













MEET: A Million-Scale Dataset for Fine-Grained Geospatial Scene Classification With Zoom-Free Remote Sensing Imagery

Yansheng Li , Senior Member, IEEE, Yuning Wu , Gong Cheng , Member, IEEE, Chao Tao , Bo Dang , Yu Wang , Jiahao Zhang , Chuge Zhang , Yiting Liu , Xu Tang , Senior Member, IEEE, Jiayi Ma , Senior Member, IEEE, and Yongjun Zhang , Member, IEEE

Abstract—Accurate fine-grained geospatial scene classification using remote sensing imagery is essential for a wide range of applications. However, existing approaches often rely on manually zooming remote sensing images at different scales to create typical scene samples. This approach fails to adequately support the fixed-resolution image interpretation requirements in real-world scenarios. To address this limitation, we introduce the million-scale fine-grained geospatial scene classification dataset (MEET), which contains over 1.03 million zoom-free remote sensing scene samples, manually annotated into 80 fine-grained categories. In MEET, each scene sample follows a scene-in-scene layout, where the central scene serves as the reference, and auxiliary scenes provide crucial spatial context for fine-grained classification. Moreover, to tackle the emerging challenge of scene-in-scene classification, we present the context-aware transformer (CAT), a model specifically designed for this task, which adaptively fuses spatial context to accurately classify the scene samples. CAT adaptively fuses spatial context to accurately classify the scene samples by learning attentional features that capture the relationships between the center and auxiliary scenes. Based on MEET, we establish a comprehensive benchmark for fine-grained geospatial scene classification, evaluating CAT against 11 competitive baselines. The results demonstrate that CAT significantly outperforms these baselines, achieving a 1.88% higher balanced accuracy (BA) with the Swin-Large backbone, and a notable 8.87% improvement with the Swin-Huge backbone. Further

experiments validate the effectiveness of each module in CAT and show the practical applicability of CAT in the urban functional zone mapping. The source code and dataset will be publicly available at <https://jerrywyn.github.io/project/MEET.html>.

Index Terms—Fine-grained geospatial scene classification (FGSC), million-scale dataset, remote sensing imagery (RSI), scene-in-scene, transformer.

I. INTRODUCTION

FINE-GRAINED geospatial scene classification (FGSC) with remote sensing imagery (RSI) aims to categorize scene samples into fine-grained geospatial scene categories. Compared to coarse-grained remote sensing scene classification [1]–[4], FGSC presents greater challenges but offers significant utility in various applications, such as water resource management [5], [6], urban planning [7], habitat conservation [8]–[10], etc. Along with increasing high-resolution RSI [11], [12] and powerful deep learning models [13], [14] available, FGSC holds substantial promise and has become a focal point of research [15].

To pursue FGSC, existing research [15]–[17] manually zoom RSI with different rates to form typical scene samples (e.g., the samples from Fig. 1(a)), which are further utilized to train FGSC models. However, in practical applications [18]–[22], the input RSI to be classified is often with fixed-resolution. In this situation, the trained FGSC model with zooming samples may perform poorly. Without no doubt, manual zooming intervention of the input imagery may improve the scene classification performance but inevitably harm automatic process. With this consideration, this paper tries to leverage the fixed-resolution RSI without zooming to form scene samples. However, the zoom-free solution may raise another issue (e.g., the samples from Fig. 1(b)) that fixed-resolution samples presents inter-class and intra-class confusion, especially for fine-grained categories. To address this issue, we introduce scene-in-scene layout to form the scene sample. As shown in Fig. 2, the center scene is the basic unit for classification, while the surrounding scene and global scene serve as auxiliary contexts. Only focusing on center scene leads to confusion between river and lake categories, while this issue can be addressed by introducing auxiliary scenes. In summary, as shown in Table I, existing scene classification datasets have the following issues: 1) Manually zooming scene samples with different rates results in a gap

Manuscript received January 1, 2025; revised January 26, 2025; accepted February 12, 2025. This work was supported by the National Natural Science Foundation of China (42030102, 42371321). Recommended by Associate Editor Hui Yu. (Corresponding authors: Jiayi Ma and Yongjun Zhang.)

Citation: Y. Li, Y. Wu, G. Cheng, C. Tao, B. Dang, Y. Wang, J. Zhang, C. Zhang, Y. Liu, X. Tang, J. Ma, and Y. Zhang, “MEET: A million-scale dataset for fine-grained geospatial scene classification with zoom-free remote sensing imagery,” *IEEE/CAA J. Autom. Sinica*, vol. 12, no. 5, pp. 1004–1023, May 2025.

Y. Li, Y. Wu, B. Dang, Y. Wang, J. Zhang, C. Zhang, Y. Liu, and Y. Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; yuning.wu@whu.edu.cn; bodang@whu.edu.cn; wangfaye@whu.edu.cn; jhzhang7@whu.edu.cn; chugezhang@whu.edu.cn; yiting.liu@whu.edu.cn; zhangyj@whu.edu.cn).

G. Cheng is with the School of Automation, Northwestern Polytechnical University, Xi’an 710129, China (e-mail: gcheng@nwpu.edu.cn).

C. Tao is with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China (e-mail: kingtaochao@csu.edu.cn).

X. Tang is with the School of Artificial Intelligence, Xidian University, Xi’an 710071, China (e-mail: tangxul28@xidian.edu.cn).

J. Ma is with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jiayima@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JAS.2025.125324

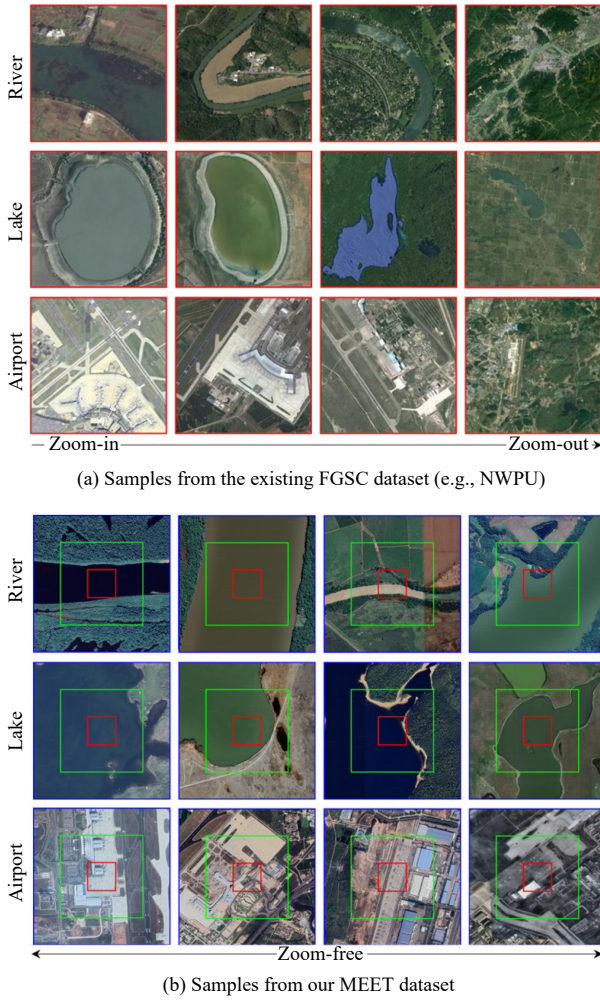


Fig. 1. Illustration of the zoom-free and fine-grained characteristics of our MEET dataset. (a) shows that the existing FGSC dataset forms typical scene samples by manually zooming remote sensing images at different rates. In (b), the center scene outlined in red, is the basic unit for classification, while the surrounding scene outlined in green and global scene outlined in blue serve as auxiliary contextual images. With zoom-free samples and auxiliary scenes, MEET addresses inter-class and intra-class confusion in zoom-free image samples.

with fixed-resolution image interpretation in real-world application; 2) Insufficient sample quantity and limited category coverage restrict the effectiveness of the dataset.

