Aerial Scene Parsing: From Tile-level Scene Classification to Pixel-wise Semantic Labeling

Yang Long[a], Gui-Song Xia[b,a,*], Liangpei Zhang[a], Gong Cheng[c] and Deren Li[a]

[a] State Key Lab. LIESMARS, Wuhan University, Wuhan 430079, China
[b] School of Computer Science, Wuhan University, Wuhan 430079, China
[c] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Keywords:
Aerial image interpretation
Million-AID
scene classification
semantic segmentation
transfer learning

ABSTRACT

Given an aerial image, aerial scene parsing (ASP) targets to interpret the semantic structure of the image content, *e.g.*, by assigning a semantic label to every pixel of the image. With the popularization of data-driven methods, the past decades have witnessed promising progress on ASP by approaching the problem with the schemes of *tile-level scene classification* or *segmentation-based image analysis*, when using high-resolution aerial images. However, the former scheme often produce results with tile-wise boundaries, while the latter one needs to handle the complex modeling process from pixels to semantics, which often requires large-scale and well-annotated image samples with pixel-wise semantic labels. In this paper, we address these issues in aerial scene parsing, with perspectives form tile-level scene classification to pixel-level semantic labeling. To this end, we first review aerial image interpretation by revisiting its development outline. We then present a large-scale scene classification dataset, termed Million-AID, which consists of more than a million aerial images. With the presented dataset, we also report benchmarking experiments using convolutional neural networks (CNNs). Finally, we perform experiments on aerial scene parsing that unifies the tile-level scene classification and object-based image analysis to achieve pixel-level semantic labeling. Intensive experiments show that Million-AID is a challenging dataset, which can serve as a benchmark for evaluating new developed algorithms. When transferring knowledge from Million-AID, CNN models pre-trained on Million-AID show considerable superiority than those on ImageNet for aerial scene classification, demonstrating the strong generalization ability of the proposed dataset. Moreover, our designed hierarchical multi-task learning methods achieves the state-of-the-art performance of pixel-level classification on the challenging GID, which is a profitable attempt to bridge the tile-level scene classification toward pixel-level semantic labeling for aerial image interpretation. We hope that our work could serve as a baseline for aerial scene classification and inspire rethinking the scene parsing of high-resolution aerial images.

1. Introduction

Aerial image understanding is a task of primary importance for a wide range of applications such as agriculture production (Weiss et al., 2020), urban planning (Wellmann et al., 2020), and environmental monitoring (Yuan et al., 2020). An essential way toward understanding an aerial image is to perform a full-scene semantic structure interpretation, also denoted as aerial scene parsing, which aims to label each pixel in the image with a semantic category to which it belongs. With more and more aerial images being available, aerial scene parsing has been a momentous but active topic in the field of remote sensing (Porway et al., 2010; Zheng et al., 2020; Zhou et al., 2021; Wang et al., 2021). Besides, pixel semantics acquired by aerial scene parsing are usually demanded as imperative prerequisites in practical applications like land use/land cover investigation (Tong et al., 2020; Liu et al., 2020; Lv et al., 2021). However, aerial images taken from bird's view with large imaging angle are always characterized with large scale, which implies that conventional computational method with pixel-wise analysis is hard to fulfill a full aerial scene parsing. Moreover, the highly complex content and image structure further increase the difficulty of identifying semantics of pixels in aerial images.

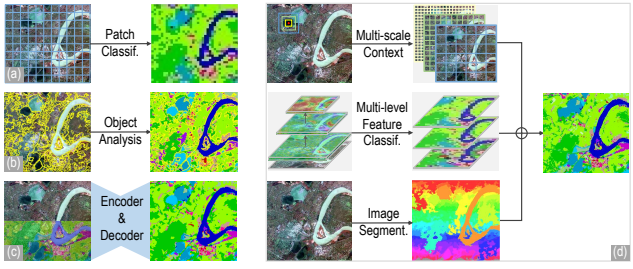


Figure 1: Aerial scene parsing based on (a) patch classification, (b) object analysis, (c) end-to-end semantic segmentation, and (d) our proposed method. Our basic idea is to utilize multi-scale contextual information and multi-level convolutional features for tile-level scene classification, where image segmentation is incorporated to extract homogeneous region for producing pixel-level semantic boundaries.

Faced with this situation, aerial scene parsing has been simplified as tile-level scene classification which integrates the complicated image features and content as a whole for semantic classification (Yang and Newsam, 2010; Sheng et al., 2012; Hu et al., 2015; Yang et al., 2015; Chen and Tian, 2015; Xia et al., 2017; Cheng et al., 2017, 2020; Chen and Tsou, 2021). However, the tile-level scene classification focuses on summarizing the thematic content within

*Corresponding author
guisong.xia@whu.edu.cn (G. Xia)
ORCID(s):

a local area while accurate semantics of individual pixels cannot be identified. To remedy this defect, the scene image can be divided into different regions for semantic identification. A feasible way is to split the scene image into regular grids, on which the patch-based classification can be performed for aerial scene parsing (Sharma et al., 2017; Paoletti et al., 2018; Sharma et al., 2018; Jean et al., 2019; Liu and Shi, 2020). Nevertheless, it usually produce coarse maps with blurred object edges. Moreover, object-based image analysis (OBIA) that extracts homogeneous entities by over-segmentation has been widely employed to achieve pixel-level semantic labeling (Blaschke and Strobl, 2001; Blaschke, 2010; Blaschke et al., 2014; Hossain and Chen, 2019; Martins et al., 2020). Even with great success, aerial scene parsing rely on OBIA suffers from parameter optimization (Ming et al., 2015; Ma et al., 2015), feature selection (Ma et al., 2017), and modeling the complex relationship among different objects when reasoning their semantic meaning (Porway et al., 2010). Recently, semantic segmentation based on fully convolutional networks (FCNs) (Long et al., 2015; Ronneberger et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2017) provides an end-to-end aerial scene parsing framework that has been intensively approached (Zhu et al., 2017; Li et al., 2019; Audebert et al., 2019; Zheng et al., 2020; Zhou et al., 2021; Wang et al., 2021). Still, the optimization of FCN methods requires a large amount of pixel-wise annotation which is extremely labor-intensive and time-consuming to produce. Figure 1(a)∼(c) illustrate these conventional methods for aerial scene parsing based on patch classification, object analysis, and end-to-end semantic segmentation, respectively. From the overall perspective, the above methods for aerial scene parsing are typically performed in a separated way, of which advantages cannot be integrated. To change this situation, a great deal of effort must be paid by addressing the following critical issues:

- *The divergence of aerial scene parsing prototypes in tile-level and pixel-wise classification.* Typically, pixel-wise classification has been intensively approached to produce fine-gained labeling result (Lu and Weng, 2007; Chen and Peter Ho, 2008; Tuia et al., 2011) while tile-level classification can only provides coarse semantic description of regions (Yang and Newsam, 2010; Yang et al., 2015; Cheng et al., 2017; Xia et al., 2017; Cheng et al., 2020). Currently, the continuous improvement of image resolution has brought us aerial images of large scale and rich detail. As a result, aerial scene parsing by pixel-wise classification becomes a challenging task because of the variant attributes of pixels and high computational cost. From the perspective of image expression, the improvement of image resolution also greatly enhances the semantic homogeneity of pixels in local regions. Thus, the semantics of individual pixels in an aerial image are closely related to their contextual information rather than rely solely on its own. In this context, it is reasonable to consider the pixel as a unit centred by tile-level scene and bridge its gap to pixel-level semantic labeling.

- *The scarcity on exploring the transferability of semantic scene knowledge of aerial images.* Currently, the lack of large-scale benchmark datasets has become a bottleneck that hamper the development of data-driven methods for aerial image interpretation, particularly deep learning-based ones. To alleviate this situation, the conventional way is to employ CNN models pre-trained on natural image archives (*e.g.*, ImageNet (Deng et al., 2009)) as feature extractor or fine-tune them on target aerial images. However, there are great differences between natural and aerial images in spectral properties, image structure, and spatial arrangement. Moreover, the semantic categories that define the scene content also vary significantly between natural and aerial images. Thus, the learned features obtained through above methods can be biased in characterizing aerial image content. With this in mind, it is of great significance to explore the transferability ability of data-driven models adapted with pure aerial scene images and free up their potential for aerial image interpretation.

With these points in mind, this paper addresses aerial scene parsing that unifies tile-level scene classification and OBIA to achieve pixel-level semantic labeling as illustrated in Figure 1(d). In doing so, we first provide a review to depict aerial image interpretation. Then, we present a large-scale aerial scene classification dataset, *i.e.*, Million-AID, on which the benchmarking experiments are performed to investigate how well the current CNNs perform for aerial scene classification. Finally, we conduct aerial scene parsing from tile-level scene classification to pixel-level semantic labeling, where knowledge transfer by Million-AID is performed to improve the accuracy. To sums up, our main contributions are as follows:

- We provide a comprehensive review on aerial image interpretation by revisiting its development outline, ranging from pixel-wise image classification, segmentation-based image analysis, and tile-level image understanding that are connected to the improvement of spatial resolution of aerial images.

- We released the large-scale dataset, i.e., Million-AID, for aerial scene recognition. We also investigated a number of classical CNN models to perform multi-class and multi-label scene classification on Million-AID. The benchmarking results indicate that Million-AID is a challenging dataset, which can provide the research community an evaluation and comparison platform for aerial scene classification algorithms.

- We conducted extensive experiments to verify the tremendous potential of transferring scene knowledge of Million-AID to advance aerial scene parsing from

Figure 2: The development outline of aerial image interpretation. The interpretation prototypes develop with the resolution improvement of aerial images and has experienced a long course of development ranging from pixel-wise classification, segmentation-based classification, to tile-level classification. This figure non-exhaustively presents some representative works concerning aerial image recognition. In this work, we will make our efforts to perform semantic classification that bridges tile-level scene classification toward pixel-level semantic labeling for high-resolution aerial scene parsing. Zoom for detail.

tile-level scene classification to pixel-level semantic labeling. Fine-tuning CNN models pre-trained on Million-AID show considerable superiority than those on ImageNet for aerial scene classification. The designed hierarchical multi-task learning method by employing Million-AID and GID achieves the state-of-the-art result for pixel-level image classification, which demonstrate the effectiveness of bridging tile-level classification to pixel-level labeling for aerial image interpretation.

The remainder of this paper is organized as follows. Section 2 presents the review of aerial image interpretation. Section 3 introduces the proposed large-scale scene classification dataset, *i.e.*, Million-AID. Section 4 presents the comprehensive benchmarking experiments on Million-AID, including multi-class and multi-label aerial scene classification. Section 5 presents experiments tile-level scene classification and pixel-level semantic labeling with knowledge transfer from Million-AID. Finally, in Section 6, we draw conclusions regarding this work.

2. Revisiting Aerial Image Interpretation

With the progress of sensor technology, the spatial resolution of aerial image has experienced a continuous improvement (Toth and Jóźków, 2016; Pu and Landry, 2012). Figure 2 presents the milestones of earth observation satellites at different times. Accordingly, the improvement of aerial image resolution has greatly promoted the development of aerial image interpretation as stated previously. In this section, we focus on a review by revisiting the interpretation tasks of aerial image and the outline is presented in Figure 2.

2.1. Pixel-wise aerial image classification

In the early 1970s, aerial images are characterized with low spatial resolution (*e.g.*, LandSat-1 and MODIS images) where each pixel describes an area of thousands of square meters of the Earth's surface. And the sizes of ground features or objects are usually smaller than the ground sampling distance of image pixels. Thus, each pixel is able present a scene of specific semantic category. Individual pixels are obviously distinct from each other owing to the difference of covered ground features. In this situation, semantic interpretation of aerial images mainly focuses on pixel-wise classification using spectral signatures (Swain and Ballard, 1991; Ojala et al., 2002) and coarse textural features (Haralick et al., 1973; Manjunath and Ma, 1996; Ojala et al., 2000; Xia et al., 2010a). To this end, sampling analysis is naturally employed (Van Genderen et al., 1978; Curran and Williamson, 1986; Khatami et al., 2017) to construct desirable classification schemes. Typically, training samples that are representative to reflect the distribution and variation of diverse semantic content of interest are selected to extract category information. Thus, classification methods based on statistical analysis are widely employed by estimating the probability of a pixel belonging to each of the possible classes (Bruzzone et al., 1999; Chen and Peter Ho, 2008;

Li et al., 2010; Camps-Valls et al., 2013; Li et al., 2011; Zhao et al., 2016).

In order to obtain reliable classification results, a number of classifiers were also developed for the pixel-wise aerial image parsing, such as maximum likelihood methods (Settle and Briggs, 1987; Ediriwickrema and Khorram, 1997; Peng et al., 2019b), minimum distance to means algorithms (Hodgson, 1988; Espínola et al., 2014), K-nearest neighbors classifiers (Ma et al., 2010; Tu et al., 2018), and tree-based techniques (Friedl and Brodley, 1997; Otukei and Blaschke, 2010). However, statistical classification methods usually show insufficient ability in discriminating pixel units due to the variation of spectral and spatial characteristics influenced by imaging condition and ground feature attributes. Faced with this situation, more sophisticated classifiers such as random forest (Belgiu and Drăguţ, 2016; Xia et al., 2018b; Izquierdo-Verdiguier and Zurita-Milla, 2020; Zafari et al., 2020), sparse representation (Lee et al., 2007; Liu et al., 2013; Feng et al., 2017; Fan et al., 2017; Peng et al., 2019b; Li et al., 2020a), artificial neural network (Kavzoglu and Mather, 2003; Shao and Lunetta, 2012), and kernel-based methods (Liu et al., 2013; Gu et al., 2017; Li et al., 2020a; Zafari et al., 2020) represented by support vector machine (Mountrakis et al., 2011; Maulik and Chakraborty, 2017; Okwuashi and Ndehedehe, 2020) were intensively explored by actively embracing the machine learning techniques. These methods have made dramatic progress in pixel-wise aerial image parsing owing to their strong ability of discriminating the complex spectral and spatial characteristics of ground features.

Pixel-wise classification approaches assume that each pixel only belong to single semantic category and different categories are mutually exclusive. However, such an assumption can be inconsistent with the reality due to the limitation of spatial resolution of aerial images (Lu and Weng, 2007; Li et al., 2014a). Specifically, more than one object or ground feature belonging to different semantic categories can be contained within a pixel as their scales are smaller than the spatial resolution of aerial images. As a result, the existence of mixed pixels comes to be a nonnegligible problem in the medium and coarse spatial resolution aerial images. This could lead to an appreciable interpretation result when employing the pixel-wise classification strategy. To overcome this problem, sub-pixel classification is considered as an alternative for more accurate aerial image interpretation (Liu and Wu, 2005; Bovolo et al., 2010; Wang et al., 2020; He et al., 2021).

A number of approaches have been derived to address the sub-pixel aerial image classification, including soft or fuzzy theory (Wang, 1990; Shackelford and Davis, 2003; Kothari et al., 2020), neural networks (Mertens et al., 2004; Li et al., 2014b; He et al., 2020), regression modeling and analysis (Fernandes et al., 2004; Gessner et al., 2013; Cooper et al., 2020), and spectral mixture analysis (Wu and Murray, 2003; Somers et al., 2011; Yu et al., 2021; Xu and Somers, 2021). Among these methods, the fuzzy technique and spectral mixture analysis are most commonly employed to overcome the mixed pixel problem. Particularly, fuzzy representation is developed to estimate multiple and partial memberships of all candidate categories within a pixel, where the corresponding areal proportion of each category can be acquired. The spectral mixture analysis assumes the value of a pixel is a linearly or non-linearly combination of a set of specific endmember spectra (Somers et al., 2011; Yu et al., 2021). Thus, the selection of endmembers becomes the key point for designing an effective classifier (Bateson and Curtiss, 1996; Small, 2004; Ozkan et al., 2019; Hong et al., 2021). Even with great improvement of classification accuracy, sub-pixel class composition estimated by fuzzy classification and spectral mixture analysis cannot provide the spatial distribution of land cover classes within pixels. To address this issue, the sub-pixel mapping approaches are developed (Atkinson, 1997; Wang and Atkinson, 2017; Wang et al., 2020; He et al., 2020). In this scheme, each pixel is divided into sub-pixels which are predicted to get single semantic labels. Limited by the spatial resolution, aerial images interpreted by pixel-wise classification still face challenges in acquiring satisfactory result due to the mixture and complexity of image content within single pixels.

2.2. Segmentation-based aerial image analysis

With the development of sensor technology, the spatial resolution of aerial images is gradually improved to be much smaller than the scales of ground features. Thus, the detail of spectrum, texture, and geometric structure in the image becomes prominent. Under the circumstances, single pixels are no longer isolated units since the ground features and objects could be composed of a certain number of pixels knitted into an image full of spatial patterns (Hay and Castilla, 2008). And the improved image quality also significantly increases the within-class variability, which decreases the potential accuracy of pixel-based approach to classification (Blaschke et al., 2014). As a result, traditional interpretation system established with pixel-wise statistics and analysis for low-resolution aerial images, to some extent, is beginning to show cracks in classifying aerial images for required accuracy and generalization ability (Blaschke, 2010).

Faced with this situation, researchers turn their attention to the new paradigm of object-based image analysis (OBIA) or geographic-object-based image analysis (GEOBIA) (Blaschke and Strobl, 2001; Hay and Castilla, 2008; Blaschke, 2010; Blaschke et al., 2014; Ma et al., 2017; Hossain and Chen, 2019; Kotaridis and Lazaridou, 2021), where geographical or image objects are considered as the basic units instead of individual pixels for image classification. The objects are considered to be homogeneous entities, located within an image and perceptually generated from pixel-groups, where each pixel-group is composed of similar digital values, and possesses an intrinsic size, shape, and geographic relationship with the real-world scene component it models (Hay et al., 2001). In general, the OBIA generates objects by image segmentation and then performs image classification on objects. Thus, image segmentation serves as the initial and critical part to produce the funda-

mental elements of OBIA (Martins et al., 2020; Kotaridis and Lazaridou, 2021).

