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

Yang Long, Gui-Song Xia, Wen Yang, Liangpei Zhang, Deren Li

***Abstract*—Aerial image recognition has become an active topic due to its crucial role in a wide range of applications. The interpretation methods for aerial image recognition have been developing with the improvement of image quality, of which the interpretation performance has been significantly promoted by transferring natural image knowledge with data-driven approaches. In this context, this paper addresses the aerial image recognition from tile-level scene classification to pixel-level semantic parsing after reviewing the aerial image interpretation research. Specifically, we first conduct the review by revisiting the development of aerial image interpretation prototypes and depict their connections with aerial image characters. We then present a large-scale aerial image recognition dataset which consists of more than a million scene instances, termed Million-AID. To provide reliable benchmark for future research, we also report multi-class and multi-label scene classification experiments on Million-AID using the widely employed convolutional neural networks (CNNs). Finally, we explore the transferability of semantic scene knowledge of Million-AID to advance aerial image interpretation from tile-level scene classification to pixel-level semantic parsing. Intensive experiments show that scene recognition on Million-AID is of great challenge and thus able to serve as evaluation benchmark for aerial scene classification algorithms. For scene knowledge transfer, CNN models pre-trained on Million-AID show considerable superiority than those on ImageNet for tile-level semantic interpretation, which demonstrate the strong generalization ability of the proposed Million-AID. Moreover, our designed hierarchical multi-task learning methods achieves the state-of-the-art performance for pixel-level semantic parsing on the challenging GID, which is a profitable attempt to bridge the tile-level scene classification toward pixel-level semantic parsing for aerial image interpretation. We hope our work could serve as a baseline for aerial scene recognition and inspire rethinking the semantic classification of high resolution aerial images.**

Remote sensing image interpretation, Million-AID, scene classification, semantic segmentation, transfer learning

## I. Introduction

With the advancement of aerospace technology, aerial images from various platforms have greatly helped us to measure the detailed features on the Earth's surface [1], [2] and achieved widespread applications in agriculture production [3], [4], urban planning [5], [6], environmental monitoring [7], [8], and disaster management [9], [10]. However, the convenient accessibility of aerial images poses great challenges and urgent demands for intelligent interpretation for the growing volume of received images [11]. To this end, aerial image recognition which aims to automatically extract local content information by tile-level semantic classification is employed for aerial image interpretation [12]–[15]. Despite of its advantage of interpretation efficiency, conventional aerial image recognition namely scene classification mainly focuses on extracting the abstract semantics from tile level while facing challenges in acquiring finer interpretation result, *e.g.*, accurate boundaries with pixel-level semantics. This makes the tile-level aerial image recognition suffer from serious accuracy problem that hamper its utilization in practical applications.

Aerial image recognition has come with the improvement of image resolution and the change of interpretation requirements, falling into per-pixel, object-based, and tile-level semantic classification. In the early years, the aerial images possess low spatial resolution which enables individual pixels contain large areas with distinct ground features [16]–[21] using spectral and textural features [22]–[27]. Until further notice, aerial image interpretation mainly focused on per-pixel classification. As the spatial resolution is gradually improved to be finer than the ground features, object-based image analysis that extracts homogeneous semantic entities was widely employed [28]–[30], particularly relying on the morphological segmentation methods [31]. However, the recognition of aerial images require complex relationship modeling for different objects or scene components [32]. Consequently, tile-level scene classification paved the way for aerial image interpretation by integrating the complicated features and image content as a whole for semantic classification [12], [15], [33]–[36]. In this scheme, the feature representation acts as one of the crucial roles in extracting the semantic information of scene content. However, due to the complexity of image content, semantic scenes are usually characterized with rich spectral, structural, and textural details, which make it difficult to design stable scene features and classifiers. Recently, data-driven methods represented by the deep learning have attracted extensive attention due to the strong capability of learning high-level representation of semantic content in aerial images [33], [37]. Thus, a variety of methods based on deep learning have been developed to cope with aerial image recognition and reported exciting performance [34]–[47]. Nevertheless, there are two major issues that hamper the further development of aerial image recognition:

- *The divergence of aerial image interpretation prototypes*

Y. Long, L. Zhang, D. Li are with the State Key Lab. LIESMARS, Wuhan University, Wuhan, 430079, China. e-mail: {*longyang, zlp62, drli*}@whu.edu.cn

G.-S. Xia is with the Department of Computer Science and the State Key Lab. LIESMARS, Wuhan University, Wuhan, 430079, China. e-mail: *guisong.xia*@whu.edu.cn

S. Li is with the Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China. e-mail: *shyli*@csa.ac.cn

W. Yang is with the School of Electronic Information and the State Key Lab. LIESMARS, Wuhan University, Wuhan, 430079, China. e-mail: *yangwen*@whu.edu.cn

*in tile-level and pixel-wise classification.* Pixel-wise classification has experienced a long way in extracting fine semantic information for aerial image interpretation while tile-level classification is usually employed to extract coarse but high-level semantic information of a local area composed of components with the same thematic meaning. However, the continuous improvement of image resolution brings aerial images of large scale and huge volume. As a result, existing pixel-wise aerial image interpretation methods that address fine classification result are confronted with serious efficiency problems because of the high computational cost. And the spatial, spectral, temporal variation over pixel units also raise increasingly higher requirements for the capacity of pixel-wise classification methods. From the perspective of image expression, the improvement of image resolution 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 on solely on its own. In this situation, now is the time to reconsider the homogeneous representation in tile-level and bridge its gap to pixel-level semantic classification.

- *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 hinder the utilization of data-driven methods for aerial image interpretation. To alleviate this issue, there are a lot of interpretation methods that directly employ the models trained on natural image archives (*e.g.*, ImageNet [48]) as the feature extractor for aerial image scenes. However, the differences in the way that the semantic category of natural and RS image scenes are defined can cause semantic bias, and thus, inevitably limit the accurate characterization of aerial image content. In addition, the strategy of fine-tuning models pre-trained on large-scale natural image archives is widely employed in RS community. Nevertheless, the essential structural and content difference of natural and RS images can make an inevitable influence on giving full play to the advantages of deep CNNs. In this context, the transferability ability of data-driven models adapted with pure aerial scene images become necessary to be explored to free up the potential of deep learning methods.

Due to the aforementioned issues, this paper first provides a review on aerial image interpretation. Then, we present a large-scale aerial image scene classification dataset, *i.e.*, Million-AID. The benchmarking experiments were conducted to investigate how well the current deep learning methods perform on Million-AID for aerial scene classification. In addition, we verified the effectiveness of knowledge transfer by utilizing the scene knowledge of Million-AID. To sums up, our main contributions are as follows:

- We released a large-scale dataset, i.e., Million-AID, which is a publicly available benchmark for aerial scene recognition. To the best of our knowledge, this is the aerial scene dataset of largest scale in the remote sensing community, characterized by diverse semantic scene categories, large number of annotated scene instances, global geographical distribution, and rich image variation.
- We investigated a set of representative aerial scene recognition approaches, *i.e.*, the classical CNN models, to explore how well the current scene recognition solutions of deep learning perform on Million-AID. The experimental results will provide the research community a benchmark foundation for the development of aerial scene recognition algorithms.
- We conducted extensive experiments to verify the tremendous potential of transferring the semantic scene knowledge of Million-AID to advance aerial image interpretation from tile-level scene classification to pixel-levle semantic parsing. Fine-tuning CNN models pre-trained on Million-AID show considerable superiority than those on ImageNet for tile-level semantic interpretation. The designed hierarchical multi-task learning method by employing Million-AID and GID achieves the state-of-the-art result for pixel-level semantic interpretation, which simultaneously demonstrate the importance of tile-level classification bridging to pixel-level semantic parsing for aerial image interpretation.

The remainder of this paper is organized as follows. Section II presents a review of aerial image interpretation from its development perspective. Section III introduces the proposed large-scale scene classification dataset, *i.e.*, Million-AID. Section IV present the comprehensive benchmarking experiments on Million-AID, including multi-class and multi-label aerial scene classification. Section V presents experiments on knowledge transfer by Million-AID for tile-level scene classification and pixel-level semantic parsing. Finally, in Section VI, we draw conclusions regarding this work.

## II. Revisiting Aerial Image Interpretation

With the progress of sensor technology, the quality particularly spatial resolution of aerial image has been continuously improved [1], [49]. Figure 1 presents the milestones of earth observation satellites at different times. Accordingly, the improvement of aerial image quality has greatly promoted the development of aerial image interpretation as stated previously. In this section, we focus on a review by revisiting the development of aerial image interpretation and the outline is also presented in Figure 1.

*1) Per-pixel aerial image classification:* In the early 1970s, aerial images are characterized with low spatial resolution where each pixel represents an area of thousands of square meters of the Earth's surface, *e.g.*, pixels in LandSat-1 and MODIS images. And the sizes of ground features or objects are usually smaller than the ground sampling distance of image pixels. In this sense, each pixel is able present a scene of specific semantic category. Thus, 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 per-pixel classification using spectral signatures [22], [24] and coarse textural features [23], [25]–[27]. To this end, sampling analysis is naturally employed [50]–[52] to construct desirable


Fig. 1. The development outline of aerial image interpretation. The interpretation prototypes develop with the quality improvement of aerial images and has experienced a long course of development ranging from per-pixel classification, object-based analysis, to scene-level understanding. This figure non-exhaustively presents some representative works concerning aerial image recognition. In this work, we will make our efforts to perform pixel-level semantic aerial image parsing that bridges tile-level scene classification toward pixel-level semantic parsing for high resolution aerial images.

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 [17], [19], [53]–[56]. To obtain reliable classification results, a number of classifiers were developed for the per-pixel aerial image parsing, such as maximum likelihood methods [57]–[59], minimum distance to means algorithms [60], [61], K-nearest neighbors classifiers [62], [63], and tree-based techniques [64], [65]. 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 [66]–[69], sparse representation [59], [70]–[74], artificial neural network [75], [76], and kernel-based methods [69], [71], [74], [77] represented by support vector machine [78]–[80] were intensively explored by actively embracing the machine learning techniques. These methods have made dramatic progress in per-pixel aerial image parsing owing to their strong ability of discriminating the complex spectral and spatial characteristics of ground features.