To address the above challenges, we introduce a million-scale fine-grained geospatial scene classification dataset (MEET). In terms of sample size and category coverage, MEET contains 1.03 million samples across 80 geospatial scene categories, surpassing existing datasets [4], [15]–[17], [23]–[39] in sample quantity, diversity, and fine granularity. This initiative is designed to provide a robust foundation for advancing scene classification methods and enhancing practical land-cover applications. To address the challenge of avoiding zoomed scene samples, we incorporate surrounding and global scenes as essential auxiliary context, enriching the classification process. Each sample includes the image to be classified, along with two ranges of surrounding images captured from different field-of-views. Grouping these surrounding and

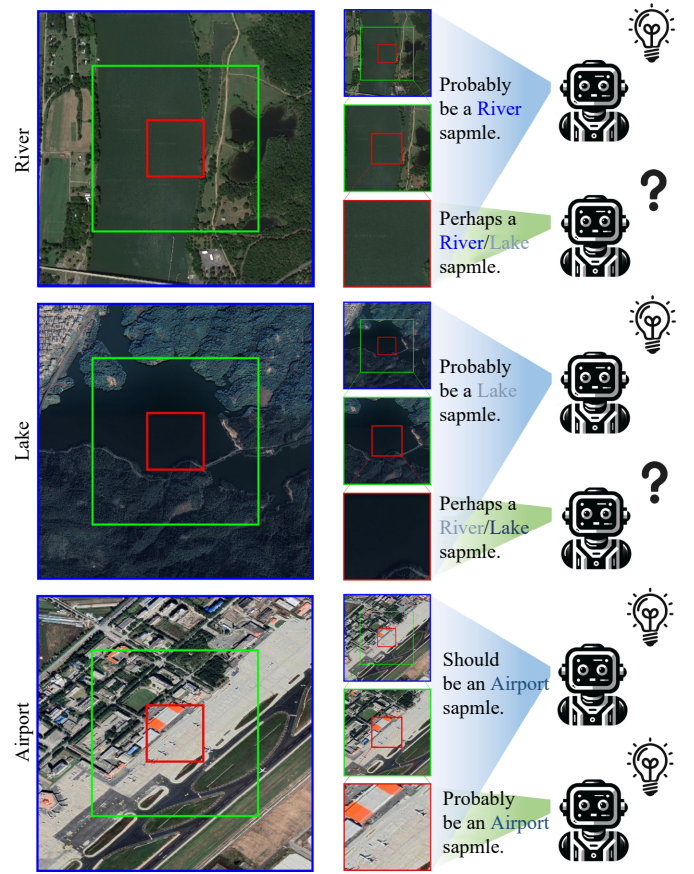


Fig. 2. Superiority illustration of the remote sensing image scene sample with the scene-in-scene layout. The sample labeled as an airport shows the case that models can successfully infer the fine-grained scene category using only the center scene and the auxiliary scenes may benefit improving the classification performance. For the samples from first row and second row, models fail to predict the fine-grained scene category using only the center scene but has a great potential to obtain the right fine-grained scene category using both the center scene and the auxiliary scenes.

global scenes together provides the necessary contextual information for accurate classification of the center scene. Furthermore, this organization offers scalability, enabling adaptive context fusion for varying classification tasks that require different ranges of context. As illustrated in Fig. 2, the center scene serves as the core unit for classification, but focusing solely on it can lead to confusion between categories, such as rivers and lakes. By incorporating auxiliary scenes, this issue can be resolved. The need for additional context depends on the specific classification task. For instance, in the airport category, the model can easily identify the salient object (i.e., an airplane) from the center scene, so no additional context is necessary. However, for categories like rivers and lakes, the introduction of auxiliary scenes becomes essential. These scenes provide a broader field-of-view, allowing the model to better discern features such as riverbanks and water bodies, which helps reduce both inter-class and intra-class confusion.

To advance the research on fundamental and practical issues, we propose a new challenging yet meaningful task: FGSC with zoom-free RSI. By contrast to existing FGSC

TABLE I

COMPARISON AMONG OPEN-SOURCE RSI SCENE CLASSIFICATION DATASETS FINE-GRAINED CHARACTERISTIC REFERS TO A DATASET THAT CONTAINS MORE THAN 30 CATEGORIES. “*” INDICATES THE STATISTIC OF THE PUBLICLY RELEASED PART FROM THE WHOLE DATASET

Dataset	Number of categories	Number of samples	Spatial resolution (m)	Image size	Fine-grained	Zoom-free
UC-Merced [23]	21	2100	0.3	256×256	×	✓
WHU-RS19 [24]	19	1013	up to 0.5	600×600	×	×
RSSCN7 [4]	7	2800	—	400×400	×	—
SAT-4 [25]	4	500 000	1–6	28×28	×	×
SAT-6 [25]	6	405 000	1–6	28×28	×	×
BCS [26]	2	2876	—	—	×	—
RSC11 [27]	11	1232	0.2	512×512	×	✓
SIRI-WHU [28]	12	2400	2	200×200	×	✓
NWPU [16]	45	31 500	0.2–30	256×256	✓	×
AID [17]	30	10 000	0.5–8	600×600	✓	×
RSD46-WHU [29]	46	117 000	0.5–2	256×256	✓	×
EuroSAT [30]	10	27 000	10	64×64	×	✓
PatternNet [31]	38	30 400	0.06–4.7	256×256	✓	×
OPTIMAL-31 [32]	31	1860	—	256×256	✓	—
BigEarthNet [33]	43	590 326	10, 20, 60	[20×20, 120×120]	✓	×
RSI-CB256 [34]	35	24 000	0.3–3	256×256	✓	×
RSI-CB128 [34]	45	36 000	0.3–3	128×128	✓	×
MLRSN [35]	46	109 161	0.1–10	256×256	✓	×
CLRS [36]	25	15 000	0.26–8.85	256×256	×	×
Million-AID* [15]	51	10 000	0.5–153	[256×256, 512×512]	✓	×
SR-RSKG [37]	70	56 000	—	256×256	✓	×
Multiscene [38]	36	100 000	0.3–0.6	512×512	✓	×
WH-MAVS [39]	14	47 137	1.2	200×200	×	✓
Our MEET	80	1 033 778	1	{256×256, 768×768, 1280×1280}	✓	✓

methods [40]–[42] that used expensive external modalities to improve the interpretation of fine-grained categories, utilizing readily available surrounding RSIs is a more suitable and natural choice. Other potential difficulties can be categorized into two main aspects: 1) Avoiding performance degradation in most cases when integrating contextual information; 2) Mitigating drastic memory consumption when applying contextual information, which limits practical usability. Based on MEET, we propose a context-aware transformer (CAT) which can flexibly exploit RSI with multiple field-of-views for FGSC. CAT offers two key advantages: 1) Leveraging spatial context information while avoiding performance degradation; 2) Being user-friendly, lightweight and compatible with pre-trained models. In summary, the MEET dataset and CAT framework proposed in this paper aim to establish a new benchmark for FGSC with zoom-free RSI. With the help of this benchmark, more innovative algorithms can be developed to facilitate the development of FGSC. The main contributions of this paper are as follows:

1) We introduce MEET, the first million-scale dataset for FGSC with zoom-free RSI. It provides over 1.03 million samples where each sample employs a scene-in-scene layout, offering a new data organization for FGSC.