In high-resolution remote sensing images, ground objects are presented with much richer spectral, textural, structural, and contextual detail that reveals the pattern characteristics (Blaschke et al., 2014). It enables the objects of interest to be extracted by spectrally-based and spatially-based segmentation approaches, among which mathematical morphology analysis plays a significant role (Cheng et al., 2001; Hossain and Chen, 2019). Hence, the thresholding (Martha et al., 2011; Yang et al., 2017; Tang et al., 2020) and feature space clustering (Amitrano et al., 2018) methods are typically employed to generate objects by spectral analysis based on the fact that homogeneous objects share similar spectral characteristics. For spatially-based segmentation, edge detection (Kaur and Garg, 2011; Yang et al., 2017; Han et al., 2018; Shang et al., 2021) and region generation (*e.g*, region growing (Martha et al., 2011; Liu et al., 2015), merging, and splitting (Yang et al., 2017; Su, 2019; Su et al., 2020)) techniques are conducted according to the discontinuity and similarity of object areas, respectively. However, edge-based methods are precise in boundary detection while facing problem in generating closed segments. By contrast, region-based methods have the advantage in generating closed regions while resulting in imprecise segment boundaries. As compensation, hybrid segmentation (Zhang et al., 2014; Yang et al., 2017; Niu et al., 2021) that consider both the boundary and spatial information between adjacent regions show significant advantages in object segmentation.

However, objects acquired by segmentation can only represent homogeneous regions lacking semantic description. Thus, the object features are then extracted and embedded into classifier to determine the semantic categories. In doing so, both the feature extraction and classifier design play crucial roles in classification performance (Ghamisi et al., 2017; Kumar et al., 2020). Owing to the overwhelming advantages in visual feature extraction and classification, CNN frameworks have recently been integrated into OBIA and triggered the new trend of object-based CNN (OCNN) for aerial image interpretation (Zhang et al., 2018a, 2020a; Martins et al., 2020). With the availability of high-resolution aerial images, object-based approaches become dominant in the task of aerial image interpretation over the past two decades. Even with significant performance advantage compared with pixel-wise classification methods, the object-based classification methods face challenges in parameter setting and optimization (*e.g.*, segmentation scale) (Ming et al., 2015; Ma et al., 2015), which affect the segmentation quality as well as the final classification accuracy. In addition, the segmentation and classification of objects falls into a complicated and multi-step implementation pipeline, which inevitably limits the efficiency and increases the difficulty of model deployment.

To overcome the above deficiencies, the solution that simultaneously produces the segmented homogeneous areas and corresponding semantic categories becomes an imperious demand. In recent years, aerial image classification have been greatly facilitated by deep convolutional neural networks (DCNNs) (Maggiori et al., 2016; Li et al., 2019; Audebert et al., 2019; Jia et al., 2021), among which the fully convolutional network (FCN) (Long et al., 2015) and the improved architectures (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2017) provide an end-to-end segmentation and classification pipeline. In contrast with conventional methods, the advantage of DCNN lies in its capacity to extract shallow visual and deep semantic features by the elaborately designed hierarchical framework (Zhu et al., 2017). However, the down-sampled features in deep layers will lead to the resolution degradation for the final classification results, in which the uncertainty of boundaries and detail of different classes is a serious issue. To deal with these issues, the multi-scale features (Zhang et al., 2019; Yang et al., 2019; Peng et al., 2019a; Sun et al., 2020) and contextual information (Zhang et al., 2018b, 2020a) are typically considered to enhance the feature representation ability. The atrous and paralleled dilation convolution (Chen et al., 2017) is utilized to preserve feature resolution (Niu et al., 2018). With the improvement of network architectures, classification accuracy of aerial images has witnessed a continuous improvement.

Apart from the evolution of networks, significant efforts have been paid to improve the performance of classification models. Typically, the spatial and spectral attributes of aerial images have been intensively addressed in pixel-wise classification (He et al., 2017; Ghamisi et al., 2018; Gao et al., 2019). Particularly, the integration of spatial and spectral information are commonly employed to address the challenge of large spatial variability of spectral signatures (Mou and Zhu, 2019; Feng et al., 2019; Li et al., 2020b; Imani and Ghassemian, 2020). Regarding the limitation of training samples and generalization ability, transfer leaning has been intensively explored to address the limitation of training samples for CNN frameworks and reported promising classification results (Tong et al., 2020; Wurm et al., 2019; Chen et al., 2019a; He and Chen, 2020). The readers may go to one of the review papers for more comprehensive perspective of semantic segmentation using deep learning techniques (Zhu et al., 2017; He et al., 2017; Ghamisi et al., 2018; Ma et al., 2019; Imani and Ghassemian, 2020; Lateef and Ruichek, 2019; Minaee et al., 2021). However, owing to the lack of large-scale datasets, many interpretation algorithms are locally-oriented, typically manifested in the validation of one or several images within local areas which would affect the generalization ability. And the CNN-based methods also suffer from computational burden when classifying aerial images of large size and huge volume due to the improvement of spectral and spatial resolution.

2.3. Tile-level aerial image classification

Even with impressive success achieved by object-based analysis, individual objects carry information independent to their neighbors and thus neglect the thematic meaning in their contextual environment, which could lead to inaccurate classification result. To alleviate this problem, the hierarchi-

cal and contextual model is developed by organizing individual objects into hierarchical groups for aerial image parsing (Porway et al., 2010). However, the implementation of object detection and hierarchical contextual representation is complicated. Thus, classification of tile-level scene, which is able to incorporate visual features, content components, and spatial arrangements as a whole, becomes an effective way for aerial image interpretation. In the last decade, a handful of visual descriptors have been employed for aerial scene classification (Yang and Newsam, 2010; Xia et al., 2010b; Chen and Tian, 2015; Zhong et al., 2015; Cheng et al., 2017; Xia et al., 2017; Cheng et al., 2020). We refer interested readers to (Yang and Newsam, 2010; Cheng et al., 2017; Xia et al., 2017) for a survey of the low-level and middle-level visual features employed for aerial scene classification.

Owing to the increasing accessibility of aerial images, the data-driven approaches particularly CNN-based ones have shown great advantage over the handcrafted-feature-based approaches for aerial scene classification. In the beginning, pre-trained CNNs are usually employed as feature extractors owing to its simplify and efficiency (Hu et al., 2015; Li et al., 2017; He et al., 2018). However, the representation of aerial scenes is a challenging task owing to the complexity of scene components and scale variation. To improve the feature discrimination ability, multi-scale images or features are extracted and fused to generate robust global representation for scene classification (Li et al., 2017; He et al., 2018; Raza et al., 2020; Bi et al., 2021b). In fact, the semantic category of an aerial scene usually depends on the spatial arrangement and class-specific objects in the image. Thus, deep local structures related to scene category are addressed to improve the classification performance (Yuan et al., 2019; Bi et al., 2021a). Recently, CNNs based on attention mechanism have been addressed to highlight more local semantics and discard the noncritical information (Wang et al., 2019; Bi et al., 2020; Fu et al., 2020; Bi et al., 2021b). In general, learning powerful feature for content representation is of great importance for aerial scene recognition.

With the improvement of spatial resolution of aerial images, the within-class diversity and between-class similarity of semantic scenes have been greatly increased, which make the scene recognition a challenging task. To relieve this issue, deep metric learning algorithms are developed to learn discriminative category features (Cheng et al., 2018; Gong et al., 2018; Kang et al., 2020). Particularly, the complex relationship pervading aerial scenes are further explored in the embedding space by learning deep graph networks (Wang et al., 2018; Khan et al., 2019; Kang et al., 2021). The main idea is to map the scene features closely to each other for the same categories while as farther apart as possible for different categories. As conventional CNNs with a fixed architecture may show limitation in grasping the scene content of large diversity, automatically learning CNN architecture specified for aerial scenes has been intensively explored (Ma et al., 2021a; Peng et al., 2021; Ma et al., 2021b; Broni-Bediako et al., 2021) and achieved encouraging results.

However, the methods based on deep learning require large-scale annotated samples for model adaption while most of them are trained and tested on relatively small-scale datasets. To overcome this problem, scene classification based on few-shot learning has recently attracted extensive attention (Cheng et al., 2021; Li et al., 2021; Liu et al., 2021). Moreover, annotated scene images from different domains are also employed to relieve the issue of data dependency by transfer learning (Othman et al., 2017; Lu et al., 2020; Zhang et al., 2020b; Zhu et al., 2021). These approaches have reported exciting performance on aerial scene classification. However, the recent scene classification algorithms have intensively reported saturation results as shown in (Cheng et al., 2020). Faced with this situation, the potential of the data-driven methods for scene classification remains to be further explored and boosted by large-scale datasets.

The aforementioned prototypes have achieved great success in aerial image interpretation. However, conventional OBIA method usually results in the complicated process for semantic reasoning while end-to-end semantic segmentation requires a large amount of pixel-level annotations for model adaption. The tile-level image classification also show deficiency in producing pixel-level semantic boundaries. From a geographical perspective, the discrimination of a ground object rely heavily on to the background environment. Hence, the semantics of individual pixels are closely related to their surrounding neighbors in high-resolution aerial images. In this situation, it is reasonable to employ the tile-level scene to incorporate the contextual information of its central pixels and then predict the semantic meaning. And the acquisition of tile-level scene labels is much easier than those of pixel-level ones. With these in mind, we aim to unify the aforementioned prototypes and perform aerial scene parsing from tile-level scene classification to pixel-level semantic labeling in this work.

3. An Introduction to Million-AID

In this section, we detail the properties of Million-AID to be released for aerial scene classification. The readers may go to (Long et al., 2021) for the construction of Million-AID.

3.1. Scene categories

The semantic scenes in Million-AID are hierarchically organized by referencing the land-use classification standards. There are 8 major classes of aerial scenes in the first level, *i.e.*, *agriculture land*, *commercial land*, *public service land*, *industrial land*, *transportation land*, *residential land*, *water area*, and *unutilized land*, covering 28 sub-classes in the second level. And more specific scene categories are organized at the third level. In total, there are 51 fine-gained scene categories, including *dry field* (DF), *greenhouse*, *paddy field* (PF), *terrace field* (TF), *meadow*, *forest*, *orchard*, *commercial area* (CA), *storage tank* (ST), *wastewater plant* (WP), *works, oil field*, *mine*, *quarry*, *solar power plant* (SPP), *wind turbine* (WT), *substation*, *swimming pool* (SP), *church*, *cemetery*, *basketball court* (BC), *tennis court* (TC), *baseball*

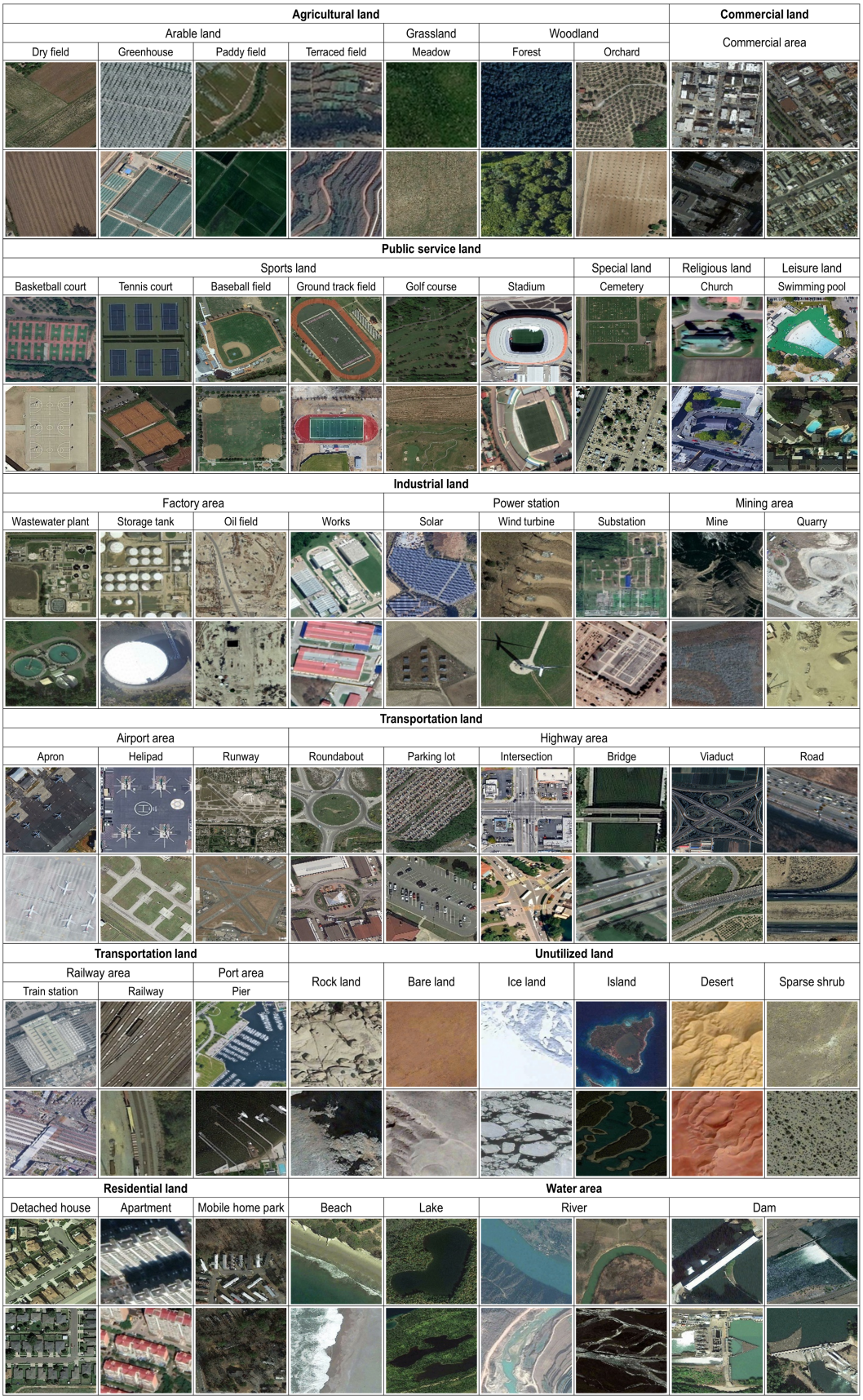


Figure 3: Scene samples of Million-AID: two or four examples of each semantic category are presented. All the semantic scenes are organized by the hierarchical system with three-level semantic labels, containing 51 fine-gained scene categories belonging to 8 major categories.

field (BF), *ground track field* (GTF), *golf course* (GC), *stadium*, *detached house* (DH), *apartment*, *mobile home park* (MHP), *apron*, *helipad*, *runway*, *road*, *viaduct*, *bridge*, *intersection*, *parking lot*, *roundabout*, *pier*, *railway*, *train station* (TS), *rock land*, *bare land*, *ice land*, *island*, *desert*, *sparse shrub land* (SSL), *lake*, *river*, *beach*, and *dam*. All labels have been checked by the specialists in the field of aerial image interpretation. Several instances in each scene class are shown in Figure 3. Moreover, each scene image can be assigned with more than on category labels according to the hierarchical semantic nodes. This property enable Million-AID to be an aerial image dataset for hierarchical multi-label scene recognition. In total, there are 73 semantic labels contained in Million-AID. In this work, we treat the 51 fine-gained scenes as independently parallel categories for multi-class (single label) scene classification and the 73 scene categories are are employed for multi-label scene classification.

3.2. Dataset scale

Apart from the wide coverage of semantic categories, Million-AID is characterized with large scale. Particularly, the total number of images in Million-AID is 1,000,848. To the best of our knowledge, this is the first aerial scene classification dataset in which the number of images exceeds a million in remote sensing community. As shown in Figure 4, the numbers of scene images varies greatly among different categories, endowing the dataset with the property of unbalanced distribution. Taking the widely used AID (Xia et al., 2017) and NWPU-RESISC45 (Cheng et al., 2017) as comparision, our proposed Million-AID surpasses them hugely in both the numbers of categories and images. Recently, data-driven methods particularly deep learning (Zhu et al., 2017; Reichstein et al., 2019; Cheng et al., 2020) have shown promising perspectives for intelligent aerial image interpretation, relying on the huge available dataset ontology. The Million-AID make it possible to further boost the design and learning of aerial scene interpretation algorithms using data-driven schemes.

3.3. Geographical distribution

With the change of geographical environment, aerial scene images usually show different patterns in appearance, structure, and content. Hence, scenes in an aerial image dataset should be as widely distributed as possible to characterize their features in the real world. To this end, we collected the aerial scenes around the world by utilizing the geographical information as introduced in (Long et al., 2021). The geographical distribution of aerial scenes in Million-AID is shown in Figure 5. It can be seen from the distribution map that the scene images are widely located all over the world. It is worth noting that most of the scene images are located on the land areas, and intensively distributed in cities or areas inhabited by humans. This is reasonable because it is in line with the reality that the semantic scenes of aerial images are usually closely associated with human production and living activities.

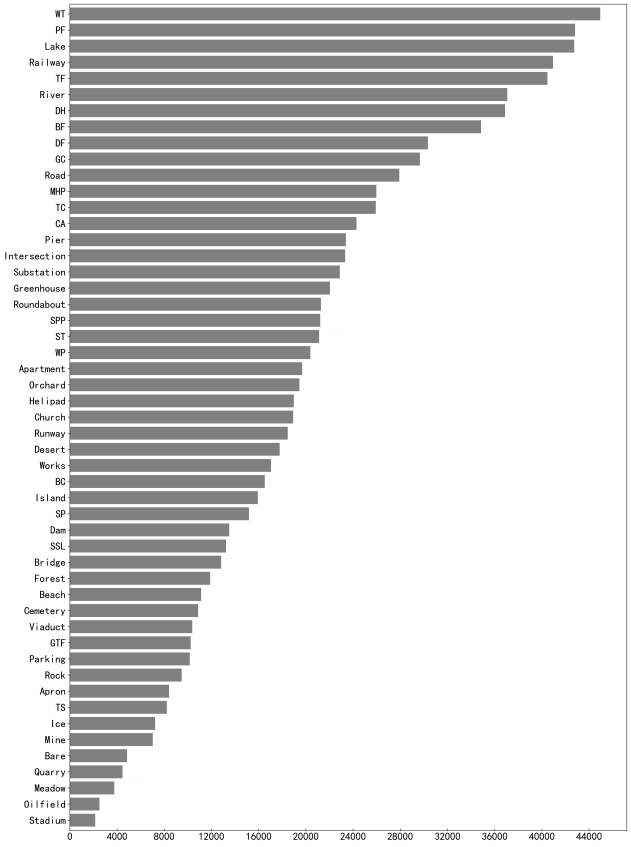


Figure 4: The number of instances in each scene category. Zoom for detail.