Typically, per-pixel 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 [20], [81]. 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 come to be a nonnegligible problem in the medium and coarse spatial resolution data, causing an appreciable effect on the use of aerial images interpreted in a per-pixel classification way. To overcome this problem, sub-pixel classification is considered as an alternative for more accurate aerial image parsing [82]–[85]. A number of approaches have been derived to address the sub-pixel aerial image classification, including soft or fuzzy theory [16], [86], [87], neural networks [88]–[90], regression modeling and analysis [91]–[93], and spectral mixture analysis [18], [94]–[96]. Among these methodologies, 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 [94], [95]. Thus, the selection of endmembers becomes one of the key points for designing an effective classifier [97]–[100]. 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 [84], [101],

[102]. In this scheme, each pixel is divided into sub-pixels which are predicted to get predefined single semantic labels. Limited by the spatial resolution, aerial images interpreted by per-pixel classification still face challenges in acquiring satisfactory result due to the mixture and complexity of image content within single pixels.

*2) Object-based Aerial Image Analysis:* With the improvement of sensor technology, the spatial resolution of aerial images is gradually improved to be much smaller than the scales of ground features and objects of interest. Apart from that, more rich detail of spectrum, texture, and particularly geometric structure becomes prominent in images. 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 [103]. And the improved image quality also significantly increases the within-class variability, which decreases the potential accuracy of purely pixel-based approach to classification [30]. As a result, traditional interpretation system established with per-pixel 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. 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) [28]–[31], [103]–[105], 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 [106]. 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 fundamental elements of OBIA [105], [107].

In high resolution of RS images, ground objects are presented with much richer spectral, textural, structural, and contextual detail that reveals pattern characteristics [30]. These enable the objects of interest to be obtained by spectrally-based and spatially-based segmentation approaches, among which mathematical morphology analysis plays a significant role [31], [108]. Hence, the thresholding [109]–[111] and feature space clustering [112] 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 [110], [113]–[115], region generation (*e.g*, region growing [109], [116], merging, and splitting [110], [117], [118]), hybrid segmentation [110], [119], [120] techniques are conducted according to the discontinuity and similarity of object areas. However, the acquired objects 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 this pipeline, the feature extraction and classifier design play the crucial roles in classification performance [121], [122]. Recently, CNN frameworks have shown overwhelming advantage in visual feature extraction and classification, which are integrated into OBIA and triggered the new trend of object-based CNN (OCNN) for aerial image interpretation [107], [123], [124]. 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 per-pixel classification methods, the object-based classification methods face challenges in parameter setting and optimization (*e.g.*, segmentation scale) [125], [126], which affect the image segmentation quality as well as the final classification accuracy. In addition, the segmentation and classification of objects falls into a multi-step pipeline, which inevitably affects the classification efficiency and increases the difficulty of model deployment.

To overcome the deficiencies of OBIA, the solution that simultaneously produces the segmented homogeneous areas and the corresponding semantic categories becomes an imperious demand. In recent years, aerial image classification have been greatly facilitated by deep convolutional neural networks (DCNNs) [127]–[130], among which the fully convolutional network (FCN) and its advanced models [131] provides 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 [132]. 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 details of different classes is a serious issue. To deal with these issues, the multi-scale features [133]–[136] and contextual information [124], [137] are typically considered to enhance the feature representation ability. The atrous and paralleled dilation convolution [138] is utilized for sampling strategy to preserve feature resolution [139]. And the ensemble of multiple networks is explored to take advantages of different frameworks for feature complementary and achieved encouraging classification performance [140], [141]. In addition, spatial features have been reported to be helpful in improving the classification performance [142]–[144]. Owing to the advantage of CNN frameworks and aerial image signatures, spatial and spectral information are inherently integrated into the pixel-wise classifiers to address the challenge of large spatial variability of spectral signatures [145]–[148]. Regarding the limitation of training samples and interpretation generalization ability, transform leaning have been intensively explored to address the limitation of training samples for CNN frameworks and reported promising classification results [140], [141], [149], [150]. The readers may go to one of the review papers for more comprehensive perspective of semantic segmentation using deep learning techniques [132], [142], [143], [148], [151]–[153]. 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 computational burden when classifying aerial images of large size and huge volume owing to the improvement of spectral and spatial resolution.

*3) Tile-level Aerial Image Understanding:* In the era of high resolution remote sensing, the content of aerial images are usually characterized with diverse visual features, complicated components, and uncertain spatial arrangements. Even with impressive success achieved by object-based analysis, the semantics obtained by object classification could be wrong because of the confused features among homogeneous regions of different objects. Besides, individual objects carry information independent to their neighbors and thus neglect the thematic meaning in their contextual environment. These make the interpretation of aerial image with high resolution a challenge task. To alleviate this problem, the hierarchical and contextual model for aerial scene image parsing is designed by organizing individual objects into hierarchical groups [32]. However, the implementation of simultaneously detecting objects and organizing them into a hierarchical contextual representation for the scene is can be difficult. Thus, the tile-level scene classification, which is able to incorporate visual features, scene components, and spatial arrangements as a whole, becomes an effective scheme for aerial image interpretation [12]. In the last decade, a handful of visual descriptors have been employed for aerial scene classification [15], [34]–[36], [154], [155]. We refer interested readers to [12], [34], [35] for a survey of the low-level and middle-level visual features employed for aerial scene classification.

Thanks to the increasing accessibility of aerial images and availability of computational resources, 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 employed as feature extractors owing to its simplify and efficiency [33], [38], [156]. However, the recognition of aerial scenes is a challenging task owing to the complexity of scene content in high resolution aerial images. To improve the feature discrimination capacity and adapt to the scale variation of ground objects, multi-scale images or features are extracted and fused to generate robust global representation for scene classification [38], [156]–[158]. In fact, semantic categories of aerial scenes usually depend on the complicated spatial arrangement and class-specific objects in images. Thus, deep local structures related to scene category are addressed to improve the classification performance [47], [159]. Furthermore, CNNs based on attention mechanism are developed to highlight more local semantics and discard the noncritical information [158], [160]–[162]. With the improvement of spatial resolution of aerial images, the within-class diversity and between-class similarity of semantic scenes are greatly increased, which make the scene recognition a challenging task. To relieve this issue, deep metric learning algorithms are developed to learn more discriminative category features [39], [163], [164]. Particularly, the complex relationship pervading aerial scenes are further explored in the embedding space by learning deep graph networks [165]–[167]. The core idea of these strategies 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 capturing the essential content of aerial scenes, automatically learning the CNN architecture are intensively explored [168]–[171]. To overcome the problem of limited annotation data, scene classification based on few-shot learning has attracted extensive attention [172]–[174]. And annotated scene images from different domains are employed to relieve the issue of data dependency [175]–[178]. These approaches have reported exciting performance on aerial scene classification. 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. Consequently, the recent scene classification algorithms have intensively reported saturation results as shown in [36]. Faced with this situation, the potential of the data-driven methods for scene classification remains to be further explored and boosted by large-scale annotated datasets.

Even with great achievements, aerial image interpretation based on tile-level scene classification is not able to provide result with accurate semantic boundaries at pixel level. With the availability of enormous amount of aerial images with improved spectral, spatial, temporal, and radiometric resolutions, the real bottleneck of aerial image interpretation has come to be the requirement of accurate and efficient image classification over large areas. On the one hand, conventional pixel-wise and object-based classification requires a large amount of high-cost pixel-level annotations for model adaption while suffering from the increased computational and memory consumption. These factors have limited the potential of current aerial image classification algorithms in either precision or efficiency to some extent. On the other hand, the existing tile-level aerial aerial classification schemes for a long time focuses on summarizing the scene content over large grid areas which neglects emphasis on the homogeneous components in pixel-level semantics. As a result, the tile-level classification could achieve only coarse interpretation result that hampers its utilization in practical applications. Currently, with the continuous improvement of aerial image resolution, the scale of homogeneous regions in high resolution aerial images is becoming increasingly large. And the semantics of individual pixels or objects in an aerial image are closely related to their contextual environment while gradually losing the significance of independent existence. Fortunately, the tile-level representation which simultaneously takes contextual and semantic meaning for the central unit becomes an reasonable alternative to perform fine pixel-level semantic classification of high resolution aerial images. And it is obvious that the acquisition of tile-level labels is much easier than those of pixel-level ones when using fully supervised classification methods. In this manuscript, we will make an attempt to demonstrate the effectiveness of tile-level representation in achieving pixel-level semantic parsing for aerial images.

## III. An Introduction to Million-AID

Due to the difficulty of aerial image collection and high cost of semantic annotation, there is a lack of large-scale aerial scene datasets which play a vital role in the development of aerial image interpretation algorithms. In this section, we detail the Million-AID dataset to be released for aerial scene classification.

### A. Dataset Construction

The aerial image datasets for scene classification have achieved significant development and become more rich than ever before, in aspects like scale and diversity. The readers may go to [179] for a comprehensive review on datasets for aerial image interpretation. Even with great achievements, the existing scene classification datasets to some extent have been faced with the problem of accuracy saturation, particularly for the data-driven methods [36]. As a result, the interpretation models usually turns out to be hard to meet the practical requirements, such as the comprehensive model evaluation and strong generalization ability in applications.

Faced with this situation, a representative and large-scale aerial scene classification dataset is desperately desired for its potential to promote the aerial image interpretation in algorithm development and practical application. In reality, aerial image scenes are typically characterized with complex content, diverse structures, and wide distribution, which make it difficult to determine the scene category and collect the abundant scene images. In the routing of Million-AID construction, a reliable category organization system is established by referring to land use and land cover standard as well as the common scene categories. For scene image acquisition, the scheme of geographical coordinates collection is developed. Specifically, the public geographic information, open geographic databases, and open source geographic data are intensively utilized to obtain the geographic location information of scene images. And the aerial scenes are generally divided into different features according to the point, line and plane structures of geographical information. Thus, the scene images closely combined with the point, line and plane features are extracted based on the aerial image database, *i.e.*, Google Earth. The whole construction process of Million-AID falls into a semi-automatic way, which enable the dataset to be created with large scale, high efficiency, and high quality assurance. We here refer the interesting readers to [179] for more details about the creation of Million-AID.

### B. Scene Categories

The semantic scenes in Million-AID are organized hierarchically by referencing the land-use classification standards as well as the common RS image scenes. 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 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 remote sensing image interpretation, and some scene instances of each class are shown in Fig. 2. Moreover, as the scenes in Million-AID are organized hierarchically as mentioned before, 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 category 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.

### C. Dataset Scale

Apart from the wide coverage of semantic scene 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. In addition, the number of scene images in each semantic category ranges from about 2,000 to 45,000 as shown in Fig. 3. As can be seen, the number of images varies greatly among different categories, endowing the dataset with the property of unbalanced distribution. Taking the widely used AID [35] and NWPU-RESISC45 [34] as comparision, our proposed Million-AID surpasses them hugely in both the numbers of scene categories and images. Recently, data-driven methods, *i.e.*, deep learning, have shown promising perspectives for intelligent image interpretation in both computer vision and RS community, benefiting form the huge available dataset ontology. The publication of Million-AID make it possible to further boost the design and learning of aerial scene interpretation algorithms using data-driven schemes.