2) To avoid excessive memory consumption, a new CAT is proposed to address FGSC with zoom-free RSI and achieves

progressive visual feature extraction through multi-level supervision.

3) We establish a new benchmark for FGSC with zoom-free RSI based on MEET. Comparisons with existing state-of-the-art algorithms demonstrate the superiority of our CAT. This benchmark may contribute to the fundamental evaluation of FGSC and promote the advancement of practical land-cover applications.

The rest of this paper is organized as follows: Section II reviews existing FGSC datasets and algorithms. Section III describes the proposed MEET dataset in detail. Section IV introduces the proposed CAT for FGSC. Section V presents the experimental results. Finally, Section VI summarizes the paper and provides insights for future work.

II. RELATED WORKS

In this section, we provide a concise review of the most pertinent studies in the field, encompassing scene classification datasets, remote sensing scene classification methods and auxiliary image context exploitation methods.

A. Scene Classification Datasets

As shown in Table I, the emergence of numerous datasets led to significant advancements in remote sensing scene classification. The earliest dataset in this field was UC Merced

[23]. Despite having a relatively small number of samples (only 100 samples per category), it played a crucial role in advancing research on geospatial scene classification tasks. In the subsequent decade, nearly 20 additional remote sensing scene classification datasets were introduced, each contributing to the evolution of the field. The total number of samples in these datasets grew significantly, from a few thousand [4], [23], [24], [26]–[28], [32] to tens of thousands [16], [17], [30], [31], [34], and even to hundreds of thousands [25], [35], [37], [38]. This expansion significantly broadened the scope and application scenarios for scene classification tasks. In terms of category diversity, the number of categories also increased over time, from fewer than 10 [4], [25], [26] to over 40 [15], [16], [29], [33], [35]. For example, the SR-RSKG dataset [37] reached 70 categories, further enhancing the richness of classification tasks. Regarding data sources, most of these datasets were primarily based on Google Earth imagery [4], [15]–[17], [24], [27], while some used freely available medium-resolution satellite imagery, such as Sentinel-2 [30], [33]. A smaller subset of datasets utilized data from other sources, including the United States Geological Survey (USGS) [23], Bing Maps [34], [36], and Tianditu [29].

Despite efforts, existing datasets with zoom-free RSI [17], [23], [27], [28], [30], [39] were limited in terms of the number of categories. To address this issue, many approaches [15], [16], [25], [29], [31], [34]–[38], [43] manually zoomed RSI to construct datasets aimed at improving class separability for FGSC. However, this data-construction technique created a mismatch with practical land-cover applications, which require fixed-resolution imagery. Therefore, developing datasets that incorporate a fine-grained and distinguishable scene category system with zoom-free RSI remained an unaddressed area in previous work.

B. Scene Classification Methods

To motivate FGSC with zoom-free RSI, in the following paragraphs, we reviewed existing deep learning methods for scene classification with RSI and explored potential technologies to address FGSC challenges using zoom-free RSI. Scene classification has been extensively studied across both natural images and RSI, with various approaches applied to both domains.

For natural images, numerous studies [44]–[49] focused on optimizing the design of general scene classification backbones, which were validated across a wide range of visual downstream tasks. Similarly, in the domain of RSI, significant advancements in scene classification were achieved through three primary approaches: 1) Training models from scratch [50]–[53]; 2) Adapting pre-trained models from ImageNet to RSI [16], [54], [55]; 3) Fine-tuning pre-trained models specifically for RSI data [14], [56]–[59]. Among the methods involving training from scratch, ARC-Net [50] incorporated residual blocks with asymmetric convolution (RBAC) to reduce computational cost and shrink the model size. Additionally, dilated convolutions and multi-scale pyramid pooling modules were used to expand the receptive field and improve accuracy. Bai *et al.* [51] proposed a multiscale feature fusion covariance network with octave convolution,

which extracted multifrequency and multiscale features from RSIs. Chen *et al.* [52] introduced GCSANet, which leveraged global context spatial attention (GCSA) and densely connected convolutional networks to capture multiscale global scene features. For methods involving fine-tuning pre-trained models specifically for RSI data, Guo *et al.* [14] introduced geo-context prototype learning to learn region-aware prototypes based on RSI's multi-modal spatiotemporal features. Each of these approaches uniquely enhanced the discriminative ability and robustness of models, driving advancements in the field. For the more challenging task of FGSC, many studies turned to auxiliary data to improve the interpretation of fine-grained categories. Srivastava *et al.* [60] optimized FGSC performance by utilizing visual cues from side-view pictures sourced from Google street view (GSV). Similarly, Fang *et al.* [40] incorporated street view images (SVI) and developed a spatial context-aware land-use classification method to enhance land-use classification accuracy. Yao *et al.* [61] introduced temporal resolution time-series electricity data to explore the relationship with socioeconomic attributes and constructed a neural network that can fuse time-series electricity data and RSIs to identify urban land-use types. Arbingner *et al.* [62] introduced geographic coordinates or geoinformation data to enable a better understanding of the image content and thus facilitate their classification. The limited availability and high acquisition cost of additional data sources posed challenges and restricted the broader application of these methods.

C. Auxiliary Image Context Exploitation Methods

In literature, incorporating auxiliary contextual information was regarded as a natural and effective approach to enhance the interpretability of RSI. In semantic segmentation of RSI, Li *et al.* [63] proposed a deep adaptive fusion network with multi-scale context, specifically designed for RSI semantic segmentation. GLNet [64] preserved both global and local information in a highly memory-efficient manner, capturing high-resolution fine structures from zoomed-in local patches and contextual dependencies from the downsampled input. CascadePSP [65] used a global step to refine the entire image, providing sufficient image context for a subsequent local step to perform full-resolution, high-quality refinement. In object detection of RSI, HBD-Net [66] addressed bridge detection by incorporating multi-scale context within the dynamic image pyramid (DIP) of large-scale images, while employing a shape-sensitive sample re-weighting (SSRW) strategy to balance regression weights for bridges with varying aspect ratios. GLGCNet [67] extracted global representations and combined them with local-context-aware features gathered from three saliency-up modules for comprehensive saliency modeling. An edge assignment module was also employed to refine preliminary detections. GeoAgent [68] enhanced performance by adaptively capturing contextual information based on geographical objects, using a feature indexing module to differentiate locations. However, to the best of our knowledge, no work on scene classification has yet utilized contextual information, let alone for FGSC. Furthermore, these semantic segmentation and object detection methods could not be directly