3.4. Image variation

Rich variation of images can greatly enhance the diversity of a dataset, so as to better represent the scene and feature distribution in the real world. In Million-AID, the widths of scene images range from 100 to 30,000 pixels to approximate the scale variation of scenes in practical situation. The spatial resolution is 0.2m to 153m per pixel. As aerial imaging is easily affected by environmental factors, scene images in Million-AID are extracted under various circumstances, *i.e.*, viewpoint, weather, illumination, season, background, scale, resolution, geographical area, *etc*. These properties reflect the real challenges in the task of aerial scene recognition.

Furthermore, owing to the high complexity of ground features, scene content in aerial images usually show remarkable difference in appearance characterized with various geometrical, structural, and textural attributes. This requires the created dataset with high intra-class diversity and inter-class similarity for developing interpretation algorithms with excellent generalization ability. The above introduced properties and variation of scene images have provided sufficient assurance of intra-class diversity for Million-AID. Besides, scene images of sub-classes are typically contained in the same major classes. This enables the scene images of the sub-classes to possess property of high inter-class similarity inherited from their common major classes. In general, the presented Million-AID is of great capacity to represent aerial scenes and feature distribution in

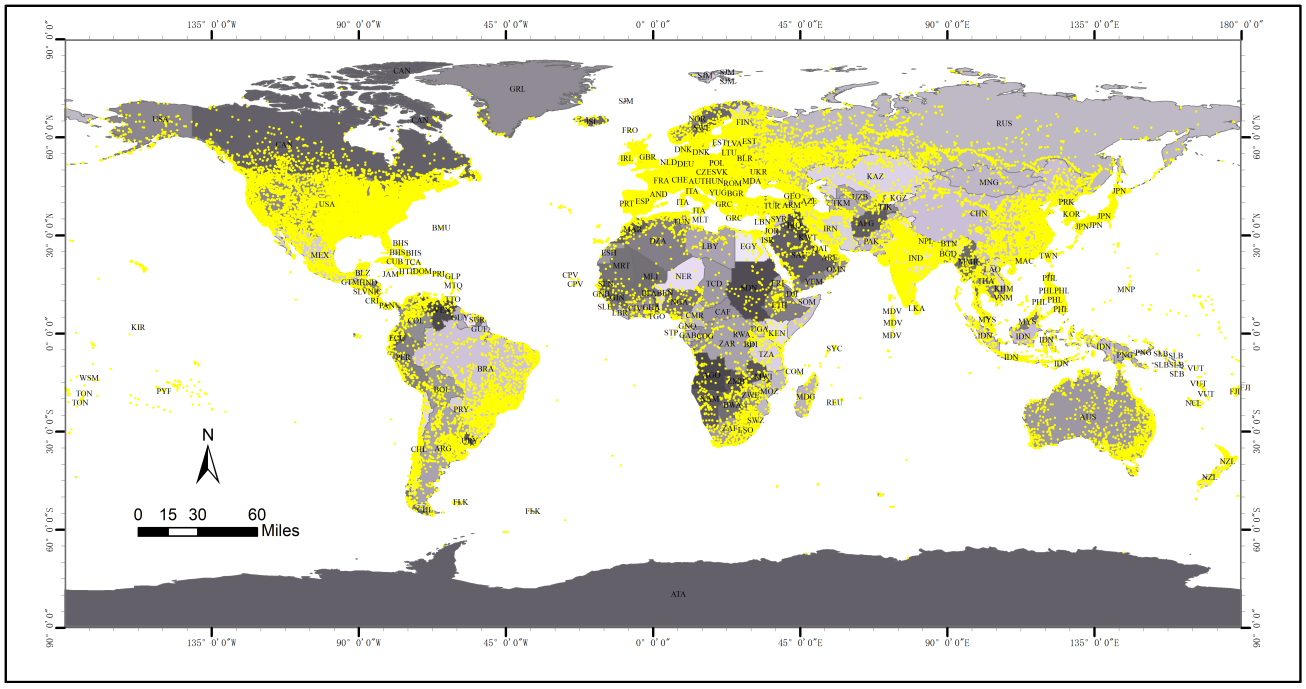


Figure 5: The distribution of scene images in Million-AID. To clearly express the image location and its distribution, more than a half of the total scene images are taken as the representative instances, and visually displayed using their geographical locations. The location of a scene image is represented by the central geographical coordinates of the image block.

the real world, and thus, facilitate the development of data-driven interpretation algorithms and establishment of public comparison platforms.

4. Scene Classification: A New Benchmark on Million-AID

Data-driven algorithms represented by deep learning have been reported with overwhelming advantages over the conventional classification methods (Xia et al., 2017; Cheng et al., 2017), and thus, dominated aerial image recognition in recent years (Cheng et al., 2020). In this section, we train a number of representative CNN models and conduct comprehensive evaluations for multi-class and multi-label scene classification on Million-AID, which we hope to provide a benchmark for future researches.

4.1. Experimental setup

Dataset partition: In order to make a comprehensive evaluation, the partition scheme is established for the baseline training and testing sets. Specifically, we extract training and test scene images located at different areas. In this configuration, we try to make the training and test data as spatially independent as possible. Consequently, there are 10,000 scene images in the whole dataset of Million-AID randomly selected as the training subset and the left images are fixed as the testing subset. Besides, the training set is characterized with long-tail distribution, which poses the great challenge to the scene classification model.

Model configuration: For image scene classification, the representative CNN models are employed for benchmarking experiments. Specifically, AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), ResNet101 (He et al., 2016), DenseNet121 (Huang et al., 2017), and DenseNet169 (Huang et al., 2017) are selected to explore their scene classification performance on Million-AID. We chose these models in consideration of their broad applications in RS image interpretation particularly in scene recognition. And it is apparently to observe that the employed CNN models consist of wide degrees of model depth and parameter scale, covering CNN frameworks from the shallow to deep ones, which can help to explore the classification performance more comprehensively and objectively. For the convenience of experimental implementation and fair performance comparison, we build an unified CNN library using PyTorch (Paszke et al., 2019) for model training and testing.

Table 1
Summary of CNN Models Used in This Work

Model	#Layers	#Param.	Acc@1 (%)	Year
AlexNet	8	60M	56.52	2012
VGG16	16	138M	73.36	2014
GoogleNet	22	6.8M	69.78	2014
ResNet101	101	44M	77.37	2015
DenseNet121	121	8M	74.43	2017
DenseNet169	169	14M	75.60	2017

Acc@1 indicates the Top-1 accuracy of CNN models tested on ImageNet.

4.2. Multi-class scene classification

4.2.1. Implementation detail

In (Xia et al., 2017) and (Cheng et al., 2017), the aerial image features were directly extracted from the CNN models

Table 2
Performance of Single-label Scene Classification with Different CNN Models (%)

Metric	AlexNet	VGG16	GoogleNet	ResNet101	DenseNet121	DenseNet169
OA	67.53	77.47	77.37	77.36	79.04	78.99
AA	63.18	74.58	74.86	74.58	76.67	76.67
Kappa	66.61	76.84	76.73	76.73	78.46	78.46

pre-trained on ImageNet and then classified by support vector machines. By contrast, we deliver an end-to-end training scheme in which the *softmax* classifiers are integrated in the original CNN models. For efficient model adaption, we employ the training strategy by fine-tuning CNN models pre-ptrined on ImageNet. For fair comparison, we keep the training parameters consistent with different models. Specifically, the number of total iteration is set to be 50 epochs for sufficient parameter adaption considering the scalable training sets and stochastic gradient descent is utilized as the optimisation strategy. The batch size is set to be 32. The initial learning rate is 0.01 and divided by 10 every 20 epochs. The weight decay and momentum are 0.005 and 0.9, respectively. The hardware is based on the Inter Xeon E5 CPU and the NVIDIA Tesla V100 GPU with 16GB memory.

4.2.2. Evaluation protocols

For performance evaluation, we employ the commonly used overall accuracy (OA), average accuracy (AA), confusion matrix (CM), and Kappa coefficient (Kappa) to measure the classification results. The OA and CM are defined as same as those in (Xia et al., 2017; Cheng et al., 2017). Specifically, the OA is defined as the number of correctly predicted images divided by the total number of predicted images in the test dataset. The OA measures the classification performance on the whole dataset from a quantitative perspective while regardless of the classification performance on the single class. By contrast, AA is calculated by the mean value of classification accuracy of all classes. The CM can present the classification performance of a model on each class. Each row of a CM represents the actual instances in a predicted class while each column reveals the predicted instances in an actual class. The CM makes it convenient to explore a model's classification capability on the confusing classes. Kappa coefficient which can be calculated on the basis of CM, is a robust measure since it takes into account the classification reliability for categorical items.

4.2.3. Experimental results

Baseline results: Table 2 illustrates the scene classification result from different CNN models. We can see that VGG16, GoogleNet, ResNet101, DenseNet121, and DenseNet169 achieve significantly better classification results when compared with AlexNet. Note that AlexNet is a shallow CNN framework with only 5 convolutional layers while the others possess more convolutional layers, which are able to extract highly abstract information for scene content representation. Thus, the deeper CNN models gains classification performance on OA, AA, and Kappa. This result demonstrates the superiority of the deep CNN frameworks, which is consistent with the classification of natural images (Russakovsky et al., 2015).

Particularly, VGG16 outperforms AlexNet and gives comparable results with some of the deeper models, *e.g.*, GoogleNet, ResNet101. This phenomenon stems from the advantage of larger scale of parameters possessed by VGG16 network. And the batch normalization operation incorporated in VGG16 network also helps to relieve the internal convariate shift problem (Ioffe and Szegedy, 2015) reflected by the complex content of aerial images. Benefiting from the elaborately designed inception module, GoogleNet is able to gather features with different receptive fields in one layer, which makes it suitable for processing aerial scene images of high variation.

Among the evaluated models, DenseNet121 and DenseNet169 outperform the others obviously. The densely connected nets can integrate features from different convolutional layers and thus enhance the representation ability of learned scene features. Note that DenseNet169 and achieves similar results with DenseNet121. This phenomenon reveals that a much deeper net would no longer bring performance improvement even with more dense connected layers. However, the OAs of all evaluated scene classification models are below 80%. Therefore, more effective algorithms are expected to be developed toward semantic scene classification of aerial images.

Analysis of different metrics: When examining the performance by different metrics, we can find that Kappa and AA perform worse than OA. This is largely caused by the heavy unbalanced instance numbers of different scene categories. By referencing the confusion matrices as shown in Figure 6, we can see that some categories with relatively large number of scene images achieves high classification. For example, *wind turbine* and *river* contain over 44k and 37k instances, respectively. And the corresponding OAs achieved by DenseNet121 are close to 1. By contrast, some categories with relatively small number of scene instances achieves lower classification accuracy. As a case in point, *stadium*, and *works* consist of only 2k and 17k instances while the corresponding OAs are only 0.49 and 0.37 by DenseNet121, respectively. As a result, the OA gains performance since it count more on the total number of instances that are correctly classified while AA and Kappa are heavily influenced by the low accuracy of poorly classified categories. The difference is evidently indicated by AA as shown in Table 2. Superficially, the unbalanced image numbers of scenes in Million-AID should be more in accordance with the scene distribution in the real world when compared with the existing scene classification datasets (Zou et al., 2015; Zhu et al., 2016; Cheng et al., 2017; Zhou et al., 2018)

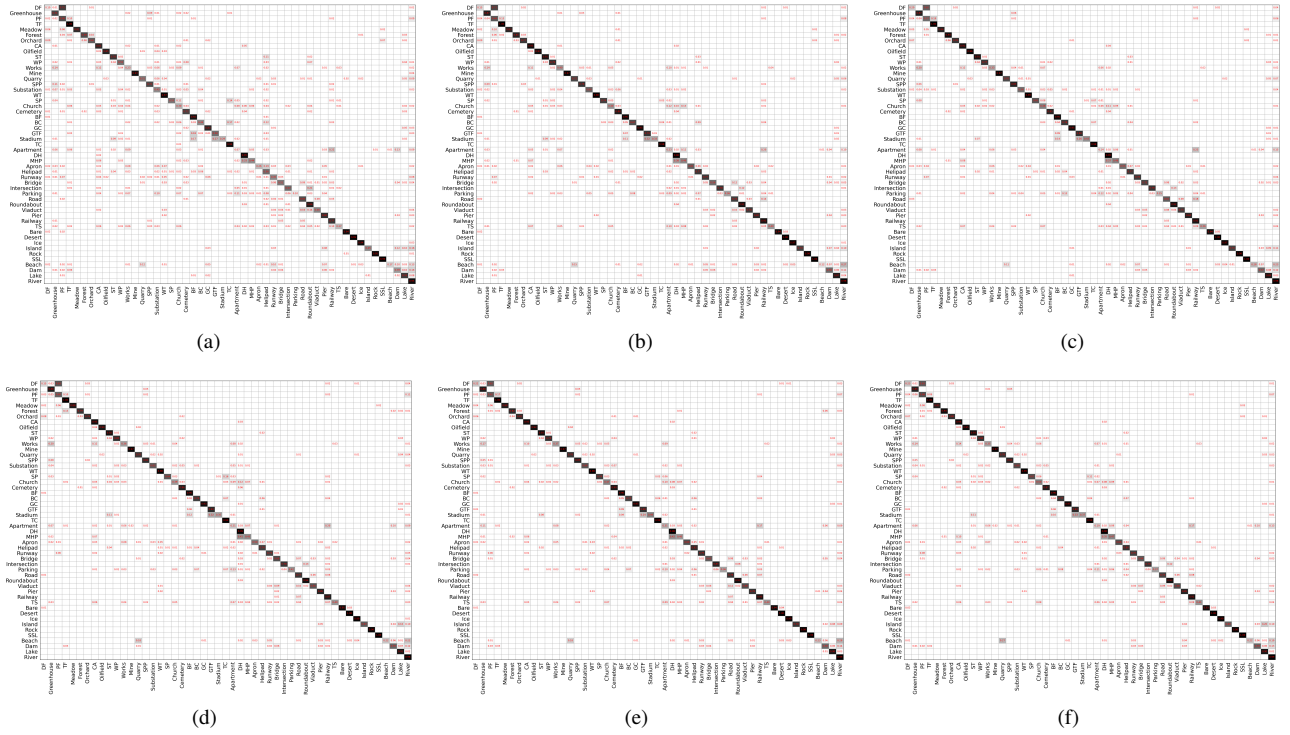


Figure 6: Confusion matrix obtained by (a) AlexNet, (b) VGG16, (c) GoogleNet, (d) ResNet101, (e) DenseNet121, and (f) DenseNet169 on Million-AID dataset. Zoom for detail.

in which each scene category share the same number of images. This implies that significant emphasis should be addressed to the property of category imbalance when developing algorithms for aerial scene classification.

Confusion matrices: By further investigating the confusion matrices (as shown in Figure 6) of different CNN models, we can see that the deep CNN models, e.g., ResNet101 and DenseNet121/169, achieve much clearer confusion matrices than those of the shallow ones, e.g., AlexNet. It indicates that the deep CNN models have better ability to distinguish different scene categories, which is consistent with the result from Table 2. Several scene categories achieve classification accuracy approximate or equal to 1 as most of them show simple color, texture, and structure features in the scene images. Specifically, scene images like *desert* and *ice land* are mainly characterized with yellow and white components, respectively. The *terrace field* scene usually consists of distinct curve texture. In most cases, natural scenes like *river* and *sparse shrub land* show single structure and monotonous content in the aerial images. Thus, these kinds of scenes can be easily distinguished from others benefiting from their highly recognizable features of image content.

Nevertheless, the majority of scene categories obtain the classification accuracy below 0.9 and quite a few categories obtain the classification accuracy below 0.5. Particularly, the *dry field* and *paddy field*, *detached house* and *mobile home park* are heavily confused as they fall into similar land cover types, respectively. Many *stadium* images are misclassified as *ground track field* because of their high similarity of scene content. Most of the *beach* scenes are wrongly classified as *river* and *quarry* owing to their commonalities in structure and texture attributes. The same situation can also be observed between *dam* and *river* scenes. Notably, some scenes are easily misclassified as many different categories, such as *train station*, *parking lot*, *church*, and *works*. This phenomenon is mainly caused by the high intra-class variation of scene images that the algorithms cannot accurately distinguish them from each other. From this result we can seen that the Million-AID is a challenging dataset characterized with strong image variation of high inter-class similarity and intra-class diversity. Therefore, effective algorithms are desired to deal with these challenges, thereby, extracting excellent representations toward distinguishing different aerial scene categories.

Table 3
OA Comparison Among Different Datasets (%)

Dataset	AlexNet	VGG16	GoogleNet
AID	86.86	86.59	83.44
AID*	88.79	93.72	92.24
NWPU-RESISC45	85.16	90.36	86.02
NWPU-RESISC45*	87.19	92.76	91.71
Million-AID	67.53	77.47	77.37

AID* indicates the average OAs of ten repeated experiments using our implemented CNN framework, so does the ***NWPU-RESISC45****. The standard deviations are omitted since their negligible influence on the final result.