### D. Geographical Distribution

With the change of geographical environment and background information, aerial image scenes will show different patterns in appearance, structure and content, which pose great challenges to aerial image scene interpretation. Therefore, the wide geographical distribution of scenes in an aerial image dataset can contribute to better represent the real-world situation. However, most scene images of the existing datasets are located within local or limited regions, which are insufficiently to reflect the scene distribution comprehensively. In order to make the Million-AID more consistent with the distribution of the scenes in the real world, we collect aerial image scenes globally. Benefiting from our semi-automatic dataset construction scheme by utilizing the geographical information, we are able to acquire the geographical positions of scene image blocks in Milllion-AID. The distribution of scene image in Million-AID is shown in Fig. 4. It can be seen from the

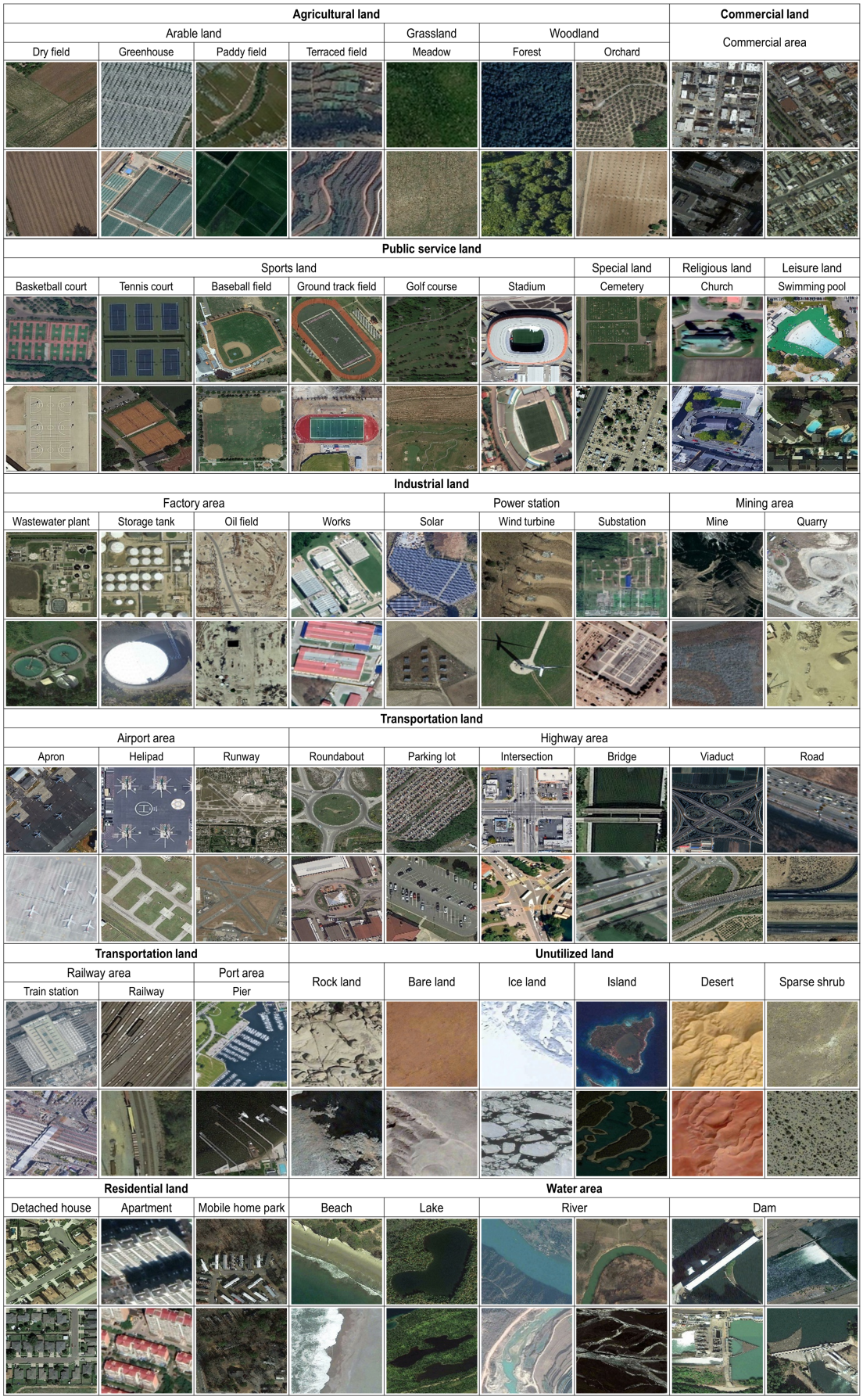



Fig. 2. Scene samples of Million-AID: two or four examples of each semantic category are presented. All the semantic scenes are organized by the hierarchical system with three-level semantic labels, containing 51 fine-gained scene categories belonging to 8 major categories.

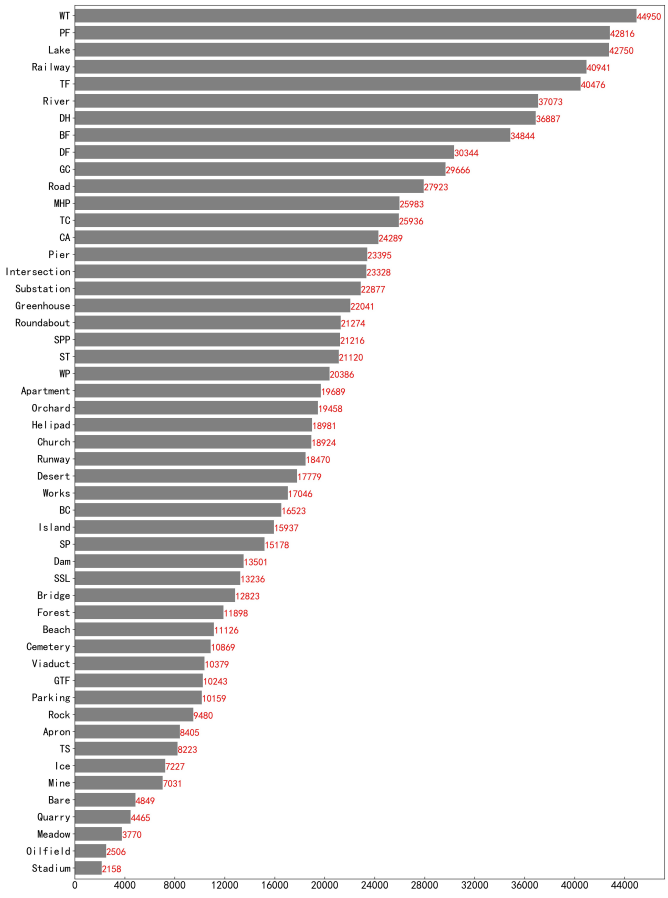



Fig. 3. The number of instances in each scene category. Zoom for detail.

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 are 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.

### *E. Image Variation*

Rich variation of RS images can greatly enhance the diversity of a dataset, so as to better represent the scene and feature distribution in the real world. With this point in mind, plenty of significant factors are taken into consideration to make Million-AID approximate practical situations as far as possible in the process of dataset construction. Aerial scenes are usually characterized with complex content and various scales. In order to keep consistent with this reality, scene images in Million-AID are with various sizes and resolutions. This is different from the existing scene classification datasets, in which scene images are resized to fixed sizes ( *i.e.*, $256\times256$ and $512\times512$) and the resolutions of images are usually with a narrow range. In Million-AID, the width of scene images ranges from 100 to 30,000 pixels and the spatial resolution ranges from about 0.2m to 153m per pixel. These features ensures the completeness and authenticity of the real-world scene content reflected by aerial images. 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 aerial image scene recognition task.

Furthermore, owing to the high complexity of earth surface features, scene content in aerial images usually show remarkable difference in appearance characterized with different geometrical shapes, structural properties, and texture attributes. This requires the constructed 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 provide sufficient assurance of intra-class diversity for Million-AID. As Million-AID contains a lot of scene classes, scene images of sub-classes are usually contained in the same major classes. This enables the scene images of the sub-classes to possess high inter-class similarity inherently because they share common characteristics of the major scene classes. In general, the presented Million-AID is of great capacity to represent aerial scenes and feature distribution in the real world, and thus, able to boost the establishment of public comparison platforms and development of data-driven interpretation algorithms for practical applications.

## IV. 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 [34], [35], and thus, dominated aerial image recognition in recent years [36]. 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.

### *A. 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. With 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 [180], VGG16 [181], GoogleNet [182], ResNet101 [183], DenseNet121 [184], and DenseNet169 [184] 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 a wide

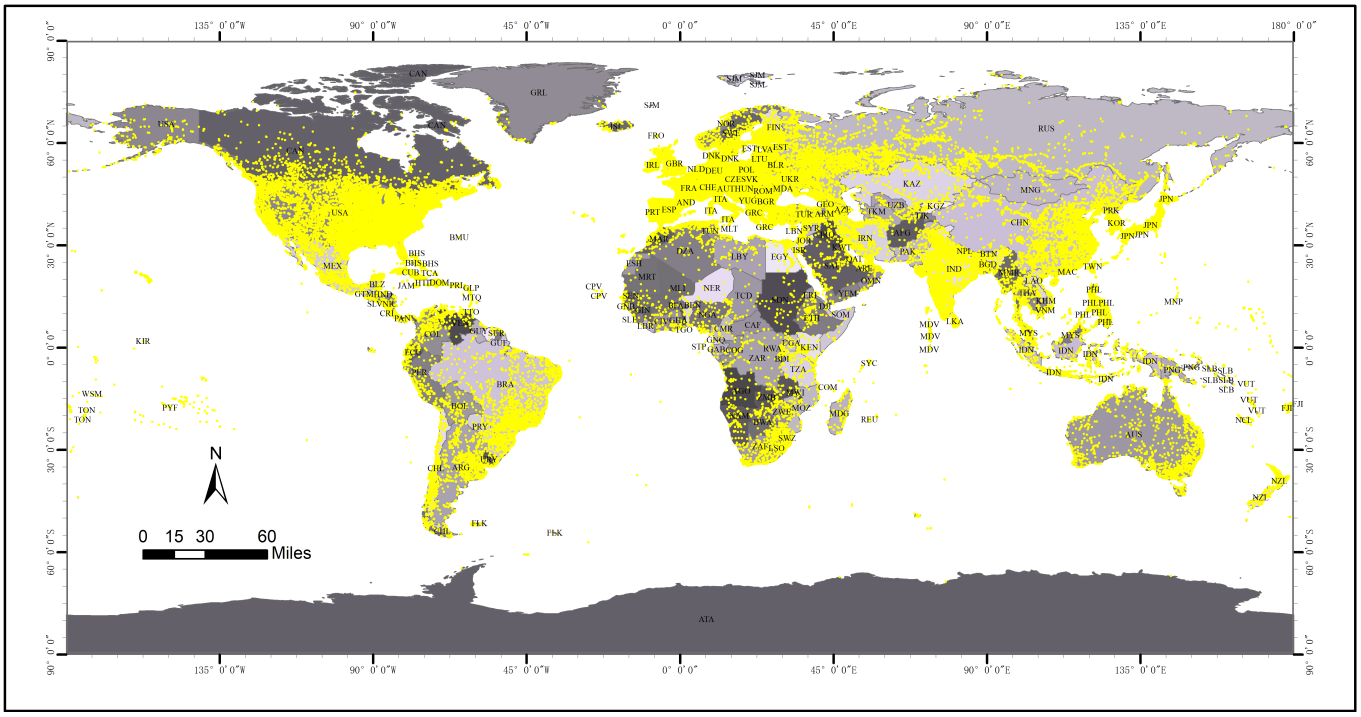



Fig. 4. The distribution of scene images in Million-AID. To clearly express the image location and its distribution, more than a half of the total scene images are taken as the representative instances, and visually displayed using their geographical locations. The location of a scene image is represented by the central geographical coordinates of the image block.

degree of model depth, 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 a unified CNN library using PyTorch [185] for model training and testing.