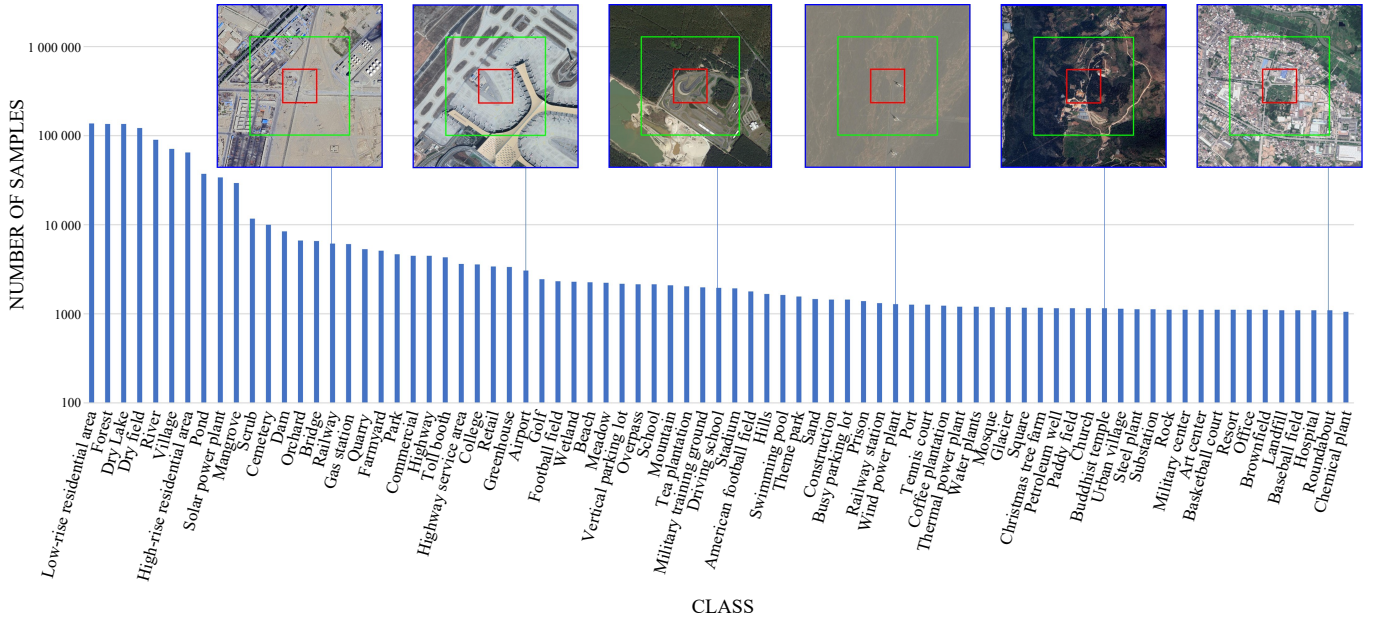


Fig. 3. Statistics and visualization of samples from MEET.

adapted to FGSC with image-level labels. This left the exploitation of contextual information in FGSC as an open and significant research space.

III. THE PROPOSED MEET DATASET

Our goals for developing a new dataset for FGSC are twofold: 1) To promote a new meaningful yet challenging task: FGSC with zoom-free RSI; 2) To occupy the niche of FGSC datasets with context image. This section provides a comprehensive overview of the MEET dataset, focusing on three key aspects: data collection and organization, data annotation, and data analysis.

A. Data Collection and Organization

We select fixed-resolution samples and supplement them with surrounding imagery as multi-scale context. To ensure data diversity, images are collected globally, covering variation in appearance, illumination and occlusion. The fine-grained geospatial scene category is determined by the center scene, with auxiliary scenes serving as contextual information. The MEET dataset provides global coverage through the collection of 1.03 million samples spanning Asia, Africa, South America, North America, and Europe, and covering 80 typical scene categories, as shown in Figs.3. and 4. The images are collected from 2018 to 2022, with each sample containing a center scene with 256×256 pixels, along with a surrounding scene with 768×768 pixels and a global scene with 1280×1280 pixels. The overall distribution and some samples are shown in Fig. 5. It is important to note that the spatial resolution of all samples is consistently set to 1.0 m. Some representative examples from all categories of the proposed MEET dataset are presented in Fig. 6.

To comprehensively obtain RSI with diverse and comprehensive scene categories on a global scale, we leverage data from OpenStreetMap (OSM), which is a collaborative project creating a free, editable map of the world. OSM provides semantic annotations by labeling geographical features and

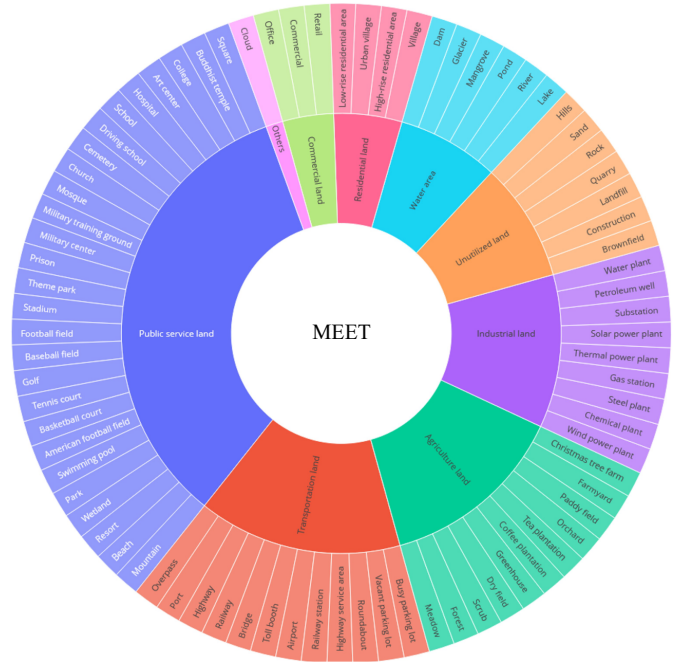


Fig. 4. Hierarchical scene category of MEET. All categories are hierarchically organized in a two-level tree: 80 leaf nodes fall into 9 parent nodes, representing 9 underlying scene categories of commercial land, residential land, water area, unutilized land, industrial land, agriculture land, transportation land and public service land.

land-use within its maps, offering detailed information about roads, buildings, and other points of interest. Subsequently, we preprocess the acquired data by performing coarse filtering or integration of scene categories based on the quality of OSM annotations, and design a series of rules to reduce noise at the image-level labels. Finally, we ensure a global distribution and high richness of samples by defining random spatial windows using geographic coordinates.