Comparison with existing benchmarks: Many datasets have been established to promote the advancement of scene classification as detailed in (Long et al., 2021). We compare the classification results of Million-AID with those of popular aerial scene classification datasets, *i.e.*, AID (Xia et al., 2017) and NWPU-RESISC45 (Cheng et al., 2017),

Table 4
Performance of Multi-label Scene Classification with Different CNN Models (%)

Model	$\tau = 0.5$						$\tau = 0.75$						mAP
	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1	
AlexNet	71.45	48.19	57.56	76.19	62.84	68.87	78.89	38.51	51.76	85.65	53.03	65.50	61.76
VGG16	82.26	62.20	70.84	86.98	75.31	80.72	84.61	54.29	66.14	91.70	69.37	78.99	79.13
GoogleNet	51.79	33.99	41.04	88.50	59.47	71.14	50.99	23.76	32.42	94.90	47.02	62.89	60.03
ResNet101	79.38	59.67	68.13	88.74	77.31	82.63	76.83	51.56	61.71	93.05	70.93	80.50	80.42
DenseNet121	79.09	56.21	65.71	89.74	75.10	81.77	76.36	47.75	58.76	94.20	67.72	78.79	78.94
DenseNet169	78.54	61.92	69.24	88.50	78.55	83.23	78.52	55.10	64.76	92.66	73.10	81.72	80.99

considering their high quality and wide application. Table 3 describes the overall accuracy of different CNN models. The results show that our implemented CNN models (indicated with *) achieve better performance than that reported in the original publications, which confirms the rationality and superiority of our implemented framework and learning schemes. Thus, we are able to acquire reliable experimental results based the our established CNN library in this work. Obviously, Million-AID achieves significantly lower accuracy than that of AID and NWPU-RESISC45 on all reported CNN models. This indicates Million-AID is a more challenging dataset than the compared ones. Note that the number of testing image in Million-AID is dozens of times larger than that of other datasets. It means that a small decline of OA indicates the large amount of incorrectly classified scene images. Thus, Million-AID has the potential to serve as a reliable benchmark dataset for comprehensively evaluating and comparing the performance of different scene interpretation algorithms.

4.3. Multi-label scene classification

4.3.1. Implementation detail

We employ the aforementioned CNNs to evaluate the performance of multi-label scene classification on Million-AID. The predicted labels via the last fully connected layer are activated by *sigmoid* function and generate confidences for each of the semantic categories similar to (Qi et al., 2020). The binary cross-entropy is employed to measure the distance between the prediction and the true label (which is either 0 or 1). All CNN models are initialized with parameters pre-trained on ImageNet. The training and testing subsets are the same with those for multi-class scene classification except for the labels that are extended according to the category organization system as shown in Figure 3. For the adaption of classification models, we transform the hierarchical multi-label scene classification problem into traditional multi-class classification problem, where each of the nodes in the category system is regarded as a single label. In this configuration, the existing classification algorithms can be extended effortlessly for multi-label scene classification.

4.3.2. Evaluation protocols

The precision and recall are employed as evaluation metrics. For each image, the predicted scene labels are considered as positive if the confidences are greater than a threshold τ. The precision is defined as the fraction of correctly annotated labels with respect to generated labels. The recall is defined as the fraction of correctly annotated labels with respect to ground-truth labels. Following conventional settings (Wang et al., 2016; Chen et al., 2019b; Lin et al., 2021), we calculate the per-class precision (CP), recall (CR), F1 (CF1) and overall precision (OP), recall (OR), F1 (OF1) for performance evaluation, where the average is calculated over all classes and all testing scene images, respectively. For fair comparison, we also compute the mean average precision (mAP), which is the mean value of average precision per class. Generally, the CF1 and, OF1, and mAP are relatively more important evaluation metrics to reflect the comprehensive performance.

4.3.3. Experimental Results

Quantitative results of multi-label scene classification on Million-AID are reported in Table 4. Obviously, the VGG16 model contains the most parameters while DenseNet169 the most convolutional layers among the employed networks. As can be seen, VGG16 and DenseNet169 achieve comparable performance and obviously outperform the other networks. For per-class metrics, VGG16 obtains the CP of 82.26% and CR of 62.20%, achieving the best performance on CF1 of 70.84% when $\tau = 0.5$. This result is slightly better than that from DenseNet169. Nevertheless, DenseNet169 achieves better performance on overall metrics, where the OP, OR, and OF1 are 88.50%, 78.55%, 83.23%, respectively. Figure 7(a) and (b) presents the average precision of each category when using VGG16 and DenseNet169, respectively. As can be seen, the two methods achieve similar classification performance for most categories. Superficially, the ResNet101 that shows superiority in both parameter scale and number of layers achieves the classification performance (mAP of 80.42%) close to that of DenseNet169 (mAP of 80.99%). However, when the scene images are assigned with the hierarchically multiple labels, the issue of data imbalance becomes prominent, which brings the problem of "catastrophic forgetting" (Goodfellow et al., 2013; Pfülb et al., 2018). As a result, the networks show weak performance on categories like *stadium* and *apartment* When increasing τ to be 0.75, the OP of all models gain significant improvement. This makes sense because the greater threshold value means the scene labels are predicted and filtered with higher confidences. However, all the recall metrics decline sharply, including the CR and OR metrics of different methods. As a result, the performance on CF1, OF1, and

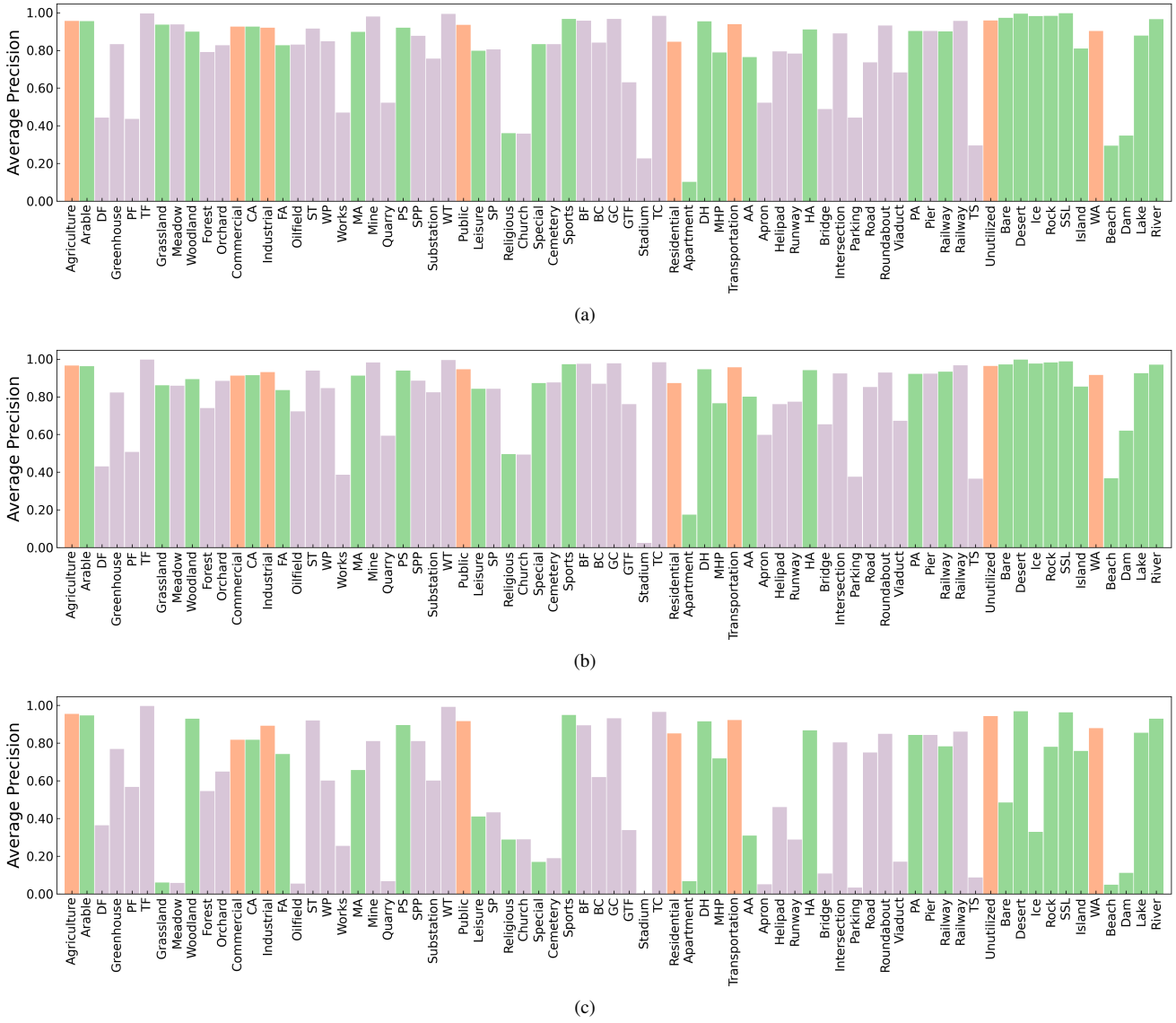


Figure 7: Average precision results of (a) VGG16, (b) DenseNet169, and (c) GoogleNet on Million-AID dataset. The orange bars indicates predicted scene labels belonging to the first-level categories, the green bars the second-level categories, and the plum bars the third-level categories.

mAP metrics decline correspondingly in comparison with the result of which $\tau = 0.5$. Nevertheless, DenseNet169 achieves the best performance on OF1 (81.72%) and mAP (80.99%), indicating its excellent ability in distinguishing different semantic scene categories.

An observation of interest is that, there are shallow CNN models that significantly outperform the deep ones. A case in point is that for $\tau = 0.5$ AlexNet achieves CF1 of 57.56% and mAP of 61.76%, which are 16.25% and 1.73% higher than those of GoogleNet, respectively. Even AlexNet achieves 2.27% lower OF1 than that of GoogleNet, the former model shows superiority on OR. Besides, VGG16 reports comprehensively better performance of mAP when compared with the deeper networks such as GoogleNet and DenseNet121. What is noteworthy is that the shallow networks, *i.e.*, AlexNet and VGG16, contain particularly large-scale parameters compared with the others as detailed in Table 1. With this superiority, the shallow networks are able to learn the relationship of scene categories at different levels of hierarchy. As a comparison, GoogleNet possesses more convolutional layers than those of AlexNet and VGG but with the least model parameters (6.8M). The experimental results show that GoogleNet provides the worst comprehensive performance (mAP of 60.03%) among the employed CNN models. Taking the result of $\tau = 0.5$ as an example, GoogleNet achieves CF1 of 41.04% and mAP of 32.96%, which are significantly poorer than those from other models. Simultaneously, the catastrophic forgetting problem become particularly prominent for GoogleNet. Figure 7(c) presents the average precision of each category when using GoogleNet. As can be seen, many of categories at the second and third semantic levels can not be recognized, resulting in poor CF1

and mAP. Therefore, GoogleNet show relatively weak ability in learning the hierarchical relationship between different semantic scenes.

Significantly, the biggest difference between VGG16 and AlexNet is that the former network possesses more convolutional layers and thus contains more than twice as many parameters as the former one. Hence, VGG16 gains remarkable improvement of classification performance. Although DenseNet121 consists of parameters at a scale comparable with that of GoogleNet, it possesses much more convolutional layers which help to significantly improve the performance of multi-label scene classification. As the depth of convolutional layers going deeper, the performance improvement is also obvious, *i.e.*, the results from DenseNet121 and DenseNet169 as shown in Table 4. With the above analysis, it is natural to argue that both the parameter scale and depth of a CNN are crucial for recognizing the scene categories with multiple semantics. Intuitively, more parameters and deeper convolutional layers can enhance the network's ability to learn the heterogeneous characteristics of different scene categories, but also the ability to learn the homogeneous characteristics of scenes belonging to the same parent categories. This to some extent helps to reveal the hierarchical relationship between different semantic categories, which greatly improves the performance of hierarchical multi-label scene classification. Nevertheless, how to model the the hierarchical rather than parallel relationships between different scene categories and further improve the performance of hierarchical multi-label scene classification remain to be further explored.

5. Transferring Knowledge From Million-AID

Million-AID consists of large-scale aerial images that characterize diverse scenes. This provides Million-AID with rich semantic knowledge of scene content. Hence, it is natural for us to explore the potential to transfer the semantic knowledge in Million-AID to other domains. To this end, we consider two basic strategies, *i.e.*, fine-tuning pre-trained networks for tile-level scene classification and hierarchical multi-task learning for pixel-level semantic parsing.

5.1. Fine-tuning pre-trained networks for scene classification

5.1.1. Implementation detail

A network trained from scratch is usually hard to capture the essential features of aerial scene content. Fine-tuning a pre-trained CNN model has proven to be useful for aerial image interpretation (Liu et al., 2018; Zhao et al., 2017a; Cheng et al., 2017; Xia et al., 2018a; Tong et al., 2020), of which performance is improved by leveraging content knowledge from other domains. Particularly, CNN models are usually pre-trained on natural image archives, *e.g.*, ImageNet (Deng et al., 2009), and then fine-tuned on the target dataset for aerial image scene classification. The fine-tuning strategy has been regarded as an common solution to relieve the data scarcity problem for scene classification model adaption. Likewise, we employ the fine-tune learning strategy to verify the generalization ability of Million-AID dataset.

To verify the superiority of Million-AID, we first train CNN models for scene classification using all images in Million-AID. The CNN models pre-trained on Million-AID are then fine-tuned with images in the target scene classification datasets, *i.e.*, AID (Xia et al., 2017) and NWPU-RESISC45 (Cheng et al., 2017). Similar to the dataset partition scheme in (Xia et al., 2017; Cheng et al., 2017), 20% images are randomly selected as training set and the rest 80% as test set. We repeat this operation ten times to reduce the influence of the randomness and obtain reliable classification results. The epidemic CNN networks presented in Section 4 are employed to comprehensively evaluate the superiority of Million-AID. The learning rate are set to be 0.01 in the pretrain phase. In order to effectively utilize the scene knowledge learned from initial datasets, the learning rates in the fine-tune phase are set to be 0.001 for all models. Through this step-wise optimization scheme, we are able to transfer the learned scene knowledge of Million-AID better to adapt to the target datasets. The other training parameters are set the same as those for multi-class scene classification in Section 4. As a comparison, all the employed CNN networks are fine-tuned with the models pre-trained on ImageNet. We also report the scene classification results from models trained from scratch, where the learning rates are also set to be 0.001 for consistency. The evaluation protocols are the same as those for multi-class scene classification as introduced in Section 4.

5.1.2. Experimental results

Tables 5 and 6 illustrate the means and standard deviations of OA, AA, and Kappa on AID and NWPU-RESISC45, respectively. For each model, the best performance among different learning schemes (*i.e.*, the models trained from scratch, fine-tuned on ImageNet, and fine-tuned on Million-AID) is reported in bold. By analyzing the tables, one can see that learning directly from scratch achieves the worst result. It indicates that optimizing CNN models for aerial scene classification can be difficult owing to the scarcity of training data and complexity of scene content. Thus, researchers often resort to extract aerial scene features by models adapted well on natural images and then recognize aerial scenes by utilizing feature classifiers (*e.g.*, SVM (Xia et al., 2017; Cheng et al., 2017; Nogueira et al., 2017)). Compared with the models trained from scratch, the models pre-trained on ImageNet and Million-AID can significantly improve the classification performance. Figures 8 and 9 provide the confusion matrices of the best results obtained by different learning schemes on AID and NWPU-RESISC45, respectively. It is shown that the classification performance of each scene category is significantly improved by the fine-tuned models. This confirms the importance of parameter initialization for CNN model adaption. In particular, it strongly demonstrates the effectiveness and positive significance of Million-AID for training CNN models toward aerial image scene classification.

Table 5
Classification Accuracy (%) on AID Dataset Using Different Training Schemes

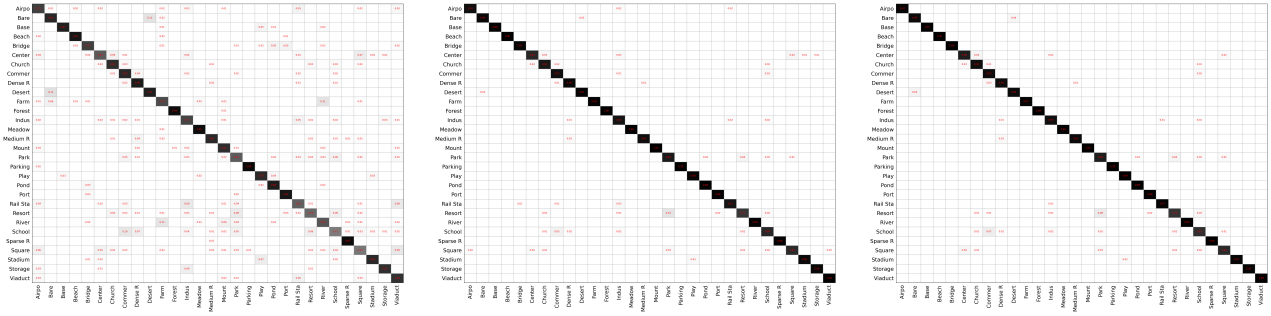
Metric	Pretrain dataset	AlexNet	VGG16	GoogleNet	ResNet101	DenseNet121	DenseNet169
OA	W/O	33.47 ± 2.15	72.18 ± 0.49	79.05 ± 0.89	49.46 ± 2.07	58.02 ± 0.74	59.16 ± 0.52
	ImageNet	88.79 ± 0.40	93.72 ± 0.21	92.24 ± 0.21	94.52 ± 0.25	94.68 ± 0.19	94.76 ± 0.21
	Million-AID	**90.70 ± 0.43**	**95.33 ± 0.28**	**94.55 ± 0.23**	**95.40 ± 0.19**	**95.22 ± 0.26**	**95.24 ± 0.35**
AA	W/O	33.85 ± 2.35	72.16 ± 0.54	78.88 ± 0.88	49.29 ± 2.06	57.88 ± 0.73	59.04 ± 0.51
	ImageNet	88.52 ± 0.39	93.38 ± 0.22	91.78 ± 0.23	94.18 ± 0.29	94.39 ± 0.21	94.44 ± 0.22
	Million-AID	**90.46 ± 0.45**	**95.14 ± 0.27**	**94.30 ± 0.23**	**95.17 ± 0.19**	**94.97 ± 0.26**	**95.00 ± 0.38**
Kappa	W/O	31.09 ± 2.24	71.19 ± 0.51	78.31 ± 0.92	47.63 ± 2.15	56.50 ± 0.76	57.69 ± 0.53
	ImageNet	88.39 ± 0.42	93.49 ± 0.21	91.96 ± 0.22	94.32 ± 0.26	94.49 ± 0.20	94.57 ± 0.22
	Million-AID	**90.37 ± 0.44**	**95.17 ± 0.29**	**94.35 ± 0.24**	**95.24 ± 0.20**	**95.05 ± 0.27**	**95.07 ± 0.37**