TABLE I
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.

### B. Multi-class scene classification

*1) Implementation detail:* In [35] and [34], the aerial image features were directly extracted from the CNN models 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 (SGD) 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.

*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 [34], [35]. 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. A CM can present the classification performance of a model on each class. Each row of the CM represents the actual instances in a predicted class while each column reveals the predicted instances in an actual class. The CM make 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.

*3) Experimental results:* **Baseline results:** Table II illustrates the performance of scene classification using different CNN models. From the classification results, we can see that all the models achieve reasonable classification performance. Generally, VGG16, GoogleNet, ResNet101, DenseNet121, and DenseNet169 present significantly better classification results when compared to AlexNet. Note the fact that AlexNet is

TABLE II
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 |

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 [186].

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 [187] 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. It is worth noting that DenseNet169 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%, which is far from being satisfactory. 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 Fig. 5, 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 II. 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 [34], [188]–[190] in which each scene category share the same number of images. This implies that significant attention should be paid to the property of category imbalance when developing scene classification algorithms.

**Confusion matrices:** By further investigating the confusion matrices (as shown in Fig. 5) 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 II. 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.

**Comparison with Existing Benchmarks:** Many datasets have been established to promote the advancement of scene classification as detailed in [179]. We compare the classification results of Million-AID with those of popular aerial

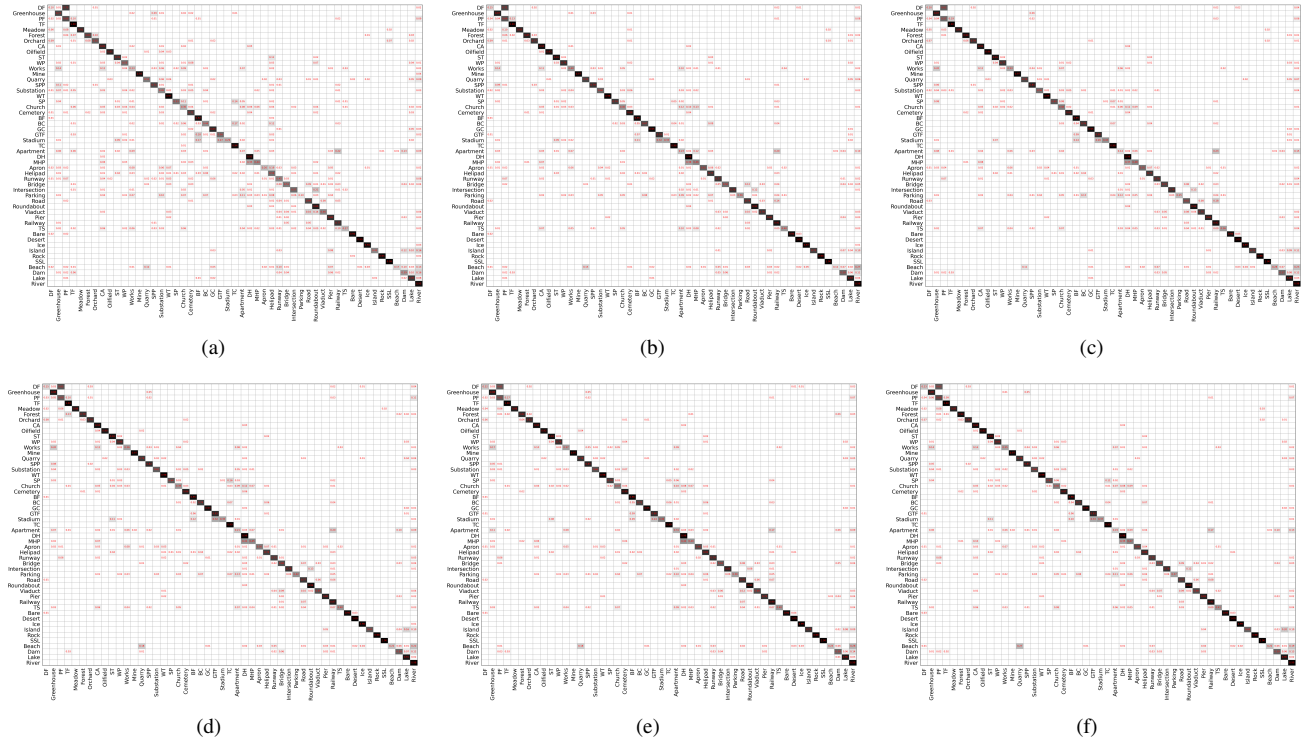


Fig. 5. Confusion matrix obtained by (a) AlexNet, (b) VGG16, (c) GoogleNet, (d) ResNet101, (e) DenseNet121, and (f) DenseNet169 on Million-AID dataset. Zoom for detail.

scene classification datasets, *i.e.* NWPU-RESISC45 [34] and AID [35], considering their high quality and wide application. Table III describes the overall accuracy of different CNN models evaluated on different datasets. The results show that our implemented CNN models (indicated with * symbols) 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 that our proposed Million-AID is more challenging than the compared datasets. 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.

TABLE III
OA COMPARISON AMONG DIFFERENT DATASETS (%)

| Dataset | AlexNet | VGG16 | GoogleNet |
|---|---|---|---|
| AID [35] | 86.86 | 86.59 | 83.44 |
| AID* | 88.79 | 93.72 | 92.24 |
| NWPU-RESISC45 [34] | 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 (details will be introduced in Section V) using our implemented CNN framework, so does the *NWPU-RESISC45**. The standard deviations are omitted since their negligible influence on the final result.

### C. Multi-label Scene Classification

*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 a *sigmoid* function and generate confidences for each of the semantic categories similar to [191]. 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 network as shown in Figure 2. 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 network is regarded as a single label. In this configuration, the existing classification algorithms can be extended effortlessly for multi-label scene classification.

*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 $\tau$. 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 [192]–[194], 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

TABLE IV
PERFORMANCE OF MULTI-LABEL SCENE CLASSIFICATION WITH DIFFERENT CNN MODELS (%)

| Model | $\tau = 0.5$ | | | | | | | $\tau = 0.75$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CP | CR | CF1 | OP | OR | OF1 | mAP | CP | CR | CF1 | OP | OR | OF1 | mAP |
| AlexNet | 71.45 | 48.19 | 57.56 | 76.19 | 62.84 | 68.87 | 44.20 | 78.89 | 38.51 | 51.76 | 85.65 | 53.03 | 65.50 | 36.52 |
| VGG16 | 82.26 | 62.20 | 70.84 | 86.98 | 75.31 | 80.72 | 60.05 | 84.61 | 54.29 | 66.14 | 91.70 | 69.37 | 78.99 | 53.10 |
| GoogleNet | 51.79 | 33.99 | 41.04 | 88.50 | 59.47 | 71.14 | 32.96 | 50.99 | 23.76 | 32.42 | 94.90 | 47.02 | 62.89 | 23.36 |
| ResNet101 | 79.38 | 59.67 | 68.13 | 88.74 | 77.31 | 82.63 | 57.78 | 76.83 | 51.56 | 61.71 | 93.05 | 70.93 | 80.50 | 50.43 |
| DenseNet121 | 79.09 | 56.21 | 65.71 | 89.74 | 75.10 | 81.77 | 54.63 | 76.36 | 47.75 | 58.76 | 94.20 | 67.72 | 78.79 | 46.86 |
| DenseNet169 | 78.54 | 61.92 | 69.24 | 88.50 | 78.55 | 83.23 | 59.75 | 78.52 | 55.10 | 64.76 | 92.66 | 73.10 | 81.72 | 53.66 |

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.

*3) Experimental Results:* Quantitative results of multi-label scene classification on Million-AID are reported in Table IV. It can be seen that the performance varies widely between different methods. 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, which consists of the most layers among the other networks. Nevertheless, DenseNet169 achieves the best performance on overall metrics, where the OP, OR, and OF1 are 88.50%, 78.55%, 83.23%, respectively. Fig. 6(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. However, when the scene images are assigned with the hierarchically multiple labels, the issue of data imbalance becomes more prominent, which brings the problem of "catastrophic forgetting" [195], [196]. As a result, the networks show weak performance on categories like *stadium*, *apartment*, *Beach*,and *Parking*. Moreover, DenseNet169 shows insufficient ability in recognizing *grassland*, *meadow*, *oilfield quarry*, *ground track field*, *viaduct*, and *apron*. Consequently, DenseNet169 reports slightly worse performance on mAP than that of VGG16 for $\tau = 0.5$.

When increasing $\tau$ to be 0.75, the OP of all models gain significant improvement. This makes sense because 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 mAP metrics decline correspondingly in comparison with the result of which $\tau = 0.5$. It is worth noting that DenseNet169 achieves the best performance on OF1 (81.72%) and mAP (53.66%), indicating its excellent ability in distinguishing different semantic categories with high reliability.