We randomly divide the entire dataset into training, valida-

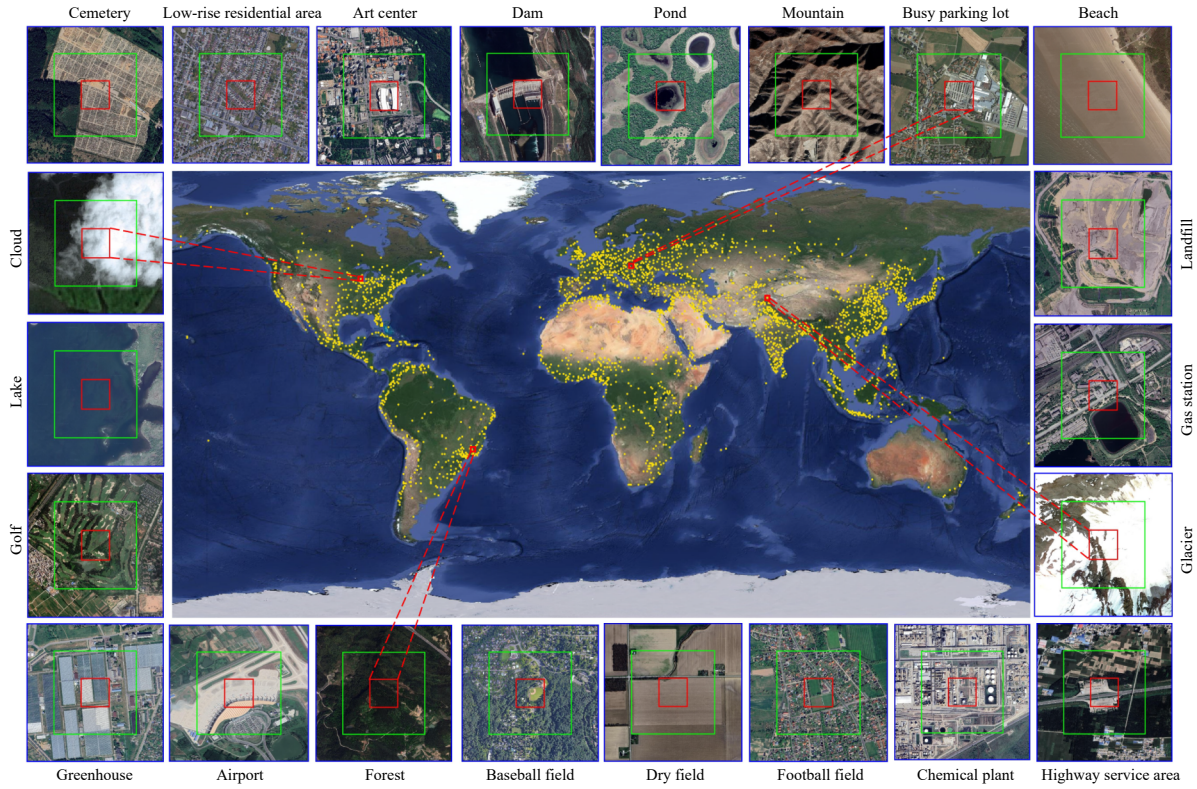


Fig. 5. The geographical distribution map of the sampled images from the proposed MEET dataset.

tion, and testing sets by category, with a ratio of 6 : 2 : 2. More specifically, the training set, validation set, and testing set contain 620 237 206 755, and 206 755 samples, respectively.

B. Data Annotation

Overall, fine-grained geospatial scene categories are defined by considering the center scene along with its auxiliary scenes. To ensure precise annotations of MEET dataset, ten trained experts in the field of remote sensing participate in the annotation process with cross-validation. During the annotation process, each annotator receives the center scene along with the corresponding contextual imagery. For samples whose categories can be determined solely based on the center scene, annotators use the predominant scene classification within the image as the scene label for the current sample. For example, a mosque located within a residential area of a city, due to its smaller footprint and relative rarity, is categorized under mosque. This strategy enhances the coverage of smaller target features, thereby enriching the holistic understanding of urban attributes. As far as samples whose categories require contextual imagery to determine, annotators can also make correct labeling choices by combining the corresponding context. For instance, for texture-poor water body region images, annotators can distinguish whether the sample belongs to the river or lake category by considering the shape of the banks in the surrounding context.

The procedure of labeling MEET encompasses a tripartite framework consisting of three stages: pre-annotation stage, expert feedback and optimization stage, and large-scale detailed annotation stage. In the initial phase of pre-annotation, we form a specialized team comprising 10 members,

each possessing extensive expertise in the field of remote sensing interpretation. This team undergoes comprehensive training in fundamental annotation techniques and subsequently conducts annotation tests on a representative subset of the dataset. In the following feedback and optimization stage, experts thoroughly review and evaluate the team's initial annotations, leading to the formulation of improved annotation standards. Subsequently, guided by these adjustments, the team embarks on the formal large-scale annotation process, accompanied by experts' random sampling inspections.

C. Dataset Analysis

The MEET dataset distinguishes itself from existing remote sensing scene classification datasets through several unique attributes: the breadth of its category coverage, the scale of its sample size, the diversity of its samples, and the incorporation of contextual information. Additionally, the dataset maintains a uniform sample resolution and is tailored to support models designed for large-scale scene classification and mapping tasks, further emphasizing its distinctive characteristics.

1) *Fine Granularity of Categories*: The MEET dataset comprises 80 fine-grained geospatial scene categories, categorized into 11 major scene types. With the introduction of auxiliary contextual information, it has become possible to annotate more fine-grained categories. These categories comprehensively cover discernible remote sensing scene categories. Therefore, our MEET dataset offers advantages over existing remote sensing scene classification datasets by providing more high-value fine-grained scene categories. Especially in urban mapping and analysis applications, these fine-grained scene categories make a wider range of scene classification applications possible.



Fig. 6. Some examples from the proposed MEET dataset, which employs a scene-in-scene layout. These images exhibit rich variations in appearance, illumination, background, occlusion, and other factors.

2) *Large Volume of Samples*: The MEET dataset includes 1 033 801 samples, covering over 3.3 billion square kilometers globally. It surpasses other publicly available datasets in both sample volume and richness of annotation data for remote sensing scene classification.

3) *High Intra-Class Variability and Inter-Class Similarity*: Intra-class variation is mainly due to differences in appearance. Inter-class similarity arises from similar appearance representations in the center scene, but it manifests differently in auxiliary scenes. As shown in Fig. 7(a), samples in the river category exhibit high richness in image quality, color variations, seasonal changes, geographic regions, and river widths. Conversely, inter-class similarities underscore the utility of contextual information in enhancing classification accuracy, as illustrated in Fig. 7(b) where incorporating context aids in discerning challenging objects within the current block.

Although the distribution of the MEET dataset exhibits a certain degree of class imbalance, as shown in Fig. 3, it closely mirrors the frequency distribution of real-world scene categories. This characteristic enhances its value for practical applications. Additionally, it is important to emphasize that

the selected samples exhibit significant variation within each category, ensuring that even among the head classes, homogeneous low-value samples are also relatively few.