*** *W/O* indicates the classification models are trained from scratch.**

Table 6
Classification Accuracy (%) on NPWU-RESISC45 Dataset Using Different Training Schemes

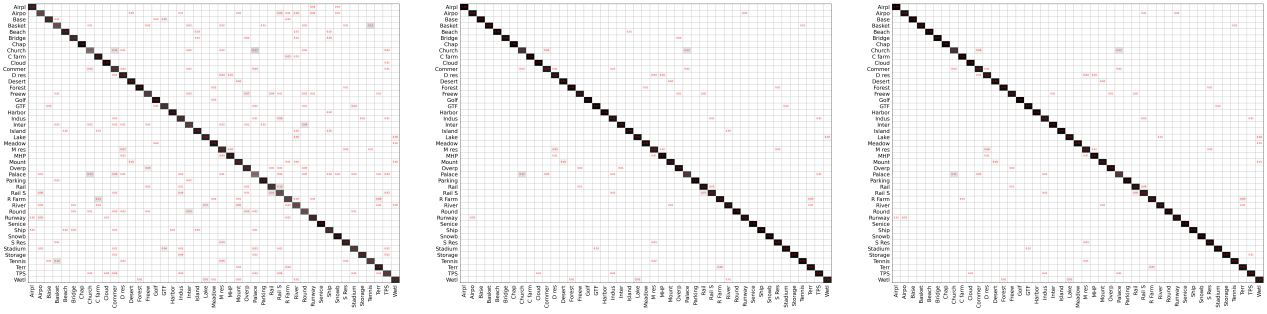
Metric	Pretrain dataset	AlexNet	VGG16	GoogleNet	ResNet101	DenseNet121	DenseNet169
OA	W/O	37.92 ± 0.70	73.19 ± 0.44	81.77 ± 0.56	58.82 ± 0.74	63.35 ± 0.34	64.51 ± 0.47
	ImageNet	87.19 ± 0.26	92.76 ± 0.18	91.71 ± 0.25	94.06 ± 0.16	93.90 ± 0.19	94.11 ± 0.20
	Million-AID	**88.24 ± 0.21**	**93.62 ± 0.20**	**93.40 ± 0.23**	**94.20 ± 0.16**	**94.21 ± 0.20**	**94.26 ± 0.21**
AA	W/O	37.92 ± 0.70	73.19 ± 0.44	81.77 ± 0.56	58.82 ± 0.74	63.35 ± 0.34	64.51 ± 0.47
	ImageNet	87.19 ± 0.26	92.76 ± 0.18	91.71 ± 0.25	94.06 ± 0.16	93.90 ± 0.19	94.11 ± 0.20
	Million-AID	**88.24 ± 0.21**	**93.62 ± 0.20**	**93.40 ± 0.23**	**94.20 ± 0.16**	**94.21 ± 0.20**	**94.26 ± 0.21**
Kappa	W/O	36.51 ± 0.72	72.59 ± 0.45	81.36 ± 0.58	57.89 ± 0.75	62.51 ± 0.35	63.70 ± 0.48
	ImageNet	86.89 ± 0.21	92.60 ± 0.19	91.52 ± 0.26	93.92 ± 0.17	93.76 ± 0.19	93.98 ± 0.20
	Million-AID	**87.97 ± 0.21**	**93.48 ± 0.20**	**93.25 ± 0.24**	**94.07 ± 0.16**	**94.08 ± 0.20**	**94.13 ± 0.21**

*** *W/O* indicates the classification models are trained from scratch.**



(a) (b) (c)

Figure 8: Confusion matrix obtained by (a) GoogleNet trained from scratch, (b) DenseNet169 pre-trained on ImageNet, and (c) ResNet101 pre-trained on Million-AID. Results are based on AID dataset. Zoom for detail.



(a) (b) (c)

Figure 9: Confusion matrix obtained by (a) GoogleNet trained from scratch, (b) DenseNet169 pre-trained on ImageNet, and (c) DenseNet169 pre-trained on Million-AID. Results are based on NWPU-RESISC45 dataset. Zoom for detail.

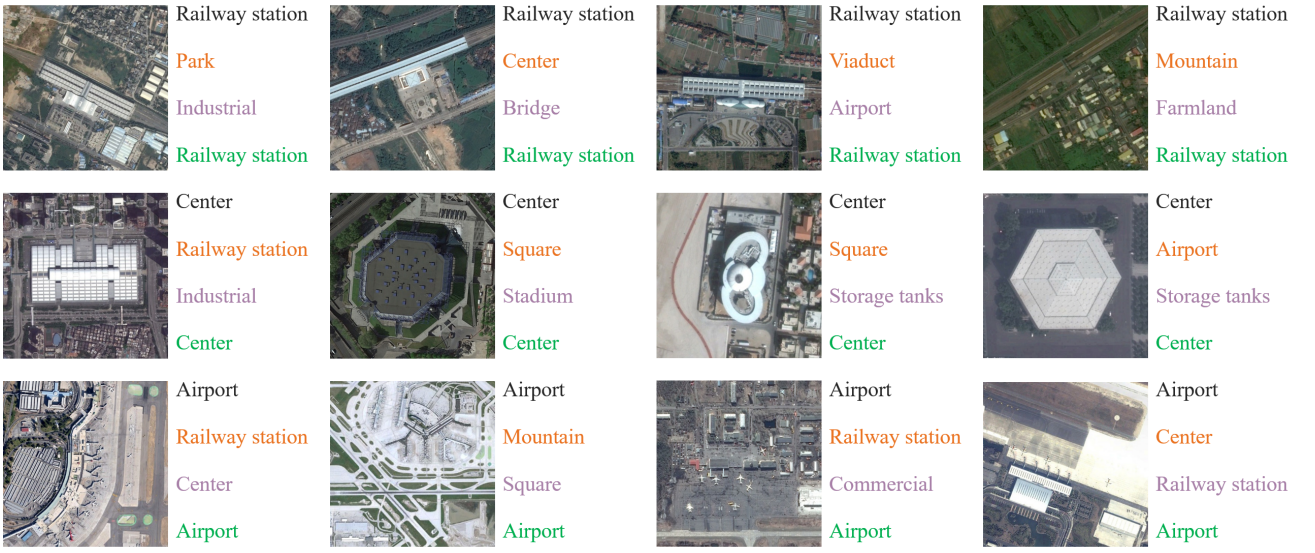


Figure 10: Example images and predictions on AID. For each training scheme, the best of the trained models is selected for classification result comparison. The black labels are the ground truth. The orange labels indicate predictions by GoogleNet trained from scratch, the plum labels the predictions by DenseNet169 pre-trained on ImageNet, and the green labels the predictions by ResNet101 pre-trained on Million-AID.

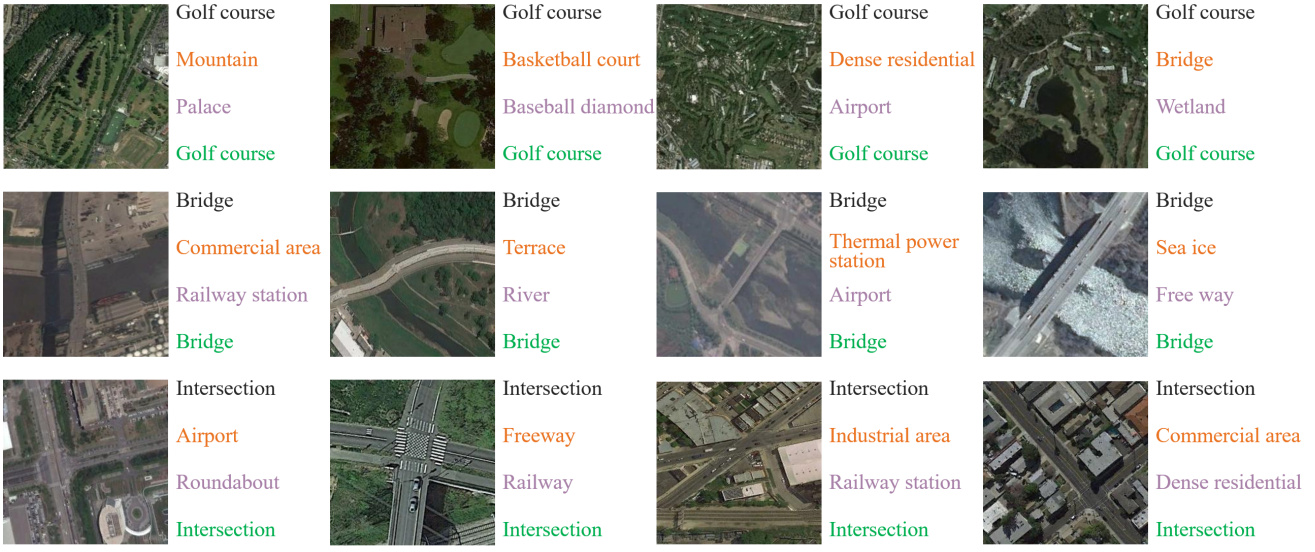


Figure 11: Example images and predictions on NWPU-RESISC45. For each training scheme, the best of the trained models is selected for classification result comparison. The black labels are the ground truth. The orange labels indicate predictions by GoogleNet trained from scratch, the plum labels the predictions by DenseNet169 pre-trained on ImageNet, and the green labels the predictions by DenseNet169 pre-trained on Million-AID.

An important observation is that all models pre-trained on Million-AID achieve obviously better performance compared with those pre-trained on ImageNet. Specifically, for both AID and NWPU-RESISC45, each considering CNN model pre-trained on Million-AID provides the maximum accuracy. As shown in Figure 8 (b) and (c), by employing Million-AID for model pre-training, the classification accuracy of *railway station*, *center*, and *airport* in AID reach 97%, 88%, and 97%, which are 6%, 4%, and 4% higher than using ImageNet, respectively. Likewise, impressive accuracy improvement can also be observed for scene categories in NWPU-RESISC45, such as *golf course*, *bridge*, and *intersection* as shown in Figure 9 (b) and (c). Figures 10 and 11 provide the corresponding example images and predictions on AID and NWPU-RESISC45, respectively. It is shown that the models pre-trained on Million-AID can better distinguish between semantic scenes with similar characteristics. Intuitively, due to the difference in spatial pattern, texture structure, and visual appearance, there is a gigantic semantic gap between the natural and aerial image content. Hence, the CNN models pre-trained with natural images may not be generally applicable to reduce this semantic gap for aerial image interpretation. By contrast, the models trained with pure large-scale aerial images can naturally grasp the unique

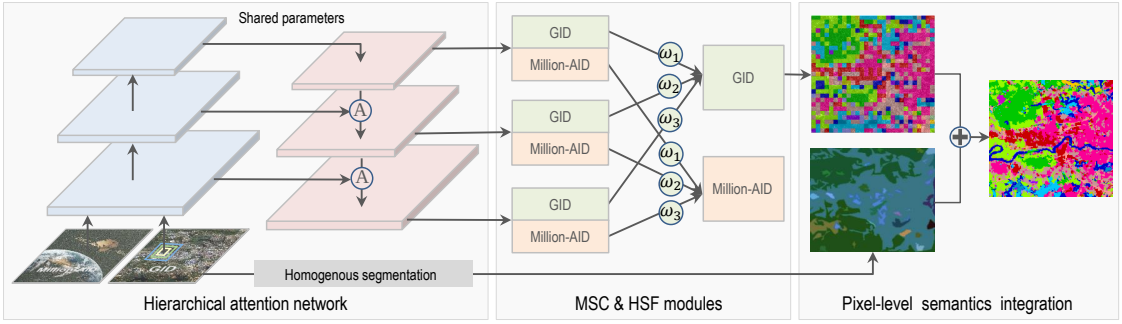


Figure 12: The framework of hierarchical multi-task learning for pixel-level semantic labeling. Each pair of framed *GID* and *Million-AID* denotes the multi-task classification branches for GID and Million-AID, respectively. The sign of Ⓐ indicates the attention module and the ⊕ the majority voting process.

characteristics and knowledge of image content. In this context, the subsequent CNN models fine-tuned with aerial images in the target datasets are able to learn better features for aerial scene content representation, and thus, outperform those using the natural images.

From the shallow networks (*e.g.*, AlexNet, VGG16, and GoogleNet) to deeper ones (*e.g.*, ResNet101, DenseNet121, and DenseNet169), the performance difference between ImageNet and Million-AID pre-trained models become smaller. As an example on NWPU-RESISC45, AlexNet, VGG16, and GoogleNet pre-trained on Million-AID achieves 1.05%, 0.86%, and 1.69% higher OAs than the results from ImageNet pre-trained models, respectively. This performance difference decreases to 0.14%, 0.31%, and 0.15% when it comes to ResNet101, DenseNet121, and DenseNet169, respectively. The results on AID also show similar phenomenon. This makes sense because the deeper the network, the more likely the learned features adapted to the target aerial images, resulting in comparable performance among different learning strategies. Nevertheless, the models pre-trained with Million-AID still show superiority, which confirms the strong generalization ability of Million-AID. To our knowledge, it is the first time to be observed that CNN models pre-trained with pure large-scale aerial images are verified to surpass those using natural images. Prior to this, many CNN models are usually pre-trained using the ImageNet dataset and then fine-tuned on the target dataset for aerial image scene classification owing to the lack of available large-scale aerial image archives. With the above observation and results, it is natural to argue that the proposed Million-AID can make a significant advancement for the use of CNN models in aerial image scene classification, opening up a promising direction to support parameter initialization of CNN models toward various aerial image interpretation tasks.

5.2. Hierarchical multi-task learning for semantic labeling

5.2.1. Framework

The conventional CNN learns scene features via stacked convolutional layers and the output of the last fully connected layer is usually employed for scene representation. However, learning stable features from single layer can be a difficult task because of the complexity of scene content. Moreover, data sparsity which is a long-standing notorious problem can easily lead to model overfitting and weak generalization ability because of the insufficient knowledge captured from limited training data. To relieve the above issues, we introduce a hierarchical multi-task learning method and further explore how well the knowledge contained in Million-AID can be transferred to boost the pixel-level semantic parsing of aerial images. To this end, the GID (Tong et al., 2020), which consists of training set with tile-level scenes and large-size test images with pixel-wise annotations, has provided us an opportunity to bridge the tile-level scene classification toward pixel-level semantic labeling. Generally, the presented framework consists four components, *e.g.*, hierarchical scene representation, multi-task scene classification, hierarchical semantic fusion, and pixel-level semantics integration, as shown in Figure 12.

Hierarchical attention network (HAN): The high-resolution features from shallow convolutional layers can learn valuable visual information of small objects, texture structures, and spatial patterns associated closely with specific scene content while the semantic clues is insufficient. To compensate for this defect, the hierarchical attention features are learned via transmitting the semantic information from the deep layers to the shallower ones inspired by (Lin et al., 2017; Zhang and Kim, 2019). Specifically, the deep-layer feature (**DF**) is firstly upsampled to the same spatial size as the shallow-layer feature (**SF**) to maintain the semantic information as much as possible. The upsampled semantic feature is processed by a 1×1 convolutional layer for dimension reduction. And the sigmoid function is then employed to generate the semantic attention map (**SAM**), which possesses the same channel number with the **SF**. Element-wise multiplication is conducted between the **SF** and **SAM** to generate local attention feature (**LAF**). Finally, the **SF** and **LAF** is assembled by element-wise summation and output the final attention feature (**AF**). The whole process is illustrated in Figure 13. By repeatably transmitting the deep-layer features to the shallower convolutional layers, we are able to construct the hierarchical attention network (as shown in Figure 12) that incorporate semantic and visual information for scene representation.

Multi-task scene classification (MSC): With the hierar-

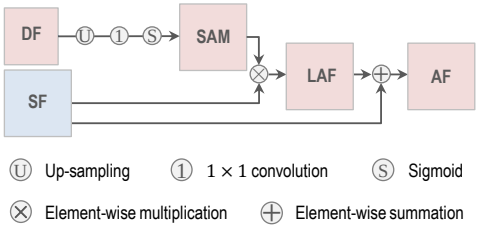


Figure 13: Attention module in hierarchical attention network.

chical attention features, a semantic scene can be represented using the features of multiple scales. As shown in Figure 12, three streams are constructed for multi-scale feature extraction. For each stream, the hierarchical attention feature is further processed by global average pooling and generate the feature for scene representation. Then, the multi-task classification branches are designed to recognize scenes from different datasets. By learning the sharing parameters for different tasks, multi-task learning enables the knowledge learned in one task to be utilized in the others, and thus, improve the generalization ability of a trained model (Pan and Yang, 2010; Zhang and Yang, 2021)l. The presented Million-AID is composed of rich semantic scenes and massive instances that characterize the land cover features. It is reasonable to transfer the land cover knowledge contained in Million-AID to boost the semantic classification of aerial images. With this in mind, the multi-task learning is conducted on Million-AID and the challenging GID (Tong et al., 2020), which is established for pixel-level semantic classification of land cover. For simplicity, two branches at each of the hierarchical output layers are designated for the scene-level classification of images in Million-AID and GID, respectively. The summation of weighted losses is used for model adaption:

$$Loss^g = \sum_{s=1}^{S} w_s CE_s^g \tag{1}$$

$$Loss^m = \sum_{s=1}^{S} w_s CE_s^m \tag{2}$$

$$Loss = \mu_g Loss^g + \mu_m Loss^m \tag{3}$$

where CE_s^g represents the cross entropy loss of scene classification using the image features of scale $s \in 1, 2, ..., S$ for GID while CE_s^m for Million-AID. w_s indicates the loss weight for the classification at scale s. μ_g and μ_m (where $\mu_g + \mu_m = 1$) indicate the weighted importance of different tasks, *i.e.*, scene classification on GID and Million-AID, respectively. In this work, we aim at improving the classification performance on GID by knowledge transfer from Million-AID. Hence, the semantic classification on GID is regarded as the main task while the scene classification on Million-AID serves as the auxiliary task (Ruder, 2017) to still reap the benefits of multi-task learning strategy.