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 44.20%, which are 16.25% and 11.24% higher than those of GoogleNet, respectively. Even AlexNet achieves 2.27% lower OF1 than that of GoogleNet, the former model shows superiority on OR. Apparently, VGG16 reports better comprehensive performance when compared with the deeper networks such as GoogleNet, RestNet, 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 I. With this superiority, the shallow networks are able to learn the relationship of scene categories at different levels of hierarchy. As a comparison, GoogleNet has more convolutional layers than that of AlexNet and VGG, but it consists of only 6.8M parameters. The experimental results show that GoogleNet provides the worst performance 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. Fig. 6(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 IV. With the above analysis, it is natural to argue that both the parameter scale and depth of convolutional layers are crucial for recognizing the scene categories with hierarchical relationships. Intuitively, more parameters and 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 actually 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

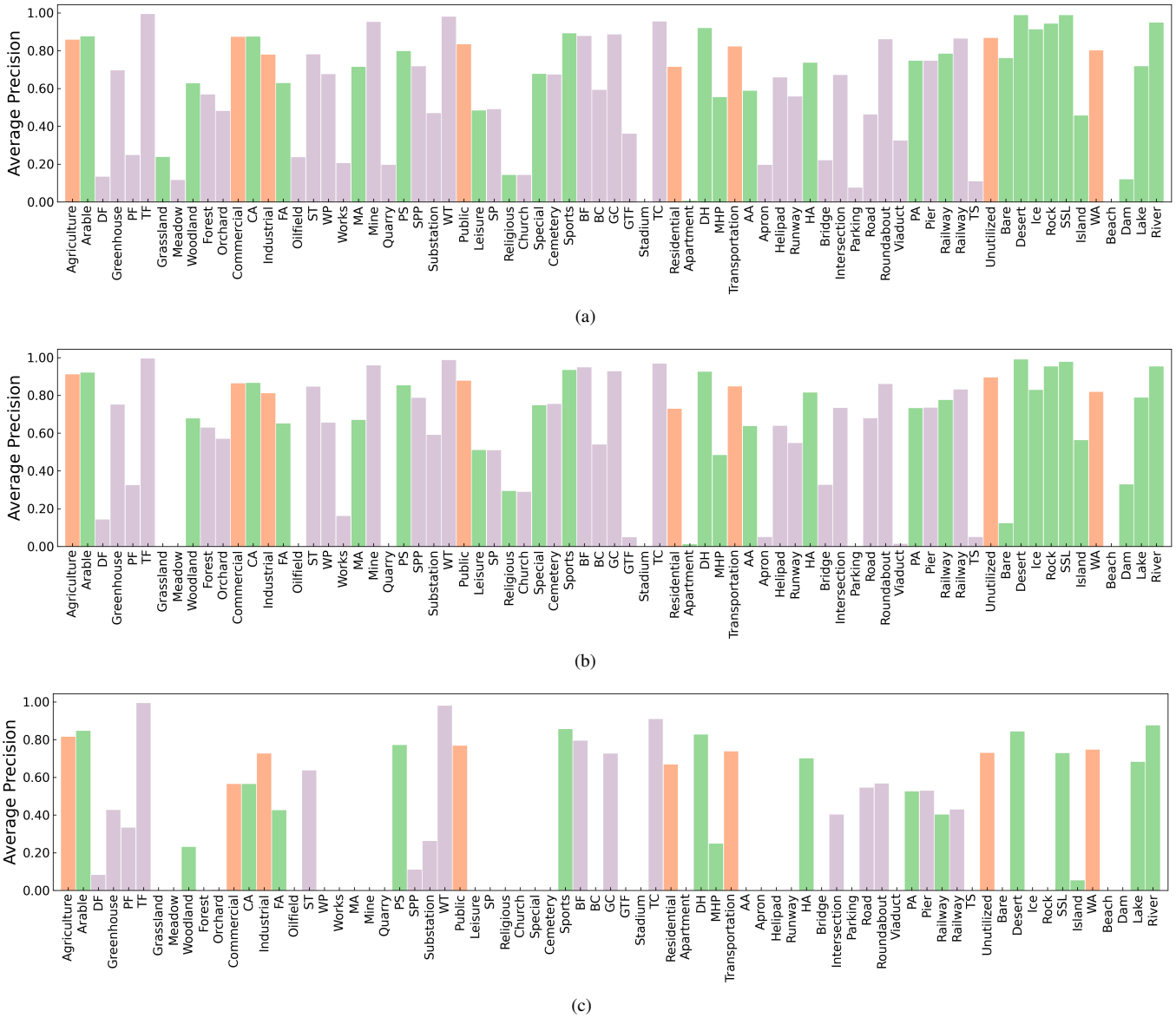



Fig. 6. Average precision results of (a) VGG16, (b) DenseNet169, and (c) GoogleNet on Million-AID dataset. The orange bars indicates predicted scene labels belonging to the first-level categories, the green bars the second-level categories, and the plum bars the third-level categories.

remain to be further explored.

## V. Knowledge Transfer By 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.

### A. Fine-tuning Pre-trained Networks for Scene Classification

*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 [34], [149], [197]–[199], 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 [48], 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 the proposed 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 [35] and NWPU-RESISC45 [34]. Similar to the dataset partition scheme in [34], [35], 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 IV 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 IV. 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 IV.

*2) Experimental Results:* Tables V and VI 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 [34], [35], [37]). Compared with the models trained from scratch, the models pre-trained on ImageNet and Million-AID can significantly improve the classification performance. Figures 7 and 8 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.

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 Fig. 7 (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 Fig. 8 (b) and (c). Figures 9 and 10 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 and then achieve accurate scene classification. 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 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 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.

### B. Hierarchical Multi-task Learning for Semantic Parsing

*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 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 much the knowledge contained in Million-AID can be transferred to boost the pixel-level semantic parsing of

TABLE V
CLASSIFICATION ACCURACY (%) ON AID DATASET USING DIFFERENT TRAINING SCHEMES

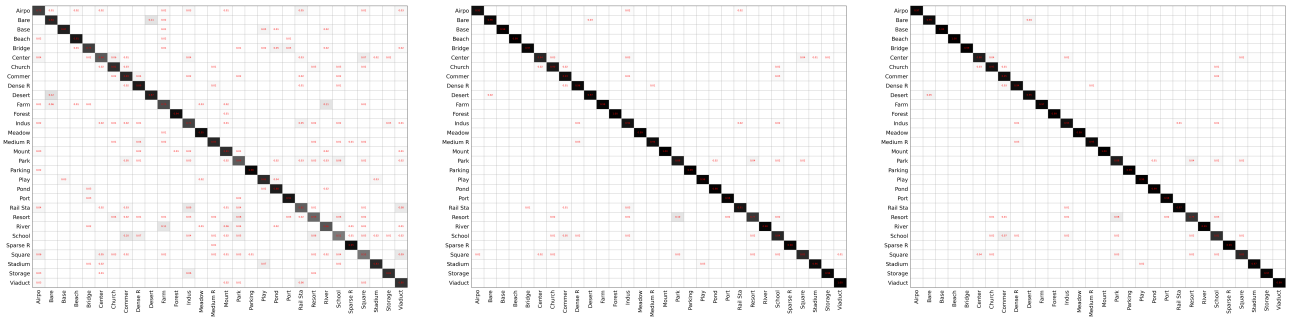
| Metric | Pretrain dataset | AlexNet | VGG16 | GoogleNet | ResNet101 | DenseNet121 | DenseNet169 |
|---|---|---|---|---|---|---|---|
| OA | W/O | $33.47 \pm 2.15$ | $72.18 \pm 0.49$ | $79.05 \pm 0.89$ | $49.46 \pm 2.07$ | $58.02 \pm 0.74$ | $59.16 \pm 0.52$ |
| | ImageNet | $88.79 \pm 0.40$ | $93.72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

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

TABLE VI
CLASSIFICATION ACCURACY (%) ON NPWU-RESISC45 DATASET USING DIFFERENT TRAINING SCHEMES

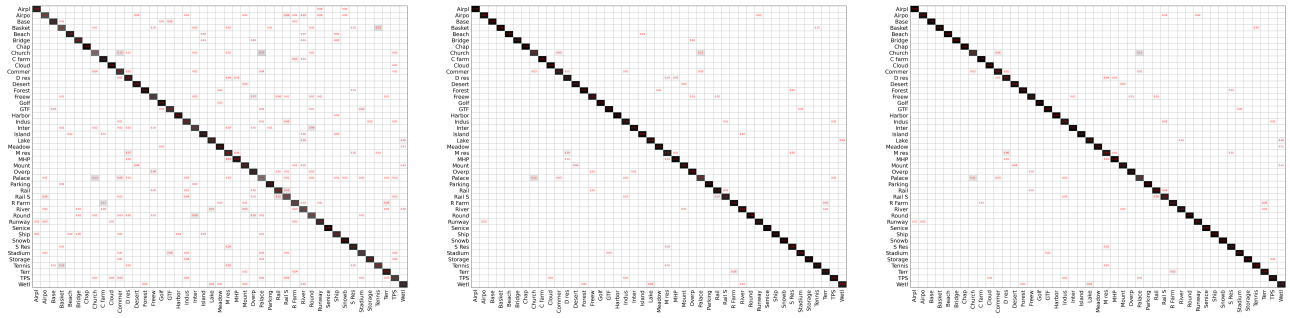
| Metric | Pretrain dataset | AlexNet | VGG16 | GoogleNet | ResNet101 | DenseNet121 | DenseNet169 |
|---|---|---|---|---|---|---|---|
| OA | W/O | $37.92 \pm 0.70$ | $73.19 \pm 0.44$ | $81.77 \pm 0.56$ | $58.82 \pm 0.74$ | $63.35 \pm 0.34$ | $64.51 \pm 0.47$ |
| | ImageNet | $87.19 \pm 0.26$ | $92.76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

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



(a) (b) (c)

Fig. 7. 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)

Fig. 8. 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.

aerial images. To this end, the GID [149], 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 parsing. Generally, the presented framework consists four components, *e.g.*, hierarchical scene representation, multi-


Fig. 9. 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.


Fig. 10. 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.

task scene classification, hierarchical semantic fusion, and pixel-level semantics integration, as shown in Figure 11.