IV. PROPOSED METHOD

To flexibly and efficiently exploit the scene-in-scene layout in FGSC with zoom-free RSI, this paper introduces CAT, a novel approach specifically tailored for this task. CAT incorporates an adaptive context fusion module to effectively extract multi-scale contextual features from the transformer backbone. To ensure performance without excessively increasing parameters, we utilize parameter-efficient fine-tuning (PEFT) methods to finetune the backbone, instead of training from scratch or parameter synchronization. Additionally, we introduce multi-level supervision through independent classification heads during training. This improves feature learning at each level and mitigates overfitting that can arise from auxiliary scenes. More specifically, we utilize the scene-in-scene layout for each sample with large range of context. However, contemporary deep networks face limitations in directly processing large-size RSI due to GPU memory constraint. To

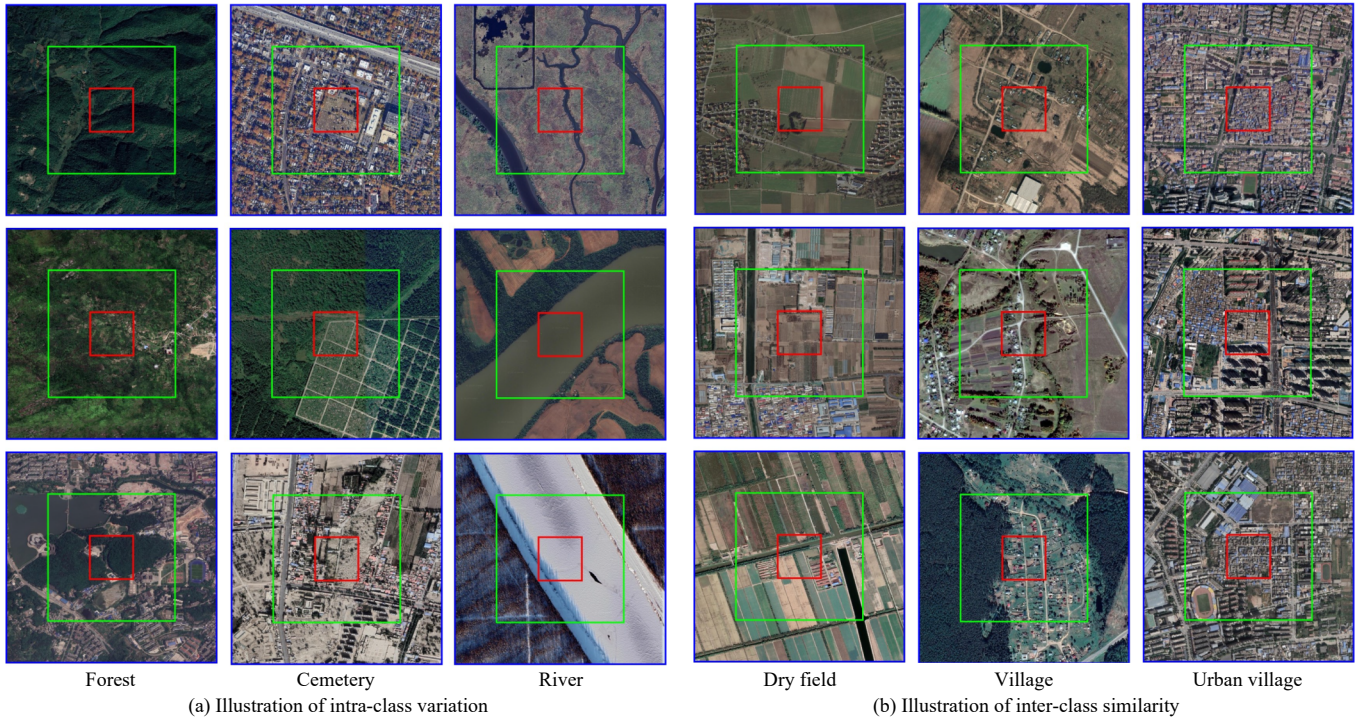


Fig. 7. Illustration of intra-class variation and inter-class similarity in MEET.

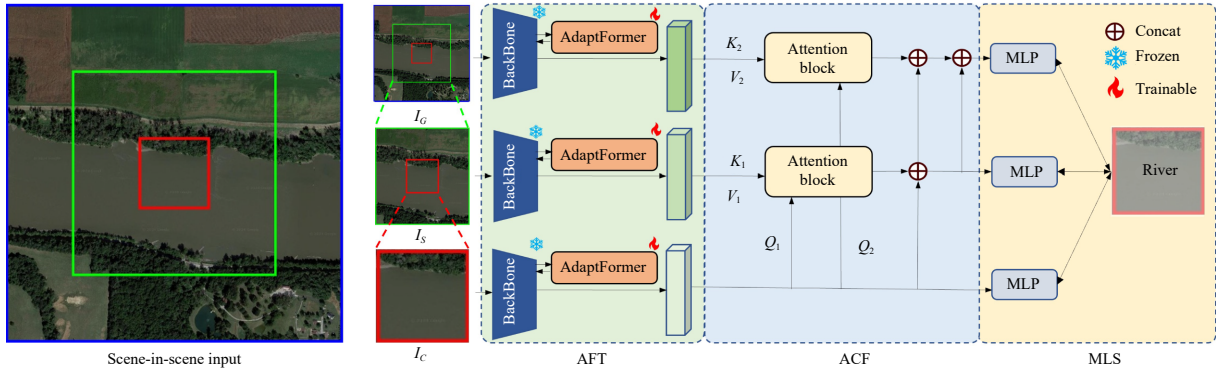


Fig. 8. Overview of CAT. The structure contains three components (from left to right): AdaptFormer tuning (AFT), adaptive context fusion (ACF), and multi-level supervision optimization (MLS).

address this challenge, we resize each image to a uniform size, denoted as I_C , I_S and I_G , respectively. Among them, I_C corresponds to center scene, while I_S and I_G correspond to the surrounding and global scene, respectively. During both the training and inference stages, the input to CAT remains consistent. The whole architecture of our method is illustrated in Fig. 8. This section is dedicated to provide a detailed explanation of the CAT.

A. AdaptFormer Tuning

For the currently constructed scene-in-scene layout, three branches are needed to perform feature extraction respectively while using pre-trained weights. Although efficiency has been improved by resizing input images, using three branches for feature extraction still poses difficulties. Using completely independent backbones for full-parameter training would significantly increase the model's parameter size, which is unacceptable given the current trend towards larger model parameters. While using shared weights does not intro-

duce additional parameters, it would hinder performance because the three branches have inputs with fixed but different spatial resolutions. To address this issue, we introduce the AdaptFormer tuning (AFT) on the transformer backbone for multi-level image feature extraction. AdaptFormer replaces the MLP modules in the transformer encoder with AdaptMLP. The computation of the AFT module in transformer can be expressed as follows:

$$z_i = W\text{-MSA}(\text{LN}(z_{i-1})) + z_{i-1} \quad (1)$$

$$\hat{z}_i = \text{MLP}_{\text{AFT}}(\text{LN}(z_i)) + z_i \quad (2)$$

$$\hat{z}_{i+1} = \text{SW-MSA}(\text{LN}(\hat{z}_i)) + \hat{z}_i \quad (3)$$

$$z_{i+1} = \text{MLP}_{\text{AFT}}(\text{LN}(\hat{z}_{i+1})) + \hat{z}_{i+1} \quad (4)$$

where z_i denotes the feature output of the i -th module in transformer. For each branch, the backbone uses shared weights and is initialized with the pre-trained model's weights, while

independent AdaptFormer modules are used for training in the three branches. During training, we freeze the weights from the pre-trained model and only update the weights of the AdaptFormer. With this design, the three branches use the AFT method to enable feature extraction from input images of different spatial resolutions. At the same time, the three branches share most of the parameters, ensuring that the total parameter size does not increase significantly, thus maintaining the model's efficiency and practicality.