Hierarchical semantic fusion (HSF): To give full play of the advantages of hierarchical attention features, the classification results with different feature scales are integrated. Using the feature at scale s, the classification probability vector $p_s(I)$ of image I is obtained by a *softmax* layer:

$$p_s(I) = \{p_{s,1}(I),\ p_{s,2}(I),\ ...,\ p_{s,N}(I)\},\ \ p_s(I) \in \mathbb{R}^N \tag{4}$$

where $p_{s,n}$ represents the probability that I belongs to class n using the feature of scale s. Essentially, the predictions at different scales reflect the probability that a classified scene belongs to individual categories from the perspective of different feature levels. Hence, it is reasonable to integrate the predictions of different scales. To this end, a summation of weighted probabilities is performed as the final prediction:

$$\widehat{p}_n(I) = \frac{\sum_{s=1}^{S} w_s p_{s,n}(I)}{\sum_{s=1}^{S} w_s} \tag{5}$$

where $\widehat{p}_n(I) \in [0, 1]$ indicates the probability that image I belongs to class n. w_s represents the weight for scale s, which serves as the loss weight for the corresponding classification stream. The integration of weighted probabilities aims to provide a more stable prediction result. Then the predicted scene category of image I is expressed as:

$$l(I) = \underset{n\in[1,2,...,n]}{\arg\max}\ \widehat{p}_n(I) \tag{6}$$

where $l(I)$ is the category label of image I.

Pixel-level semantics integration (PSI): With above procedures, we are able to obtain the semantic grid map by tile-level classification for interpreting a large-size aerial image. Here, each semantic grid corresponds to a tile-level classification result. For more accurate result with pixel-level semantics, the boundary information of ground objects in high-resolution aerial images appears the great importance. We therefore employ object-based segmentation and majority voting strategy to generate pixel-level semantic labeling result. To this end, the selective search algorithm (Van de Sande et al., 2011) is conducted on the raw aerial image and produce homogeneous segmentation map. Let r be a homogeneous region in the segmentation map. The majority voting algorithm is then performed to determine the semantic class of r denoted as $l(r)$ by referencing the semantic grid map:

$$l(r) = \underset{n\in[1,2,...,n]}{\arg\max}\ |r_n| \tag{7}$$

where $|r_n|$ denotes the number of pixels enclosed in r and labeled as class n in the semantic grid map. To enhance the discrimination ability of scene content, we represent the scene with multi-scale context, of which semantic meaning is identified by the central point as illustrated in Figure 1(d). By integrating multi-scale contextual information with tile-level classification, the whole pipeline falls into a hybrid classification framework (Tong et al., 2020). Nevertheless, our method focuses more on advancing the pixel-level semantic labeling via improving the tile-level semantic interpretation performance and transferring aerial scene knowledge from Million-AID.

Table 7
Weights Influence on Different Tasks

μ_g	μ_m	GID			Million-AID		
		Kappa (%)	OA (%)	mIoU (%)	Kappa (%)	OA (%)	AA (%)
0.1	0.9	62.85	69.06	39.88	**90.36**	**90.62**	**89.55**
0.3	0.7	65.15	71.00	41.85	89.44	89.72	88.91
0.5	0.5	**66.65**	**72.38**	**42.71**	**89.67**	**89.94**	**89.14**
0.7	0.3	66.14	72.02	41.75	88.98	89.27	87.84

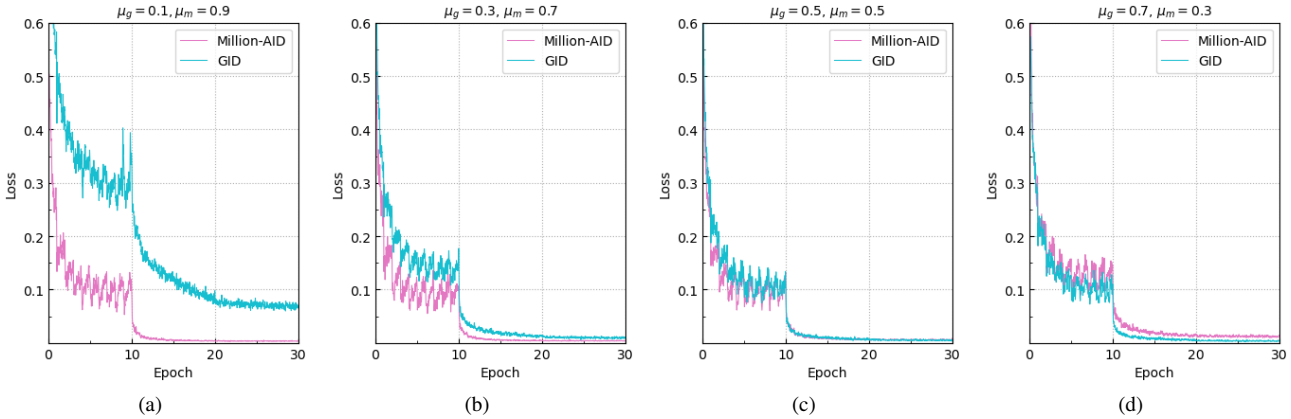


Figure 14: Training loss with respect to different setups of learning weights. The models converge successfully for both classification tasks with different learning weights, which confirms the validity of the learned classification model.

5.2.2. Implementation detail

ResNet50 (He et al., 2016) is employed as the backbone, where the last three residual blocks of conv3_x, cov4_x, and conv5_x are utilized to extract the hierarchical attention features in three streams. The loss weights for the three streams are empirical set to be $w_1 = 0.25$, $w_2 = 0.5$, and $w_3 = 1.0$ according to the ratios of channel numbers of the hierarchical layers, respectively. The *fine classification set* of GID, which contains 15 challenging semantic categories, is employed for performance evaluation. Specifically, the proposed model are trained with the subset of 30k tile-level scene patches and then tested on the subset of 10 Gaofen-2 images (6800×7200) with pixel-level semantic labels. The multi-scale contextual information are extracted with the windows of 56×56, 112×112, and 224×224, which are consistent with the sizes of training samples. As training samples in GID are highly overlapped, image flipping and rotation (90°, 180°, and 270°) are conducted for data augmentation, resulting in 120k tile-level scene samples. Correspondingly, 120k scene images in Million-AID are randomly selected as training set, according to the ratios of instance number of each scene category. The training parameters are set the same as those for multi-class scene classification in Section 4 except the number of total iteration set as 30 epochs. The OA, Kappa, and mean-Intersection-over-Union (mIoU) (Minaee et al., 2021) are employed for performance evaluation.

5.2.3. Experimental results

Ablation study: For multi-task learning, the weights of different tasks can make a big influence on the classification performance. Table 7 shows the classification result under different weight setups, where the MSC strategy is embedded in the baseline network for tile-level scene classification on Million-AID (T_m) and pixel-level semantic parsing on GID (T_g). Figure 14 illustrates the corresponding changes of training losses with respect to different setups of learning weights. As can be seen, the performance of T_g increases gradually as $\mu_g : \mu_m$ changes from 0.1 : 0.9 to 0.5 : 0.5. This makes sense because the model tend to optimize T_m when μ_g is smaller than μ_m. As μ_g increases, the model pays incremental attentions to optimize the T_g and borrows knowledge learned from Million-AID concurrently. Particularly, when $\mu_g : \mu_m = 0.1 : 0.9$, the model can be well optimized for T_m, and thus, T_m achieves the best performance with this setup (Figure 14(a)). When $\mu_g : \mu_m = 0.3 : 0.7$, the optimization for T_m is slightly insufficient while T_g gains significant improvement as shown in Figure 14(b). When $\mu_g : \mu_m$ changes to 0.5 : 0.5, both T_g and T_M can be well optimized (Figure 14(c)) and T_m can also benefit from the knowledge learned from GID. Thus, the model gains obvious performance improvement. However, as $\mu_g : \mu_m$ continues to increase, the performance of both T_g and T_m begin to decline because there is a risk of overfitting for T_g and insufficient optimzaiton for T_m as shown in Figure 14(d). Generally, the best performance for T_g can be acquired when $\mu_g = \mu_m = 0.5$, which is employed in subsequent experiments. As our purpose is to explore the possibility of transferring knowledge in Million-AID to GID for semantic classification of land cover, we will focus on reporting the performance of the main task (*i.e.*, pixel-level semantic classification on GID) in the following context.

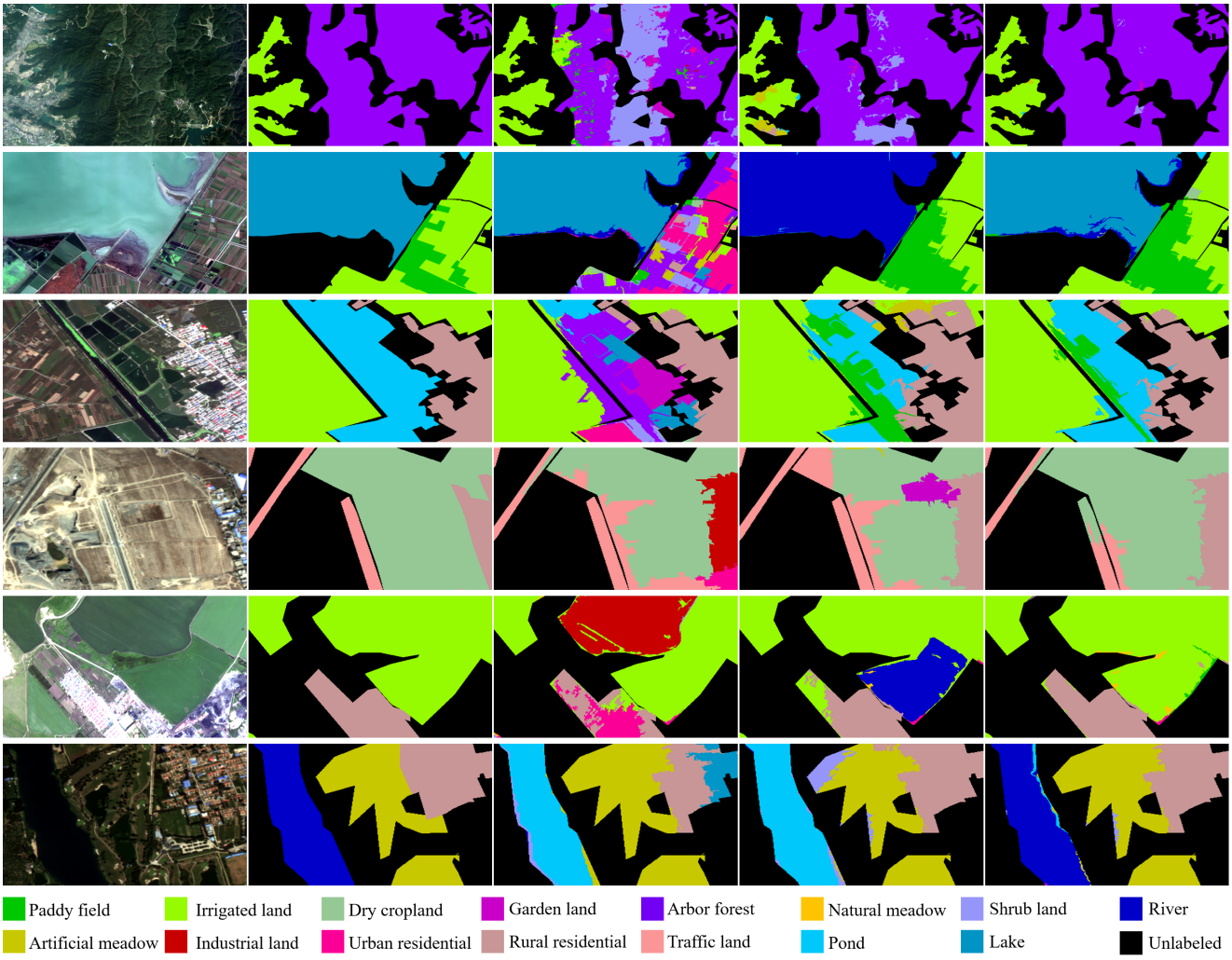


Figure 15: Qualitative comparisons among different classification schemes. Images in the first to fifth columns indicate the original images, ground truth annotations, classification maps from baseline, MSC, and the full implementation of our method, respectively.

For better understanding of our presented hierarchical multi-task learning method, detailed ablation studies are conducted with different settings. Specifically, the ResNet50 is employed as the baseline as introduced before. Then we gradually attach the multi-task scene classification (MSC) strategy, hierarchical scene representation (HAN) learning, and hierarchical semantic integration (HSF) scheme to the baseline model. Table 8 shows the performance comparison of different setups tested on GID. As can be seen, the results achieved by the baseline is far from satisfactory due to the sparsity of training samples. When employing the MSC strategy, the classification performance reaches 72.38% of OA, 42.71% of mIoU, and 66.65% of Kappa, which are 13.29%, 11.92%, and 15.06% higher than those of baseline, respectively. This strongly verifies the effectiveness of MSC, where diverse scene samples of Million-AID can bring implicit data augmentation and greatly boost the semantic feature learning for GID content representation. Under the circumstances, semantic knowledge contained in Million-AID can be effectively transferred to improve the performance of land cover classification on GID.

When the HAN is introduced, the classification is conducted with the hierarchical features from different streams, respectively. For a scene image, the highest classification score among different streams is adopted to output the corresponding semantic category. As shown in Table 8, the classification performance is further improved with HAN. This mainly benefits from hierarchical attention mechanism, where the essential features of a specific scene can be learned within individual layers. With the HSF scheme integrated, the advantage of attention features at different levels is significantly improved and the performance reaches 73.03% of OA, 43.68% of mIoU, and 67.33% of Kappa. As can be seen, the MSC strategy helps most in improving the classification performance while the full implementation of our method achieves the best result.

Figure 15 shows the qualitative comparisons of different classification schemes. It is shown that the similar categories are easy to be confused by the baseline method. By employing the MSC strategy, many confused categories can be distinguished, which verifies the effectiveness of the multi-task learning for transferring the scene knowledge contained

Table 8
Performance Comparison of Different Setups

Baseline	MSC	HAN	HSF	Kappa (%)	OA (%)	mIoU (%)
✓				51.59	59.09	30.79
✓	✓			66.65	72.38	42.71
✓	✓	✓		66.79	72.52	43.07
✓	✓	✓	✓	**67.33**	**73.03**	**43.68**

in Million-AID. With the full implementation of our developed method, the misclassification within some local areas is further corrected, which is consistent to the performance improvement in Table 8. In general, the designed modules greatly help to grasp the essential semantic knowledge of scene images in Million-AID and GID, thus, improve the generalization ability of the semantic classification model.

Performance comparison: The presented method is compared with several object-based classification methods provided by (Tong et al., 2020), where four typical features including spectral feature (SF), co-occurence matrix (GLCM), different morphological profiles (DMP), and local binary patterns (LBP) were fused to obtain the scene representation denoted as SGDL for simplicity. Then, maximum likelihood classification (MLC), random forest (RF), support vector machine (SVM), and multi-layer perception (MLP) are used as classifier for scene classification, respectively. Besides, we compare our method with the CNN model pretrained on the *large-scale classification set* of GID (PT-GID) (Tong et al., 2020). For comprehensive comparison, the presented method was also compared with the end-to-end semantic segmentation models, such as U-Net (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017b), DeepLab V3+ (Chen et al., 2018), and its variations like DeepLab V3+ Mixed loss function (DeopLab V3+ MLF), DeepLab v3+ MobileNet provided by (Ren et al., 2020).

The quantitative results of different methods are summarized in Table 9. As can be seen, our method significantly outperforms the object-based ones, showing the superiority of our presented method for semantic content understanding of aerial images. The best result achieved by image segmentation models is 59.8% of Kappa and 69.16% of OA from DeepLab V3+ Mixed Loss Functions (MLF). Note that the segmentation models and its variations are based on fully convolutional networks which learn pixel-level semantics in an end-to-end way. Nevertheless, our method achieves 7.5% higher Kappa and 3.87% higher OA than those achieved by DeepLab V3+ MLF, indicating the effectiveness of our method for pixel-level semantic parsing of aerial images. Particularly, PT-GID was achieved by transferring knowledge in the *large-scale classification set* of GID, which contains 150k samples relevant to the *fine classification set* of GID. Thus, PT-GID achieves remarkable performance on the fine land-cover classification set. In spite of this, our presented method achieves 6.8% higher Kappa and about 3% higher OA, showing the strong transferability and effectiveness of our presented method.

Figure 16 provides the intuitive visualization of the pixel-level classification results on the *fine classification set* of GID. To save space, we compare the best two results achieved by PT-GID and our designed method. As can be seen from the first row, the *irrigated land* is heavily misclassified as *dry cropland* by PT-GID because of the difficulty in distinguishing their similar visual features, such as the texture and structural information. By contrast, our method can discriminate the *irrigated land* more accurately even it is widely distributed. This benefits from our hierarchically fused features, which can simultaneously incorporate the visual features and semantic information toward specific scene content. In city areas shown in the second row, many semantic categories, such as the *traffic land*, *urban residential*, *industrial land*, and *dry cropland*, are heavily confused by PT-GID while our method obtains more accurate classification result. This contributes to the semantic attention feature learning in our method, which helps to grasp the essential information for discriminating content of different categories. Likewise, the extraction of *urban residential* areas is significantly improved by our method as shown in the third row. On the whole, the classification maps of our method present more homogenous areas and provide more smoother classification result than those of PT-GID. Thus, our method provide much better classification result than the others, which is consistent with the result in Table 9.