**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 [200], [201]. 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 $1 \times 1$ 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

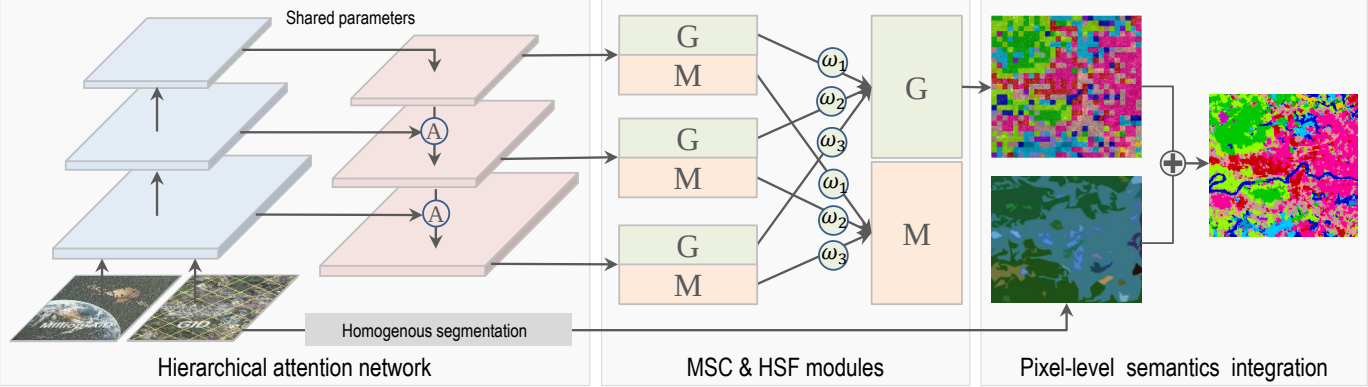



Fig. 11. The framework of hierarchical multi-task learning for pixel-level semantic parsing. Each pair of framed *G* and *M* denotes the multi-task classification branches for GID and Million-AID, respectively. The circled plus sign indicates the majority voting process and the circled *A* the attention module.

output the final attention feature (**AF**). The whole process is illustrated in Figure 12. 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 11) that incorporate semantic and visual information for scene representation.

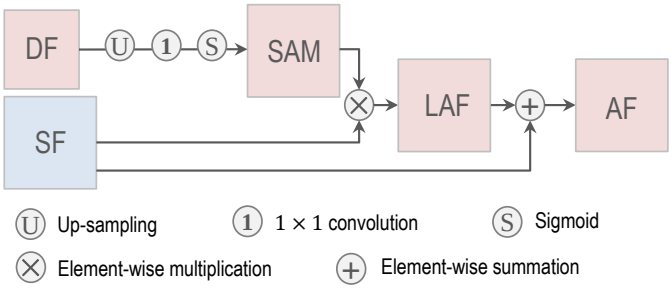



Fig. 12. The attention module in the hierarchical attention network.

**Multi-task scene classification (MSC):** With the hierarchical attention features, a semantic scene can be represented using the features of multiple scales. As shown in Figure 11, 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 [202], [203]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 [149], 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 $s$. $\mu_g$ and $\mu_m$ (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 transferring knowledge contained in Million-AID. Thus, the semantic classification on GID is regarded as the main task while the scene classification on Million-AID serves as the auxiliary task [204] 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:

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

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:

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

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

**Pixel-level semantics integration (PSI):** Through above procedures, we are able to achieve tile-level interpretation and generate the semantic grid map of a large-size aerial image. Here, each semantic grid corresponds a tile-level classification result. For more accurate result with pixel-level semantics, the boundary information of ground objects in the high resolution aerial images appears the great importance. We therefore employ object-based segmentation and majority voting strategy to generate pixel-level semantic parsing result by referencing [149]. Specifically, the selective search algorithm [205] is conducted on the raw aerial image and generate 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:

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

where $|r_n|$ denotes the number of pixels enclosed in $r$ and labeled as class $n$ in the semantic grid map. By integrating multi-scale contextual information with tile-level classification, the whole pipeline falls into a hybrid classification framework [149]. Nevertheless, our method focuses more on advancing the pixel-level semantic parsing via improving the tile-level semantic interpretation and transferring aerial scene knowledge from Million-AID.

*2) Implementation detail:* ResNet50 [183] 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 empirically 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 30k multi-scale image patches, 10 pixel-level annotated Gaofen-2 images, and 15 challenging semantic categories, is employed for model training and testing. Considering that the training samples of GID are highly overlapped, image flipping and rotation (90°, 180°, and 270°) are conducted for data augmentation. 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 IV except the number of total iteration set as 30 epochs. The OA, Kappa, and mean-Intersection-over-Union (mIoU) [153] are employed for performance evaluation.

*3) Experimental Results:* **Ablation study:** For multi-task learning, the weights of different tasks can make a big influence on the classification performance. Table VII 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 ($T_g$). Fig. 13 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 $\mu_g$ is smaller than $\mu_m$. As $\mu_g$ 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 (Fig. 13(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 Fig. 13(b). When $\mu_g : \mu_m$ changes to $0.5 : 0.5$, both $T_g$ and $T_M$ can be well optimized (Fig. 13(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 Fig. 13(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.

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 VIII 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 VIII, 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

TABLE VII
WEIGHTS INFLUENCE ON DIFFERENT TASKS

| $\mu_g$ | $\mu_m$ | 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 |

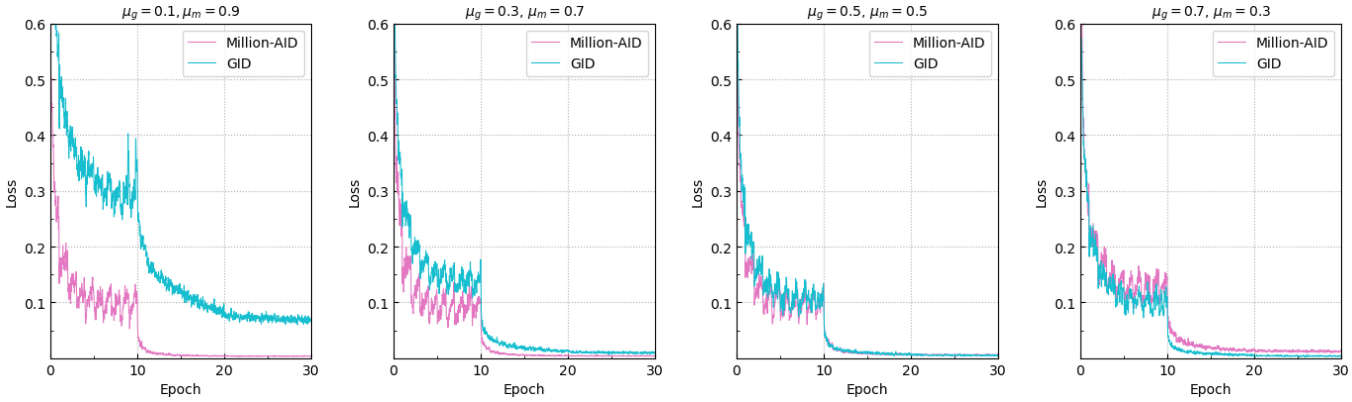



Fig. 13. Training loss with respect to different setups of learning weights. The models converge successfully for both classification tasks with different learning weights, which confirms the validity of the learned classification model.

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.

TABLE VIII
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** |

Fig. 14 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 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 VIII. 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 [149], of which 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 pre-trained on the large-scale classification set of GID (PT-GID). For comprehensive comparison, the presented method was also compared with some image segmentation models, such as the U-Net [206], PSPNet [207], DeepLab V3+ and its variations [208].

The quantitative results of different methods are summarized in Table IX. 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.

Fig. 15 provides the intuitive visualization of the pixel-

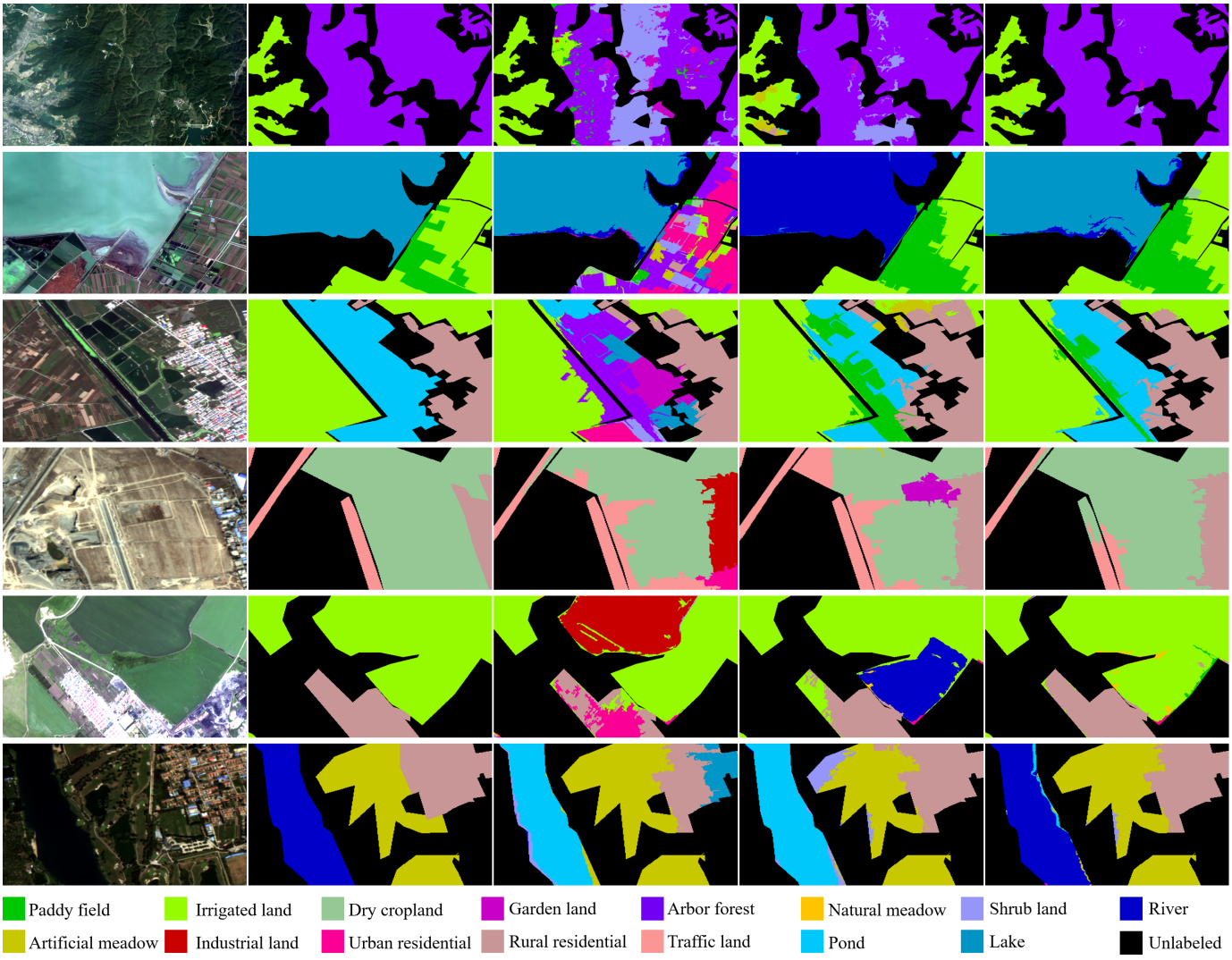



Fig. 14. Qualitative comparisons of different classification schemes. Images in the first to fifth columns indicate the original image, ground truth annotations, classification maps of baseline, MSC, and the full implementation of our method, respectively.