B. Adaptive Context Fusion

The model uses a backbone combined with AFT to extract features from I_C , I_S and I_G , denoted as F_C , F_S and F_G , respectively. Among them, feature F_C contains rich semantic information most relevant to the labels, while F_S and F_G serve as contextual features supplementing F_G .

These contextual features often contain redundant information and are not entirely correlated with the labels. As mentioned in [69], excessive redundant contextual information can impair classification results for certain samples. Therefore, inspired by the design of multi-head self-attention modules, we propose the adaptive context fusion (ACF) module to adaptively integrate features from the center scene with either the surrounding scene or the global scene. For the features extracted from each level of contextual images, the most relevant and valuable features associated with the center scene are further extracted, reducing the redundant information brought by large-scale geographic areas.

Specifically, we employ two multi-head attention modules for adaptive contextual image feature fusion on surrounding scene and global scene. The query feature F_C retrieves features from the current block, while the keys and values, derived from F_S or F_G , are obtained from the corresponding contextual blocks. This process facilitates adaptive feature extraction from the context based on the visual feature of the current block, thereby enhancing focus on the most relevant features. The ACF module is implemented using the Multi-HeadAttention module. The F_{ACF}^S and F_{ACF}^G are defined as follows:

$$F_{ACF}^S = ACF(F_C, F_S) \quad (5)$$

$$F_{ACF}^G = ACF(F_C, F_S, F_G) \quad (6)$$

where F_{ACF}^S and F_{ACF}^G are the high-value visual features adaptively extracted from the surrounding scene and global scene, respectively, based on the center scene. For the median branch, features F_{ACF}^C and F_{ACF}^S are concatenated to obtain F_{Fused}^S . For the global branch, features F_C , F_{ACF}^S , F_{ACF}^G are concatenated to obtain F_{ACF}^G . These factors can be expressed as

$$F_{fused}^S = Concat(F_{ACF}^C, F_{ACF}^S) \quad (7)$$

$$F_{fused}^G = Concat(F_{ACF}^C, F_{ACF}^S, F_{ACF}^G). \quad (8)$$

The ACF Module outputs two contextual fusion features, F_{Fused}^S and F_{ACF}^G , along with the center scene feature F_C . Features at each level contain high-value visual information rele-

vant to the center scene, with F_C as the primary feature for that scale.

C. Optimization With Multi-Level Supervision

Intuitively, directly utilizing the visual features richest in F_{ACF}^G might achieve the highest classification performance. However, introducing auxiliary scenes may overlook discriminative features of the center scene itself and lead to model overfitting, thus undermining the performance. Therefore, utilizing only the branch features rich in contextual information for supervised learning is expected to be insufficient and may also reduce the model's generalization capability. To address this, we propose a multi-level supervision (MLS) strategy. Specifically, the features extracted from the three branches are subsequently fed into three classification heads for prediction P_C , P_S and P_G . These factors can be expressed as

$$P_C = HEAD_C(F_C) \quad (9)$$

$$P_S = HEAD_S(F_{fused}^S) \quad (10)$$

$$P_G = HEAD_G(F_{fused}^G). \quad (11)$$

MLS strategy uses ground truth to constrain predictions from all branches to calculate the loss. This ensures that the model extracts effective features even at smaller field-of-view, thereby reducing the risk of associating category semantics with erroneous visual features from the context, thus preventing overfitting. The total loss of FGSC is defined as follows:

$$Loss_{all} = Loss_C + Loss_S + Loss_G. \quad (12)$$

During inference, we use P_G as the model's predictor, which possesses the complete auxiliary scene information and gets sufficient generalization by MLS.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this subsection, we first introduce the evaluation metrics, and then describe our implementation details and mainstream methods for FGSC. Finally, extensive evaluation of our proposed CAT are performed on the MEET dataset.

A. Evaluation Metrics

Considering the natural long-tail distribution of different scene categories in real-world scenarios, this study uses overall accuracy (OA) and balance accuracy (BA) as the primary evaluation metrics. The overall accuracy (OA) is defined as the number of correctly classified images divided by the total number of images in the dataset. The score of OA reflects the overall performance of classification models instead of per class as follows:

$$OA = \frac{N_c}{N_t} \quad (13)$$

where N_c represents the number of correctly classified images, and N_t represents the total number of images in the dataset. The balance accuracy (BA) is defined as the average OA across all classes in the dataset. The BA score reflects the average performance of the classification model across each class as follows:

$$BA = \frac{1}{C} \sum_{i=1}^C OA_i \quad (14)$$

where OA_i represents the OA of the i -th class. To further understand the performance on the dataset, we categorize the MEET dataset based on sample quantities. Specifically, we define a set of sample ranges as (0, 1500], (1500, 10 000], (10 000, 150 000]. Categories are classified based on their sample counts into many, medium (Med), and few. The corresponding BAs are denoted as BA_{many} , BA_{med} , and BA_{few} .

B. Implementation Details

Generally, most algorithms used in our experiments are sourced from the open-source PyTorch-based library TIMM. This library integrates various state-of-the-art computer vision models, along with their respective backbones, feature extractors, and classification heads. They are capable of reproducing the original accuracy of their respective algorithms within a unified framework, ensuring fairness. Additionally, for other models, we use official open-source code as much as possible to ensure experimental rigor. The experiments are conducted on a server with 1 NVIDIA GeForce RTX 3090 GPU and 24 GB of memory. To ensure a fair comparison, we apply the most consistent pre-trained model parameters across all methods. We use the Adam optimizer with a learning rate of 0.00005. For some remote sensing models, which are relatively smaller, a learning rate of 0.0005 is used to avoid significantly reducing training efficiency, except for SkySense. In all experiments, the batch size is set to 16, except for those using Swin-Huge as the backbone, where it is set to 8.

C. Mainstream Methods

To establish a benchmark for FGSC with zoom-free RSI, we re-implement scene classification methods. In the remote sensing field, we select several representative works: ARC-Net [50], MF2CNet [51], GCSANet [52], DOFA [59] and SkySense [14]. Given the rapid progress in exploring backbone models in the general computer vision field, we also incorporate many widely validated methods as strong comparison benchmarks, including ResNet [45], HRNet [46], Inception-Next [44], MaxViT [47], DAVit [48], and swin transformer (Swin) [49]. To ensure the performance of baseline methods, we train all baseline methods with full parameters. While considering the practical usability of our CAT, we use AFT for parameter-efficient fine-tuning, which means performance of CAT could potentially be further improved with full-parameter training.

Considering the substantial benefits of pre-trained model weights for downstream tasks, we use pre-trained model weights on ImageNet-22K for initialization wherever possible, and thus use center scenes or global scenes as model input to meet the three-channel input requirement. For our CAT, since it is specifically designed for scene-in-scene layout, both the center scene and auxiliary scenes are used as inputs.

