[90] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European Conference on Computer Vision*, pp. 646–661, Springer, 2016.

[91] X. Jia, S. Song, W. He, Y. Wang, H. Rong, F. Zhou, L. Xie, Z. Guo, Y. Yang, L. Yu, *et al.*, "Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes," *arXiv preprint arXiv:1807.11205*, 2018.

[92] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[93] H. D. Couture, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer, "Multiple instance learning for heterogeneous images: Training a cnn for histopathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 254–262, Springer, 2018.

[94] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[95] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[96] V. P. Bhapkar, "A note on the equivalence of two test criteria for hypotheses in categorical data," *Journal of the American Statistical Association*, vol. 61, no. 313, pp. 228–235, 1966.