TABLE IX
CLASSIFICATION PERFORMANCE COMPARISON WITH DIFFERENT METHODS ON GID

| Methods | Kappa | OA (%) |
|---|---|---|
| MLC + SGDL [149] | 0.145 | 23.61 |
| SVM + SGDL [149] | 0.148 | 23.92 |
| MLP + SGDL [149] | 0.199 | 30.57 |
| RF + SGDL [149] | 0.237 | 33.70 |
| DeepLab V3+ Mobilenet [208] | 0.357 | 54.64 |
| U-Net [206], [208] | 0.439 | 56.59 |
| PSPNet [207], [208] | 0.458 | 60.73 |
| DeepLab V3+ [208] | 0.478 | 62.19 |
| DeepLab V3+ MLF [208] | 0.598 | 69.16 |
| PT-GID [149] | 0.605 | 70.04 |
| ***Ours*** | **0.673** | **73.03** |

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 IX.

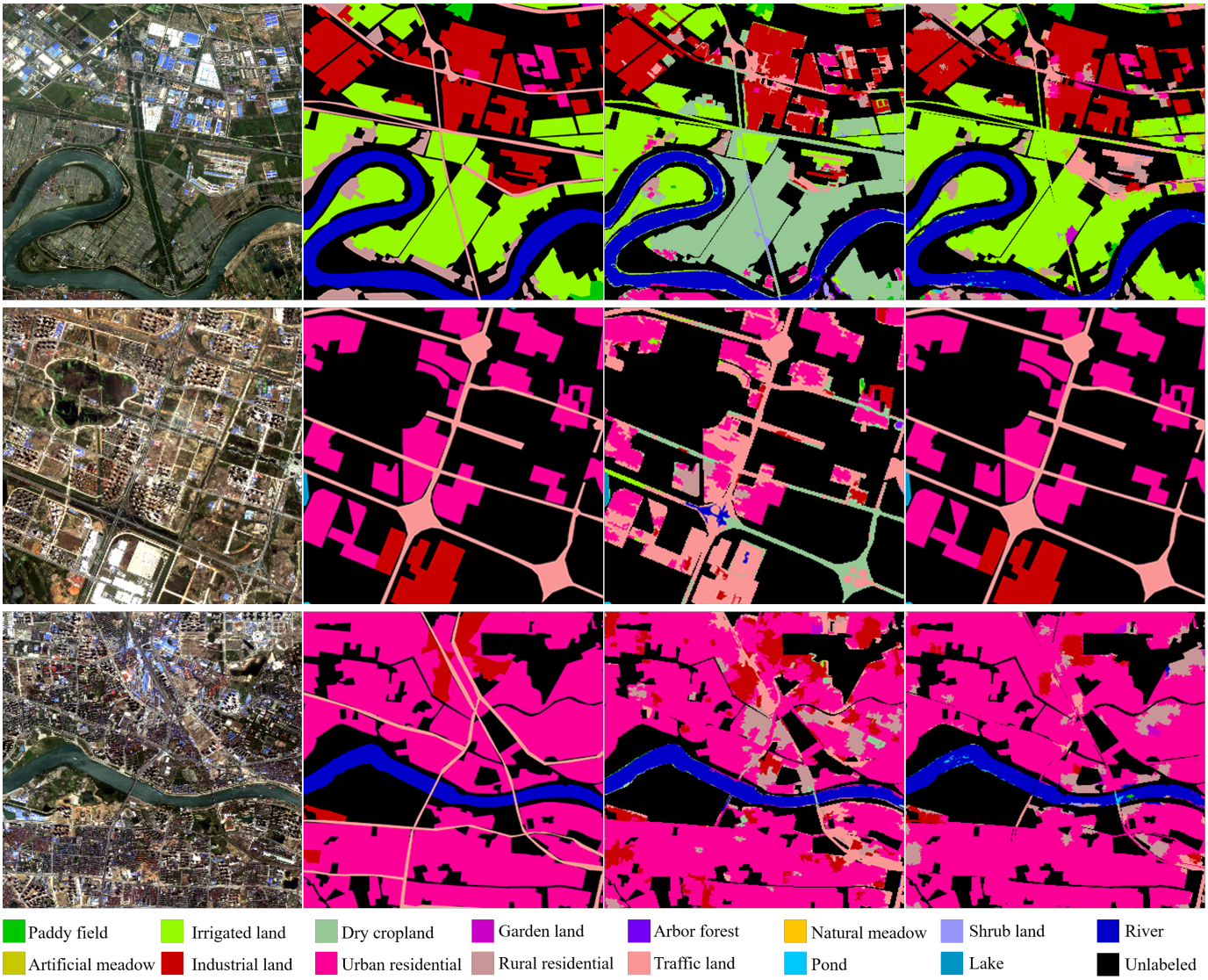



Fig. 15. Visualization of the land cover classification results on the fine classification set of GID. Images in the first to fourth columns indicate the original image, ground truth annotations, classification maps of PT-GID [149], and classification maps of Our method, respectively.

## VI. Conclusions

The progress of remote sensor technology has significantly improved human's ability to observe the ground features. However, the development of interpretation algorithms is lack of support from large-scale annotated datasets and facing challenges in the changing requirements. In this paper, we conducted a review by revisiting the development of aerial image interpretation. It is shown that the interpretation prototype of aerial images develop with the improvement of image quality and has experienced the stages from per-pixel image classification, object-based image analysis, and tile-level image understanding. Then, we introduced the large-scale large-scale dataset, i.e., Million-AID, to be released publicly available for aerial scene recognition. Intensive experiments with classical CNN frameworks indicate that Million-AID is 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 than those on ImageNet for tile-level semantic interpretation, which demonstrates the strong generalization ability of Million-AID and reveals the essential content knowledge difference between aerial and natural images, to some extent. Besides, we designed a hierarchical multi-task learning framework and achieved the state-of-the-art result for pixel-level semantic classification, 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 be extended to other related tasks, such as object detection and semantic segmentation, to further explore the transferability of large-scale aerial scene images. We hope that is work can enhance the development of content interpretation algorithms in the field of remote sensing images.

## REFERENCES


[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, 2016.

[2] T.-Z. Xiang, G.-S. Xia, and L. Zhang, "Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 3, pp. 29–63.

[3] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, p. 111402, 2020.

[4] J. Jung, M. Maeda, A. Chang, M. Bhandari, A. Ashapure, and J. Landivar-Bowles, "The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems," *Curr. Opin. Biotechnol.*, vol. 70, pp. 15–22, 2021.

[5] C. Qiu, M. Schmitt, C. Geiß, T.-H. K. Chen, and X. X. Zhu, "A framework for large-scale mapping of human settlement extent from sentinel-2 images via fully convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 152–170, 2020.

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

[7] X. Chen, Y. Xu, J. Yang, Z. Wu, and H. Zhu, "Remote sensing of urban thermal environments within local climate zones: A case study of two high-density subtropical chinese cities," *Urban Clim.*, vol. 31, p. 100568, 2020.

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

[9] M. Kucharczyk and C. H. Hugenholtz, "Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities," *Remote Sens. Environ.*, vol. 264, p. 112577, 2021.

[10] L. Moya, C. Geiß, M. Hashimoto, E. Mas, S. Koshimura, and G. Strunz, "Disaster intensity-based selection of training samples for remote sensing building damage classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–17, 2021.

[11] D. Li, M. Wang, Z. Dong, X. Shen, and L. Shi, "Earth observation brain (eob): An intelligent earth observation system," *Geo-spatial Information Science*, vol. 20, no. 2, pp. 134–140, 2017.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification."

[13] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[14] W. Yang, X. Yin, and G. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, 2015.

[15] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, 2015.

[16] F. Wang, "Fuzzy supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 2, pp. 194–201, 1990.

[17] L. Bruzzone, D. F. Prieto, and S. B. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1350–1359, 1999.

[18] C. Wu and A. T. Murray, "Estimating impervious surface distribution by spectral mixture analysis," *Remote Sens. Environ.*, vol. 84, no. 4, pp. 493–505, 2003.

[19] C. H. Chen and P.-G. Peter Ho, "Statistical pattern recognition in remote sensing," *Pattern Recognit.*, vol. 41, no. 9, pp. 2731–2741, 2008.

[20] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *Int. J. Remote Sens.*, vol. 28, no. 5, pp. 823–870, 2007.

[21] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 3, pp. 606–617, 2011.

[22] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

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

[24] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[25] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man. Cybern. Syst.*, no. 6, pp. 610–621, 1973.

[26] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, 1996.

[27] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 382–403, 2010.

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

[29] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

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

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

[32] J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 254–283, 2010.

[33] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[34] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[35] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.

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

[37] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017.

[38] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, 2017.

[39] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, 2018.

[40] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Trans. Big Data*, pp. 1–1, 2019.

[41] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, 2019.

[42] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, 2019.

[43] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined cnn and gcn for remote sensing scene classification," *IEEE J. Sel. Top. Appl. Earth Obs .Remote Sens.*, vol. 13, pp. 4325–4338, 2020.

[44] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4911–4926, 2020.

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

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

[47] Q. Bi, K. Qin, H. Zhang, and G.-S. Xia, "Local semantic enhanced convnet for aerial scene recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 6498–6511, 2021.

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

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

[50] J. Van Genderen, B. Lock, and P. Vass, "Remote sensing: Statistical testing of thematic map accuracy," *Remote Sens. Environ.*, vol. 7, no. 1, pp. 3–14, 1978.

[51] P. Curran and H. Williamson, "Sample size for ground and remotely sensed data," *Remote Sens. Environ.*, vol. 20, no. 1, pp. 31–41, 1986.

[52] R. Khatami, G. Mountrakis, and S. V. Stehman, "Mapping per-pixel predicted accuracy of classified remote sensing images," *Remote Sens. Environ.*, vol. 191, pp. 156–167, 2017.

[53] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, 2010.

[54] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, 2013.

[55] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, 2011.

[56] J. Zhao, Y. Zhong, H. Shu, and L. Zhang, "High-resolution image classification integrating spectral-spatial-location cues by conditional random fields," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4033–4045, 2016.

[57] J. Settle and S. Briggs, "Fast maximum likelihood classification of remotely-sensed imagery," *Int. J. Remote Sens.*, vol. 8, no. 5, pp. 723–734, 1987.

[58] J. Ediriwickrema and S. Khorram, "Hierarchical maximum-likelihood classification for improved accuracies," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 4, pp. 810–816, 1997.

[59] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1790–1802, 2019.