Table 9
Comparison of Classification Accuracy Among Different Methods on GID

Methods	Kappa	OA (%)
MLC + SGDL	0.145	23.61
SVM + SGDL	0.148	23.92
MLP + SGDL	0.199	30.57
RF + SGDL	0.237	33.70
DeepLab V3+ Mobilenet	0.357	54.64
U-Net	0.439	56.59
PSPNet	0.458	60.73
DeepLab V3+	0.478	62.19
DeepLab V3+ MLF	0.598	69.16
PT-GID	0.605	70.04
Ours	**0.673**	**73.03**

6. Conclusions

In this paper we address aerial scene parsing from tile-level scene classification to pixel-level semantic Labeling. Specifically, a review of aerial image interpretation was firstly conducted from its development perspective. It is shown that the interpretation prototype of aerial images has been progressing with the improvement of image resolution and experienced the stages from pixel-wise image classification, segmentation-based image analysis, and tile-level image understanding. Then, we detailed the large-scale dataset, i.e., Million-AID, to be released for aerial scene recognition. Intensive experiments with popular CNN frameworks indicate that Million-AID is a challenging dataset which can be employed as a benchmark for multi-class and multi-label aerial scene classification. Fine-tuning CNN models pre-trained on Million-AID show considerable superiority

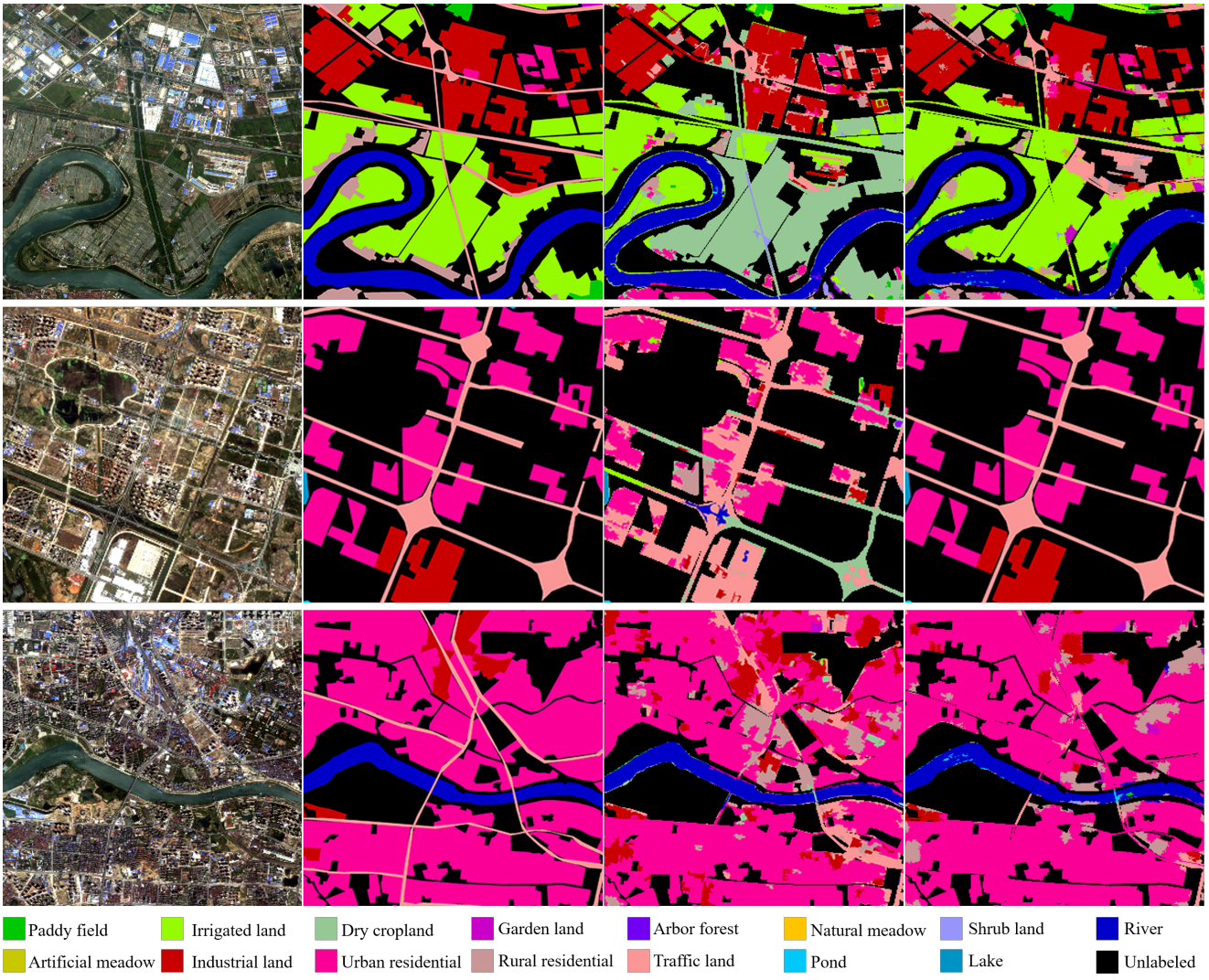


Figure 16: Visualization of the land cover classification results on the *fine classification set* of GID. Images in the first to fourth columns indicate the original image, ground truth annotations, classification maps of PT-GID (Tong et al., 2020), and classification maps of our method, respectively.

than those on ImageNet for tile-level aerial scene classification, which demonstrates the strong generalization ability of Million-AID. Besides, we designed a hierarchical multi-task learning framework and achieved the state-of-the-art result for pixel-level semantic labeling, which is an profitable attempt to bridge the tile-level scene classification toward pixel-level semantic parsing for aerial image interpretation.

In the future work, we will dedicate our efforts to enrich the Million-AID with more semantic categories and expand the scale of aerial scene images. Knowledge transfer by Million-AID will also be extended to other related tasks, such as object detection and semantic segmentation, to further explore the transferability of Million-AID. We hope that is work can enhance the development of content interpretation algorithms in the field of remote sensing.

References

Amitrano, D., Cecinati, F., Di Martino, G., Iodice, A., Mathieu, P.P., Riccio, D., Ruello, G., 2018. Feature extraction from multitemporal sar images using selforganizing map clustering and object-based image analysis. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 1556–1570.

Atkinson, P.M., 1997. Mapping sub-pixel boundaries from remotely sensed images, in: Innovations in GIS, pp. 184–202.

Audebert, N., Le Saux, B., Lefèvre, S., 2019. Deep learning for classification of hyperspectral data: A comparative review. IEEE Geosci. Remote Sens. Mag. 7, 159–173.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495.

Bateson, A., Curtiss, B., 1996. A method for manual endmember selection and spectral unmixing. Remote Sens. Environ. 55, 229–243.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogrammetry Remote Sens. 114, 24–31.

Bi, Q., Qin, K., Zhang, H., Li, Z., Xu, K., 2020. Radc-net: A residual attention based convolution network for aerial scene classification. Neurocomputing 377, 345–359.

Bi, Q., Qin, K., Zhang, H., Xia, G.S., 2021a. Local semantic enhanced convnet for aerial scene recognition. IEEE Trans. Image Process. 30, 6498–6511.

Bi, Q., Zhang, H., Qin, K., 2021b. Multi-scale stacking attention pooling for remote sensing scene classification. Neurocomputing 436, 147–161.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS

J. Photogrammetry Remote Sens. 65, 2–16.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., Van der Meer, F., Van der Werff, H., Van Coillie, F., et al., 2014. Geographic object-based image analysis–towards a new paradigm. ISPRS J. Photogrammetry Remote Sens. 87, 180–191.

Blaschke, T., Strobl, J., 2001. What's wrong with pixels? some recent developments interfacing remote sensing and gis. GIS-Zeitschrift für Geoinformationssysteme , 12–17.

Bovolo, F., Bruzzone, L., Carlin, L., 2010. A novel technique for subpixel image classification based on support vector machine. IEEE Trans. Image Process. 19, 2983–2999.

Broni-Bediako, C., Murata, Y., Mormille, L.H.B., Atsumi, M., 2021. Searching for cnn architectures for remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. , 1–13.

Bruzzone, L., Prieto, D.F., Serpico, S.B., 1999. A neural-statistical approach to multitemporal and multisource remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 37, 1350–1359.

Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J.A., 2013. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. IEEE Signal Process. Mag. 31, 45–54.

Chen, C.H., Peter Ho, P.G., 2008. Statistical pattern recognition in remote sensing. Pattern Recognit. 41, 2731–2741.

Chen, F., Tsou, J.Y., 2021. Drsnet: Novel architecture for small patch and low-resolution remote sensing image scene classification. Int. J. Appl. Earth Obs. Geoinf. 104, 102577.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proc. Eur. Conf. Comput. Vis., pp. 801–818.

Chen, S., Tian, Y., 2015. Pyramid of spatial relatons for scene-level land use classification. IEEE Trans. Geosci. Remote Sens. 53, 1947–1957.

Chen, Y., Wang, Y., Gu, Y., He, X., Ghamisi, P., Jia, X., 2019a. Deep learning ensemble for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12, 1882–1897.

Chen, Z.M., Wei, X.S., Wang, P., Guo, Y., 2019b. Multi-label image recognition with graph convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5177–5186.

Cheng, G., Cai, L., Lang, C., Yao, X., Chen, J., Guo, L., Han, J., 2021. SPNet: Siamese-prototype network for few-shot remote sensing image scene classification. IEEE Trans. Geosci. Remote Sens. , 1–11.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE 105, 1865–1883.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs .Remote Sens. 13, 3735–3756.

Cheng, G., Yang, C., Yao, X., Guo, L., Han, J., 2018. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. IEEE Trans. Geosci. Remote Sens. 56, 2811–2821.

Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J., 2001. Color image segmentation: advances and prospects. Pattern Recognit. 34, 2259–2281.

Cooper, S., Okujeni, A., Jänicke, C., Clark, M., van der Linden, S., Hostert, P., 2020. Disentangling fractional vegetation cover: Regression-based unmixing of simulated spaceborne imaging spectroscopy data. Remote Sens. Environ. 246, 111856.

Curran, P., Williamson, H., 1986. Sample size for ground and remotely sensed data. Remote Sens. Environ. 20, 31–41.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. Imagenet: A large-scale hierarchical image database, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 248–255.

Ediriwickrema, J., Khorram, S., 1997. Hierarchical maximum-likelihood classification for improved accuracies. IEEE Trans. Geosci. Remote Sens. 35, 810–816.

Espínola, M., Piedra-Fernández, J.A., Ayala, R., Iribarne, L., Wang, J.Z., 2014. Contextual and hierarchical classification of satellite images based on cellular automata. IEEE Trans. Geosci. Remote Sens. 53, 795–809.

Fan, J., Chen, T., Lu, S., 2017. Superpixel guided deep-sparse-representation learning for hyperspectral image classification. IEEE Trans. Circuits Syst. Video Technol. 28, 3163–3173.

Feng, J., Chen, J., Liu, L., Cao, X., Zhang, X., Jiao, L., Yu, T., 2019. CNN-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12, 1299–1313.

Feng, Z., Wang, M., Yang, S., Liu, Z., Liu, L., Wu, B., Li, H., 2017. Superpixel tensor sparse coding for structural hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 1632–1639.

Fernandes, R., Fraser, R., Latifovic, R., Cihlar, J., Beaubien, J., Du, Y., 2004. Approaches to fractional land cover and continuous field mapping: A comparative assessment over the boreas study region. Remote Sens. Environ. 89, 234–251.

Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.

Fu, L., Zhang, D., Ye, Q., 2020. Recurrent thrifty attention network for remote sensing scene recognition. IEEE Trans. Geosci. Remote Sens. , 1–12.

Gao, Q., Lim, S., Jia, X., 2019. Spectral–spatial hyperspectral image classification using a multiscale conservative smoothing scheme and adaptive sparse representation. IEEE Trans. Geosci. Remote Sens. 57, 7718–7730.

Gessner, U., Machwitz, M., Conrad, C., Dech, S., 2013. Estimating the fractional cover of growth forms and bare surface in savannas. a multi-resolution approach based on regression tree ensembles. Remote Sens. Environ. 129, 90–102.

Ghamisi, P., Maggiori, E., Li, S., Souza, R., Tarablaka, Y., Moser, G., De Giorgi, A., Fang, L., Chen, Y., Chi, M., et al., 2018. New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology, markov random fields, segmentation, sparse representation, and deep learning. IEEE Geosci. Remote Sens. Mag. 6, 10–43.

Ghamisi, P., Plaza, J., Chen, Y., Li, J., Plaza, A.J., 2017. Advanced spectral classifiers for hyperspectral images: A review. IEEE Geosci. Remote Sens. Mag. 5, 8–32.

Gong, Z., Zhong, P., Yu, Y., Hu, W., 2018. Diversity-promoting deep structural metric learning for remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 56, 371–390.

Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y., 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211 .

Gu, Y., Chanussot, J., Jia, X., Benediktsson, J.A., 2017. Multiple kernel learning for hyperspectral image classification: A review. IEEE Trans. Geosci. Remote Sens. 55, 6547–6565.

Han, X., Huang, X., Li, J., Li, Y., Yang, M.Y., Gong, J., 2018. The edge-preservation multi-classifier relearning framework for the classification of high-resolution remotely sensed imagery. ISPRS J. Photogrammetry Remote Sens. 138, 57–73.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man. Cybern. Syst. , 610–621.

Hay, G., Marceau, D., Dube, P., Bouchard, A., 2001. A multiscale framework for landscape analysis: object-specific analysis and upscaling. Landsc Ecol. 16, 471–490.

Hay, G.J., Castilla, G., 2008. Geographic object-based image analysis (geobia): A new name for a new discipline, in: Object-based Image Analysis. Springer, pp. 75–89.

He, D., Shi, Q., Liu, X., Zhong, Y., Zhang, X., 2021. Deep subpixel mapping based on semantic information modulated network for urban land use mapping. IEEE Trans. Geosci. Remote Sens. , 1–19.

He, D., Zhong, Y., Wang, X., Zhang, L., 2020. Deep convolutional neural network framework for subpixel mapping. IEEE Trans. Geosci. Remote Sens. , 1–22.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778.

He, L., Li, J., Liu, C., Li, S., 2017. Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines. IEEE Trans. Geosci. Remote Sens. 56, 1579–1597.

He, N., Fang, L., Li, S., Plaza, A., Plaza, J., 2018. Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans. Geosci. Remote Sens. 56, 6899–6910.

He, X., Chen, Y., 2020. Transferring cnn ensemble for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 18, 876–880.

Hodgson, M.E., 1988. Reducing the computational requirements of the minimum-distance classifier. Remote Sens. Environ. 25, 117–128.

Hong, D., Gao, L., Yao, J., Yokoya, N., Chanussot, J., Heiden, U., Zhang, B., 2021. Endmember-guided unmixing network (egu-net): A general deep learning framework for self-supervised hyperspectral unmixing. IEEE Trans. Neural Netw. Learn. Syst. .

Hossain, M.D., Chen, D., 2019. Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective. ISPRS J. Photogrammetry Remote Sens. 150, 115–134.

Hu, F., Xia, G.S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7, 14680–14707.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4700–4708.

Imani, M., Ghassemian, H., 2020. An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges. Inf. Fusion 59, 59–83.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proc. Int. Conf. Machine learn., pp. 448–456.

Izquierdo-Verdiguier, E., Zurita-Milla, R., 2020. An evaluation of guided regularized random forest for classification and regression tasks in remote sensing. Int. J. Appl. Earth Obs. Geoinf. 88, 102051.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., Ermon, S., 2019. Tile2vec: Unsupervised representation learning for spatially distributed data, in: Proc. AAAI Conf. Artif. Intell., pp. 3967–3974.

Jia, S., Jiang, S., Lin, Z., Li, N., Xu, M., Yu, S., 2021. A survey: Deep learning for hyperspectral image classification with few labeled samples. Neurocomputing 448, 179–204.

Kang, J., Fernandez-Beltran, R., Hong, D., Chanussot, J., Plaza, A., 2021. Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval. IEEE Trans. Geosci. Remote Sens. 59, 4355–4369.

Kang, J., Fernandez-Beltran, R., Ye, Z., Tong, X., Ghamisi, P., Plaza, A., 2020. Deep metric learning based on scalable neighborhood components for remote sensing scene characterization. IEEE Trans. Geosci. Remote Sens. 58, 8905–8918.

Kaur, B., Garg, A., 2011. Mathematical morphological edge detection for remote sensing images, in: Proc. Int. Conf. Electron. Comput. Technol., pp. 324–327.

Kavzoglu, T., Mather, P., 2003. The use of backpropagating artificial neural networks in land cover classification. Int. J. Remote Sens. 24, 4907–4938.

Khan, N., Chaudhuri, U., Banerjee, B., Chaudhuri, S., 2019. Graph convolutional network for multi-label vhr remote sensing scene recognition. Neurocomputing 357, 36–46.

Khatami, R., Mountrakis, G., Stehman, S.V., 2017. Mapping per-pixel predicted accuracy of classified remote sensing images. Remote Sens. Environ. 191, 156–167.

Kotaridis, I., Lazaridou, M., 2021. Remote sensing image segmentation advances: A meta-analysis. ISPRS J. Photogrammetry Remote Sens. 173, 309–322.

Kothari, N.S., Meher, S.K., Panda, G., 2020. Improved spatial information based semisupervised classification of remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 329–340.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Kumar, B., Dikshit, O., Gupta, A., Singh, M.K., 2020. Feature extraction for hyperspectral image classification: a review. Int. J. Remote Sens. 41, 6248–6287.

Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. Neurocomputing 338, 321–348.

Lee, H., Battle, A., Raina, R., Ng, A.Y., 2007. Efficient sparse coding algorithms, in: Adv. Neural Inf. Process. Syst., pp. 801–808.

Li, D., Wang, Q., Kong, F., 2020a. Adaptive kernel sparse representation based on multiple feature learning for hyperspectral image classification. Neurocomputing .

Li, E., Xia, J., Du, P., Lin, C., Samat, A., 2017. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 55, 5653–5665.

Li, H., Cui, Z., Zhu, Z., Chen, L., Zhu, J., Huang, H., Tao, C., 2021. Rs-metanet: Deep metametric learning for few-shot remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 59, 6983–6994.

Li, J., Bioucas-Dias, J.M., Plaza, A., 2010. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. IEEE Trans. Geosci. Remote Sens. 48, 4085–4098.

Li, J., Bioucas-Dias, J.M., Plaza, A., 2011. Hyperspectral image segmentation using a new bayesian approach with active learning. IEEE Trans. Geosci. Remote Sens. 49, 3947–3960.