[60] M. E. Hodgson, "Reducing the computational requirements of the minimum-distance classifier," *Remote Sens. Environ.*, vol. 25, no. 1, pp. 117–128, 1988.

[61] M. Espínola, J. A. Piedra-Fernández, R. Ayala, L. Iribarne, and J. Z. Wang, "Contextual and hierarchical classification of satellite images based on cellular automata," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 795–809, 2014.

[62] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based $k$-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, 2010.

[63] B. Tu, S. Huang, L. Fang, G. Zhang, J. Wang, and B. Zheng, "Hyperspectral image classification via weighted joint nearest neighbor and sparse representation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 11, pp. 4063–4075, 2018.

[64] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, 1997.

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

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

[67] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, 2018.

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

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

[70] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Adv. Neural Inf. Process. Syst.*, 2007, pp. 801–808.

[71] J. Liu, Z. Wu, Z. Wei, L. Xiao, and L. Sun, "Spatial-spectral kernel sparse representation for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 6, no. 6, pp. 2462–2471, 2013.

[72] Z. Feng, M. Wang, S. Yang, Z. Liu, L. Liu, B. Wu, and H. Li, "Superpixel tensor sparse coding for structural hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 4, pp. 1632–1639.

[73] J. Fan, T. Chen, and S. Lu, "Superpixel guided deep-sparse-representation learning for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3163–3173, 2017.

[74] D. Li, Q. Wang, and F. Kong, "Adaptive kernel sparse representation based on multiple feature learning for hyperspectral image classification," *Neurocomputing*, 2020.

[75] T. Kavzoglu and P. Mather, "The use of backpropagating artificial neural networks in land cover classification," *Int. J. Remote Sens.*, vol. 24, no. 23, pp. 4907–4938, 2003.

[76] Y. Shao and R. S. Lunetta, "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.*, vol. 70, pp. 78–87, 2012.

[77] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6547–6565, 2017.

[78] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[79] U. Maulik and D. Chakraborty, "Remote sensing image classification: A survey of support-vector-machine-based advanced techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, 2017.

[80] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognit.*, p. 107298, 2020.

[81] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, "A review of remote sensing image classification techniques: The role of spatio-contextual information," *Eur. J. Remote Sens.*, vol. 47, no. 1, pp. 389–411, 2014.

[82] W. Liu and E. Y. Wu, "Comparison of non-linear mixture models: sub-pixel classification," *Remote Sens. Environ.*, vol. 94, no. 2, pp. 145–154, 2005.

[83] F. Bovolo, L. Bruzzone, and L. Carlin, "A novel technique for subpixel image classification based on support vector machine," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2983–2999, 2010.

[84] Q. Wang, C. Zhang, and P. M. Atkinson, "Sub-pixel mapping with point constraints," *Remote Sens. Environ.*, vol. 244, p. 111817, 2020.

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

[86] A. K. Shackelford and C. H. Davis, "A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1920–1932, 2003.

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

[88] K. C. Mertens, L. P. Verbeke, T. Westra, and R. R. De Wulf, "Sub-pixel mapping and sub-pixel sharpening using neural network predicted wavelet coefficients," *Remote Sens. Environ.*, vol. 91, no. 2, pp. 225–236, 2004.

[89] X. Li, F. Ling, Y. Du, Q. Feng, and Y. Zhang, "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.*, vol. 93, pp. 76–87, 2014.

[90] D. He, Y. Zhong, X. Wang, and L. Zhang, "Deep convolutional neural network framework for subpixel mapping," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–22, 2020.

[91] R. Fernandes, R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien, and Y. Du, "Approaches to fractional land cover and continuous field mapping: A comparative assessment over the boreas study region," *Remote Sens. Environ.*, vol. 89, no. 2, pp. 234–251, 2004.

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

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

[94] B. Somers, G. P. Asner, L. Tits, and P. Coppin, "Endmember variability in spectral mixture analysis: A review," *Remote Sens. Environ.*, vol. 115, no. 7, pp. 1603–1616, 2011.

[95] J. Yu, B. Wang, Y. Lin, F. Li, and J. Cai, "A novel inequality-constrained weighted linear mixture model for endmember variability," *Remote Sens. Environ.*, vol. 257, p. 112359, 2021.

[96] F. Xu and B. Somers, "Unmixing-based sentinel-2 downscaling for urban land cover mapping," *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 133–154, 2021.

[97] A. Bateson and B. Curtiss, "A method for manual endmember selection and spectral unmixing," *Remote Sens. Environ.*, vol. 55, no. 3, pp. 229–243, 1996.

[98] C. Small, "The landsat etm+ spectral mixing space," *Remote Sens. Environ.*, vol. 93, no. 1-2, pp. 1–17, 2004.

[99] S. Ozkan, B. Kaya, and G. B. Akar, "EndNet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 482–496, 2019.

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

[101] P. M. Atkinson, "Mapping sub-pixel boundaries from remotely sensed images," in *Innovations in GIS*, 1997, pp. 184–202.

[102] Q. Wang and P. M. Atkinson, "The effect of the point spread function on sub-pixel mapping," *Remote Sens. Environ.*, vol. 193, pp. 127–137, 2017.

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

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

[105] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 309–322, 2021.

[106] G. Hay, D. Marceau, P. Dube, and A. Bouchard, "A multiscale framework for landscape analysis: object-specific analysis and upscaling," *Landsc Ecol.*, vol. 16, no. 6, pp. 471–490, 2001.

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

[108] H.-D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, "Color image segmentation: advances and prospects," *Pattern Recognit.*, vol. 34, no. 12, pp. 2259–2281, 2001.

[109] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4928–4943, 2011.

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

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

[112] D. Amitrano, F. Cecinati, G. Di Martino, A. Iodice, P.-P. Mathieu, D. Riccio, and G. Ruello, "Feature extraction from multitemporal sar images using selforganizing map clustering and object-based image analysis," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 5, pp. 1556–1570, 2018.

[113] B. Kaur and A. Garg, "Mathematical morphological edge detection for remote sensing images," in *Proc. Int. Conf. Electron. Comput. Technol.*, vol. 5, 2011, pp. 324–327.

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

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

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

[117] T. Su, "Scale-variable region-merging for high resolution remote sensing image segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 319–334, 2019.

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

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

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

[121] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, 2017.

[122] B. Kumar, O. Dikshit, A. Gupta, and M. K. Singh, "Feature extraction for hyperspectral image classification: a review," *Int. J. Remote Sens.*, vol. 41, no. 16, pp. 6248–6287, 2020.

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

[124] C. Zhang, P. Yue, D. Tapete, B. Shangguan, M. Wang, and Z. Wu, "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.*, vol. 88, p. 102086, 2020.

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

[126] L. Ma, L. Cheng, M. Li, Y. Liu, and X. Ma, "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.*, vol. 102, pp. 14–27, 2015.

[127] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, 2016.

[128] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, 2019.

[129] N. Audebert, B. Le Saux, and S. Lefèvre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, 2019.

[130] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, 2021.

[131] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[132] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[133] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, 2019.

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

[135] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "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.*, vol. 12, no. 8, pp. 2612–2626, 2019.

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

[137] M. Zhang, W. Li, and Q. Du, "Diverse region-based cnn for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, 2018.

[138] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[139] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "Deeplab-based spatial feature extraction for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 251–255, 2018.

[140] Y. Chen, Y. Wang, Y. Gu, X. He, P. Ghamisi, and X. Jia, "Deep learning ensemble for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 6, pp. 1882–1897, 2019.

[141] X. He and Y. Chen, "Transferring cnn ensemble for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 876–880, 2020.

[142] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, 2017.

[143] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarablaka, G. Moser, A. De Giorgi, L. Fang, Y. Chen, M. Chi *et al.*, "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.*, vol. 6, no. 3, pp. 10–43, 2018.

[144] Q. Gao, S. Lim, and X. Jia, "Spectral–spatial hyperspectral image classification using a multiscale conservative smoothing scheme and adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7718–7730, 2019.

[145] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, 2019.

[146] J. Feng, J. Chen, L. Liu, X. Cao, X. Zhang, L. Jiao, and T. Yu, "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.*, vol. 12, no. 4, pp. 1299–1313, 2019.

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

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

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

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

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

[152] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.

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

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

[155] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, 2015.

[156] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, 2018.

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

[158] Q. Bi, H. Zhang, and K. Qin, "Multi-scale stacking attention pooling for remote sensing scene classification," *Neurocomputing*, vol. 436, pp. 147–161, 2021.

[159] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, 2019.

[160] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, 2019.

[161] Q. Bi, K. Qin, H. Zhang, Z. Li, and K. Xu, "Radc-net: A residual attention based convolution network for aerial scene classification," *Neurocomputing*, vol. 377, pp. 345–359, 2020.

[162] L. Fu, D. Zhang, and Q. Ye, "Recurrent thrifty attention network for remote sensing scene recognition," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–12, 2020.

[163] Z. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-promoting deep structural metric learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 371–390, 2018.

[164] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, 2020.

[165] Y. Wang, L. Zhang, X. Tong, F. Nie, H. Huang, and J. Mei, "LRAGE: Learning latent relationships with adaptive graph embedding for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 621–634, 2018.

[166] N. Khan, U. Chaudhuri, B. Banerjee, and S. Chaudhuri, "Graph convolutional network for multi-label vhr remote sensing scene recognition," *Neurocomputing*, vol. 357, pp. 36–46, 2019.

[167] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, 2021.

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

[169] C. Peng, Y. Li, L. Jiao, and R. Shang, "Efficient convolutional neural architecture search for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6092–6105, 2021.

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

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

[172] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rs-metanet: Deep metametric learning for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6983–6994, 2021.

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

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

[175] E. Othman, Y. Bazi, F. Melgani, H. Alhichri, N. Alajlan, and M. Zuair, "Domain adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4441–4456, 2017.

[176] X. Lu, T. Gong, and X. Zheng, "Multisource compensation network for remote sensing cross-domain scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2504–2515, 2020.

[177] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7920–7930, 2020.

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

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

[180] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[181] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

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

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

[185] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035.

[186] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

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

[188] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, 2015.

[189] Q. Zhu, Y. Zhong, B. Zhao, G. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, 2016.

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

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

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

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

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

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

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

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

[198] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

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

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

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

[202] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[203] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, pp. 1–20, 2021.

[204] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

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

[206] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assis.t Interv.* Springer, 2015, pp. 234–241.

[207] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[208] Y. Ren, X. Zhang, Y. Ma, Q. Yang, C. Wang, H. Liu, and Q. Qi, "Full convolutional neural network based on multi-scale feature fusion for the class imbalance remote sensing image classification," *Remote Sens.*, vol. 12, no. 21, p. 3547, 2020.