Li, M., Zang, S., Zhang, B., Li, S., Wu, C., 2014a. A review of remote sensing image classification techniques: The role of spatio-contextual information. Eur. J. Remote Sens. 47, 389–411.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019. Deep learning for hyperspectral image classification: An overview. IEEE Trans. Geosci. Remote Sens. 57, 6690–6709.

Li, X., Ling, F., Du, Y., Feng, Q., Zhang, Y., 2014b. A spatial–temporal hopfield neural network approach for super-resolution land cover mapping with multi-temporal different resolution remotely sensed images. ISPRS J. Photogrammetry Remote Sens. 93, 76–87.

Li, Z., Wang, T., Li, W., Du, Q., Wang, C., Liu, C., Shi, X., 2020b. Deep multilayer fusion dense network for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1258–1270.

Lin, D., Lin, J., Zhao, L., Wang, Z.J., Chen, Z., 2021. Multilabel aerial image classification with a concept attention graph neural network. IEEE Trans. Geosci. Remote Sens. , 1–12.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2117–2125.

Liu, J., Li, P., Wang, X., 2015. A new segmentation method for very high resolution imagery using spectral and morphological information. ISPRS J. Photogrammetry Remote Sens. 101, 145–162.

Liu, J., Wu, Z., Wei, Z., Xiao, L., Sun, L., 2013. Spatial-spectral kernel sparse representation for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 2462–2471.

Liu, Q., Kampffmeyer, M., Jenssen, R., Salberg, A.B., 2020. Dense dilated convolutions' merging network for land cover classification. IEEE Trans. Geosci. Remote Sens. 58, 6309–6320.

Liu, S., Shi, Q., 2020. Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan china. ISPRS J. Photogrammetry Remote Sens. 164, 229–242.

Liu, W., Wu, E.Y., 2005. Comparison of non-linear mixture models: subpixel classification. Remote Sens. Environ. 94, 145–154.

Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C., 2018. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS J. Photogrammetry Remote Sens. 145, 78–95.

Liu, Y., Zhang, L., Han, Z., Chen, C., 2021. Integrating knowledge distillation with learning to rank for few-shot scene classification. IEEE Trans. Geosci. Remote Sens. , 1–12.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3431–3440.

Long, Y., Xia, G.S., Li, S., Yang, W., Yang, M.Y., Zhu, X.X., Zhang, L., Li, D., 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 4205–4230.

Lu, D., Weng, Q., 2007. A survey of image classification methods and tech-

niques for improving classification performance. Int. J. Remote Sens. 28, 823–870.

Lu, X., Gong, T., Zheng, X., 2020. Multisource compensation network for remote sensing cross-domain scene classification. IEEE Trans. Geosci. Remote Sens. 58, 2504–2515.

Lv, Z., Liu, T., Benediktsson, J.A., Falco, N., 2021. Land cover change detection techniques: Very-high-resolution optical images: A review. IEEE Geosci. Remote Sens. Mag. .

Ma, A., Wan, Y., Zhong, Y., Wang, J., Zhang, L., 2021a. Scenenet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. ISPRS J. Photogrammetry Remote Sens. 172, 171–188.

Ma, A., Wan, Y., Zhong, Y., Wang, J., Zhang, L., 2021b. Scenenet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. ISPRS J. Photogrammetry Remote Sens. 172, 171–188.

Ma, L., Cheng, L., Li, M., Liu, Y., Ma, X., 2015. Training set size, scale, and features in geographic object-based image analysis of very high resolution unmanned aerial vehicle imagery. ISPRS J. Photogrammetry Remote Sens. 102, 14–27.

Ma, L., Crawford, M.M., Tian, J., 2010. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 48, 4099–4109.

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017. A review of supervised object-based land-cover image classification. ISPRS J. Photogrammetry Remote Sens. 130, 277–293.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogrammetry Remote Sens. 152, 166–177.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657.

Manjunath, B.S., Ma, W.Y., 1996. Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell. 18, 837–842.

Martha, T.R., Kerle, N., van Westen, C.J., Jetten, V., Kumar, K.V., 2011. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. IEEE Trans. Geosci. Remote Sens. 49, 4928–4943.

Martins, V.S., Kaleita, A.L., Gelder, B.K., da Silveira, H.L., Abe, C.A., 2020. Exploring multiscale object-based convolutional neural network (multi-ocnn) for remote sensing image classification at high spatial resolution. ISPRS J. Photogrammetry Remote Sens. 168, 56–73.

Maulik, U., Chakraborty, D., 2017. Remote sensing image classification: A survey of support-vector-machine-based advanced techniques. IEEE Geosci. Remote Sens. Mag. 5, 33–52.

Mertens, K.C., Verbeke, L.P., Westra, T., De Wulf, R.R., 2004. Sub-pixel mapping and sub-pixel sharpening using neural network predicted wavelet coefficients. Remote Sens. Environ. 91, 225–236.

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. .

Ming, D., Li, J., Wang, J., Zhang, M., 2015. Scale parameter selection by spatial statistics for geobia: Using mean-shift based multi-scale segmentation as an example. ISPRS J. Photogrammetry Remote Sens. 106, 28–41.

Mou, L., Zhu, X.X., 2019. Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 58, 110–122.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS J. Photogrammetry Remote Sens. 66, 247–259.

Niu, R., Sun, X., Tian, Y., Diao, W., Chen, K., Fu, K., 2021. Hybrid multiple attention network for semantic segmentation in aerial images. IEEE Trans. Geosci. Remote Sens. .

Niu, Z., Liu, W., Zhao, J., Jiang, G., 2018. Deeplab-based spatial feature extraction for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 16, 251–255.

Nogueira, K., Penatti, O.A., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognit. 61, 539–556.

Ojala, T., Pietikäinen, M., Mäenpää, T., 2000. Gray scale and rotation invariant texture classification with local binary patterns, in: Proc. Eur. Conf. Comput. Vis., pp. 404–420.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24, 971–987.

Okwuashi, O., Ndehedehe, C.E., 2020. Deep support vector machine for hyperspectral image classification. Pattern Recognit. , 107298.

Othman, E., Bazi, Y., Melgani, F., Alhichri, H., Alajlan, N., Zuair, M., 2017. Domain adaptation network for cross-scene classification. IEEE Trans. Geosci. Remote Sens. 55, 4441–4456.

Otukei, J.R., Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. Int. J. Appl. Earth Obs. Geoinf. 12, S27–S31.

Ozkan, S., Kaya, B., Akar, G.B., 2019. EndNet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing. IEEE Trans. Geosci. Remote Sens. 57, 482–496.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359.

Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A., 2018. A new deep convolutional neural network for fast hyperspectral image classification. ISPRS J. Photogrammetry Remote Sens. 145, 120–147.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Proc. Adv. Neural Inf. Process. Syst., pp. 8024–8035.

Peng, C., Li, Y., Jiao, L., Chen, Y., Shang, R., 2019a. Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12, 2612–2626.

Peng, C., Li, Y., Jiao, L., Shang, R., 2021. Efficient convolutional neural architecture search for remote sensing image scene classification. IEEE Trans. Geosci. Remote Sens. 59, 6092–6105.

Peng, J., Li, L., Tang, Y.Y., 2019b. Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images. IEEE Trans. Neural Netw. Learn. Syst. 30, 1790–1802.

Pfülb, B., Gepperth, A., Abdullah, S., Kilian, A., 2018. Catastrophic forgetting: still a problem for dnns, in: Proc. Int. Conf. Artif. Neural Netw., pp. 487–497.

Porway, J., Wang, Q., Zhu, S.C., 2010. A hierarchical and contextual model for aerial image parsing. Int. J. Comput. Vis. 88, 254–283.

Pu, R., Landry, S., 2012. A comparative analysis of high spatial resolution ikonos and worldview-2 imagery for mapping urban tree species. Remote Sens. Environ. 124, 516–533.

Qi, X., Zhu, P., Wang, Y., Zhang, L., Peng, J., Wu, M., Chen, J., Zhao, X., Zang, N., Mathiopoulos, P.T., 2020. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. ISPRS J. Photogrammetry Remote Sens. 169, 337–350.

Raza, A., Huo, H., Sirajuddin, S., Fang, T., 2020. Diverse capsules network combining multiconvolutional layers for remote sensing image scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 5297–5313.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204.

Ren, Y., Zhang, X., Ma, Y., Yang, Q., Wang, C., Liu, H., Qi, Q., 2020. Full convolutional neural network based on multi-scale feature fusion for the class imbalance remote sensing image classification. Remote Sens. 12, 3547.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation, in: Proc. Int. Conf. Med. Image Comput. Comput. Assist Interv., Springer. pp. 234–241.

Ruder, S., 2017. An overview of multi-task learning in deep neural net-

works. arXiv preprint arXiv:1706.05098 .

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Van de Sande, K.E., Uijlings, J.R., Gevers, T., Smeulders, A.W., 2011. Segmentation as selective search for object recognition, in: Proc. IEEE Int. Conf. Comput. Vis., pp. 1879–1886.

Settle, J., Briggs, S., 1987. Fast maximum likelihood classification of remotely-sensed imagery. Int. J. Remote Sens. 8, 723–734.

Shackelford, A.K., Davis, C.H., 2003. A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas. IEEE Trans. Geosci. Remote Sens. 41, 1920–1932.

Shang, R., Liu, M., Lin, J., Feng, J., Li, Y., Stolkin, R., Jiao, L., 2021. Sar image segmentation based on constrained smoothing and hierarchical label correction. IEEE Trans. Geosci. Remote Sens. .

Shao, Y., Lunetta, R.S., 2012. Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points. ISPRS J. Photogrammetry Remote Sens. 70, 78–87.

Sharma, A., Liu, X., Yang, X., 2018. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. Neural Netw. 105, 346–355.

Sharma, A., Liu, X., Yang, X., Shi, D., 2017. A patch-based convolutional neural network for remote sensing image classification. Neural Netw. 95, 19–28.

Sheng, G., Yang, W., Xu, T., Sun, H., 2012. High-resolution satellite scene classification using a sparse coding based multiple feature combination. Int. J. Remote Sens. 33, 2395–2412.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Small, C., 2004. The landsat etm+ spectral mixing space. Remote Sens. Environ. 93, 1–17.

Somers, B., Asner, G.P., Tits, L., Coppin, P., 2011. Endmember variability in spectral mixture analysis: A review. Remote Sens. Environ. 115, 1603–1616.

Su, T., 2019. Scale-variable region-merging for high resolution remote sensing image segmentation. ISPRS J. Photogrammetry Remote Sens. 147, 319–334.

Su, T., Liu, T., Zhang, S., Qu, Z., Li, R., 2020. Machine learning-assisted region merging for remote sensing image segmentation. ISPRS J. Photogrammetry Remote Sens. 168, 89–123.

Sun, G., Zhang, X., Jia, X., Ren, J., Zhang, A., Yao, Y., Zhao, H., 2020. Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images. Int. J. Appl. Earth Obs. Geoinf. 91, 102157.

Swain, M.J., Ballard, D.H., 1991. Color indexing. Int. J. Comput. Vis. 7, 11–32.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–9.

Tang, Y., Qiu, F., Jing, L., Shi, F., Li, X., 2020. Integrating spectral variability and spatial distribution for object-based image analysis using curve matching approaches. ISPRS J. Photogrammetry Remote Sens. 169, 320–336.

Tong, X.Y., Xia, G.S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322.

Toth, C., Jóźków, G., 2016. Remote sensing platforms and sensors: A survey. ISPRS J. Photogrammetry Remote Sens. 115, 22–36.

Tu, B., Huang, S., Fang, L., Zhang, G., Wang, J., Zheng, B., 2018. Hyperspectral image classification via weighted joint nearest neighbor and sparse representation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 4063–4075.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J., 2011. A survey of active learning algorithms for supervised remote sensing image classification. IEEE J. Sel. Top. Signal Process. 5, 606–617.

Van Genderen, J., Lock, B., Vass, P., 1978. Remote sensing: Statistical testing of thematic map accuracy. Remote Sens. Environ. 7, 3–14.

Wang, D., Du, B., Zhang, L., 2021. Fully contextual network for hyperspectral scene parsing. IEEE Trans. Geosci. Remote Sens. .

Wang, F., 1990. Fuzzy supervised classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. 28, 194–201.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W., 2016. CNN-RNN: A unified framework for multi-label image classification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2285–2294.

Wang, Q., Atkinson, P.M., 2017. The effect of the point spread function on sub-pixel mapping. Remote Sens. Environ. 193, 127–137.

Wang, Q., Liu, S., Chanussot, J., Li, X., 2019. Scene classification with recurrent attention of vhr remote sensing images. IEEE Trans. Geosci. Remote Sens. 57, 1155–1167.

Wang, Q., Zhang, C., Atkinson, P.M., 2020. Sub-pixel mapping with point constraints. Remote Sens. Environ. 244, 111817.

Wang, Y., Zhang, L., Tong, X., Nie, F., Huang, H., Mei, J., 2018. LRAGE: Learning latent relationships with adaptive graph embedding for aerial scene classification. IEEE Trans. Geosci. Remote Sens. 56, 621–634.

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: A meta-review. Remote Sens. Environ. 236, 111402.

Wellmann, T., Lausch, A., Andersson, E., Knapp, S., Cortinovis, C., Jache, J., Scheuer, S., Kremer, P., Mascarenhas, A., Kraemer, R., et al., 2020. Remote sensing in urban planning: Contributions towards ecologically sound policies? Landsc. Urban Plan. 204, 103921.

Wu, C., Murray, A.T., 2003. Estimating impervious surface distribution by spectral mixture analysis. Remote Sens. Environ. 84, 493–505.

Wurm, M., Stark, T., Zhu, X.X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. ISPRS J. Photogrammetry Remote Sens. 150, 59–69.

Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018a. DOTA: A large-scale dataset for object detection in aerial images, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3974–3983.

Xia, G.S., Delon, J., Gousseau, Y., 2010a. Shape-based invariant texture indexing. Int. J. Comput. Vis. 88, 382–403.

Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55, 3965–3981.

Xia, G.S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Maître, H., 2010b. Structural high-resolution satellite image indexing, in: Proc. ISPRS TC VII Symposium - 100 Years ISPRS, pp. 298–303.

Xia, J., Ghamisi, P., Yokoya, N., Iwasaki, A., 2018b. Random forest ensembles and extended multiextinction profiles for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 56, 202–216.

Xu, F., Somers, B., 2021. Unmixing-based sentinel-2 downscaling for urban land cover mapping. ISPRS J. Photogrammetry Remote Sens. 171, 133–154.

Yang, J., He, Y., Caspersen, J., 2017. Region merging using local spectral angle thresholds: A more accurate method for hybrid segmentation of remote sensing images. Remote Sens. Environ. 190, 137–148.

Yang, K., Liu, Z., Lu, Q., Xia, G.S., 2019. Multi-scale weighted branch network for remote sensing image classification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, pp. 1–10.

Yang, W., Yin, X., Xia, G.S., 2015. Learning high-level features for satellite image classification with limited labeled samples. IEEE Trans. Geosci. Remote Sens. 53, 4472–4482.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification, in: Proc. Int. Conf. Adv. Geographic Inf. Syst., pp. 270–279.

Yu, J., Wang, B., Lin, Y., Li, F., Cai, J., 2021. A novel inequality-constrained weighted linear mixture model for endmember variability. Remote Sens. Environ. 257, 112359.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., et al., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716.

Yuan, Y., Fang, J., Lu, X., Feng, Y., 2019. Remote sensing image scene classification using rearranged local features. IEEE Trans. Geosci. Remote Sens. 57, 1779–1792.

Zafari, A., Zurita-Milla, R., Izquierdo-Verdiguier, E., 2020. A multiscale random forest kernel for land cover classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 2842–2852.

Zhang, C., Kim, J., 2019. Object detection with location-aware deformable convolution and backward attention filtering, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 9452–9461.

Zhang, C., Li, G., Du, S., 2019. Multi-scale dense networks for hyperspectral remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 57, 9201–9222.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018a. An object-based convolutional neural network (ocnn) for urban land use classification. Remote Sens. Environ. 216, 57–70.

Zhang, C., Yue, P., Tapete, D., Shangguan, B., Wang, M., Wu, Z., 2020a. A multi-level context-guided classification method with object-based convolutional neural network for land cover classification using very high resolution remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 88, 102086.

Zhang, J., Liu, J., Pan, B., Shi, Z., 2020b. Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification. IEEE Trans. Geosci. Remote Sens. 58, 7920–7930.

Zhang, M., Li, W., Du, Q., 2018b. Diverse region-based cnn for hyperspectral image classification. IEEE Trans. Image Process. 27, 2623–2634.

Zhang, X., Xiao, P., Feng, X., Wang, J., Wang, Z., 2014. Hybrid region merging method for segmentation of high-resolution remote sensing images. ISPRS J. Photogrammetry Remote Sens. 98, 19–28.

Zhang, Y., Yang, Q., 2021. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng. , 1–20.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017a. Pyramid scene parsing network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2881–2890.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017b. Pyramid scene parsing network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2881–2890.

Zhao, J., Zhong, Y., Shu, H., Zhang, L., 2016. High-resolution image classification integrating spectral-spatial-location cues by conditional random fields. IEEE Trans. Image Process. 25, 4033–4045.

Zheng, X., Huan, L., Xia, G.S., Gong, J., 2020. Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss. ISPRS J. Photogrammetry Remote Sens. 170, 15–28.

Zhong, Y., Zhu, Q., Zhang, L., 2015. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 53, 6207–6222.

Zhou, W., Jin, J., Lei, J., Hwang, J.N., 2021. Cegfnet: Common extraction and gate fusion network for scene parsing of remote sensing images. IEEE Trans. Geosci. Remote Sens. .

Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogrammetry Remote Sens. 145, 197–209.

Zhu, Q., Zhong, Y., Zhao, B., Xia, G.S., Zhang, L., 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci. Remote Sens. Lett. 13, 747–751.

Zhu, S., Du, B., Zhang, L., Li, X., 2021. Attention-based multiscale residual adaptation network for cross-scene classification. IEEE Trans. Geosci. Remote Sens. , 1–15.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 12, 2321–2325.