D. Results and Analysis

The benchmark and experimental results of FGSC on the

MEET dataset are shown in Table II. The experimental results indicate that when using Swin-Large as the backbone, our method outperforms comparison methods on the MEET dataset benchmark. Our CAT achieves an OA of 95.87% and a BA of 83.38%. Compared to all baselines with a similar number of parameters (excluding Swin-Huge), our method shows an improvement of nearly 1% in OA and over 4% in BA compared to methods using center scenes as input. Compared to methods using global scenes as input, our method shows an improvement of over 0.3% in OA and over 1.8% in BA. These results highlight a significant advantage across all evaluated metrics, surpassing both methods specifically designed for scene classification and those generally proposed for image recognition. Additionally, compared to some other backbone networks [44], [46], [48], the Swin-Large model outperforms other methods significantly by incorporating the ACF to fully utilize contextual image information and using MLS and AFT methods to further enhance performance. Specifically, the performance gains come from the model's more powerful feature extraction capabilities. The model can incorporate complementary cues from surrounding imagery for the center scene, especially for cases where the center scene lacks prominent visual features. Additionally, the model does not overfit due to the large amount of redundant information in the surrounding imagery. This is reflected in the performance gains for the tail classes in terms of BA.

To further demonstrate the effectiveness and generalization capabilities of our CAT, we also conduct experiments using the large foundation model, namely Swin-Huge from SkySense [14]. It can be observed that the model's performance is further enhanced when employing our CAT, achieving the best performance on both OA and BA, with scores of 97.74% and 89.37%, respectively, thanks to the pre-training parameter methodology of the SkySense model. Compared to methods using Swin-Huge as backbone, CAT shows an improvement of over 2.8% in OA and over 9.5% in BA. Therefore, it can be further verified that CAT achieves stable performance improvements across backbone models of different sizes. For the following ablation study, we opt to use the Swin-Large version to reduce the cost of training resources.

As shown in Fig. 9, the top 3 predictions from various comparison methods are presented, as well as those provided by CAT. It can be seen that CAT not only performs the best in classification but also achieves the highest prediction confidence. This is due to CAT's strong capability in adaptive feature extraction of spatial context, ultimately leading to more stable and accurate classification results.

To illustrate the superiority of our CAT in fusing multi-level contexts, we design several classical context fusion strategies, as shown in Fig. 10. These strategies fuse at the input, feature, and decision levels, respectively. The performance results in Table III demonstrate that our CAT achieves the best performance through ACF.

Overall, comparison results on multiple metrics demonstrate that our CAT is highly effective. It makes full use of contextual information for feature extraction, achieving the best performance. Moreover, each component of our method is independent of specific scene classification methods, allow-

TABLE II
PERFORMANCE (%) COMPARISON OF THE PROPOSED CAT AND OTHER METHODS ON THE MEET DATASET

Method	Type	Backbone	Pretrain Dataset	Input	Performance metrics				
					OA	BA _{many}	BA _{med}	BA _{few}	BA
Methods Generally Proposed for Natural Image Recognition									
InceptionNext [44]	CNN	InceptionNext-Base	ImageNet-1K	Center scene	90.37	93.66	67.99	66.71	71.02
				Global scene	92.98	96.72	73.00	62.88	72.34
Resnet [45]	CNN	Resnet101	ImageNet-22K	Center scene	81.58	91.21	52.71	47.44	55.96
				Global scene	82.80	95.46	51.53	39.57	52.94
HRNet [46]	CNN	HRNet-w64	ImageNet-22K	Center scene	91.66	93.93	71.49	73.25	75.26
				Global scene	94.27	96.52	78.80	67.25	76.76
MaxViT [47]	Transformer	Maxvit-Large	ImageNet-22K	Center scene	91.22	94.94	69.43	69.79	73.08
				Global scene	93.57	97.34	74.72	65.90	74.41
DAViT [48]	Transformer	Davitt-Base	ImageNet-22K	Center scene	90.85	94.64	69.02	66.54	71.58
				Global scene	94.26	97.51	76.65	68.91	76.52
Swin [49]	Transformer	Swin-Large	ImageNet-22K	Center scene	92.23	94.89	73.95	71.24	75.78
				Global scene	95.58	97.82	<u>83.04</u>	73.82	81.50
Methods Specifically Proposed for RSI Scene Classification									
ARCNet [50]	CNN	Resnet101	–	Center scene	88.55	93.27	63.04	59.94	65.99
				Global scene	89.78	96.18	65.59	52.84	64.85
MF2CNet [51]	CNN	Resnet50	–	Center scene	67.52	85.75	31.90	25.18	36.70
				Global scene	88.43	88.43	34.93	20.39	36.65
GCSANet [52]	CNN	Densenet121	–	Center scene	85.44	92.39	59.04	54.10	61.71
				Global scene	89.04	95.72	63.55	50.39	62.88
DOFA [59]	Transformer	Vit-Large	DOFA	Center scene	94.88	96.88	80.67	71.41	79.31
				Global scene	93.31	94.37	77.52	78.97	80.40
SkySense [14]	Transformer	Swin-Huge	SkySense-21.5M	Center scene	94.52	97.55	79.57	73.76	79.79
				Global scene	94.93	<u>98.41</u>	81.43	67.72	78.45
Our CAT	Transformer	Swin-Large	ImageNet-22K	Scene-in-scene	<u>95.87</u>	97.04	82.14	<u>80.05</u>	<u>83.38</u>
Our CAT	Transformer	Swin-Huge	SkySense-21.5M	Scene-in-scene	97.74	99.00	90.80	84.19	89.37

TABLE III
PERFORMANCE (%) COMPARISON OF THE PROPOSED CAT AND OTHER FUSION STRATEGIES

Strategy	OA	BA _{many}	BA _{med}	BA _{few}	BA
Input-level fusion	81.41	91.19	49.31	39.44	51.24
Feature-level fusion	<u>86.26</u>	<u>95.37</u>	<u>57.99</u>	<u>46.74</u>	<u>58.77</u>
Decision-level fusion	85.78	95.24	56.95	43.46	56.99
Our CAT	95.87	97.04	82.14	80.05	83.38

ing it to seamlessly adapt and enhance performance across most proposed backbones without encountering specific limitations. This observation highlights the versatility and applicability of our proposed method.

E. Ablation Study of Our CAT

The ablation experiment on the MEET dataset is to assess the impact of three key components of our proposed method: ACF, MLS and AFT. As shown in Table IV, we explore the effectiveness of the proposed ACF module. The ACF module significantly improves scene classification performance by adaptively extracting high-value graphical feature informa-

tion from contextual data. Compared to the baseline, OA increases by 2.06% and BA improves by 1.48%. This module provides the greatest performance boost by introducing incremental information and effectively merging features. The introduction of MLS further improves BA, indicating it effectively reduces overfitting on contextual information for most categories. Compared to using only the ACF module, MLS results in a slight decrease in head classes, due to most head classes' strong reliance on contextual information, which is different from the minority classes. The introduction of the AFT module results in increases of 0.65% and 2.03% in OA and BA, respectively, reflecting that we successfully enhanced the feature extraction capabilities of the three branches for different inputs without adding excessive parameters. These results highlight the capability of our method to enhance the performance of existing state-of-the-art scene classification methods. Additionally, we evaluate the running time per sample and the parameter number for CAT, demonstrating that each module of CAT is lightweight and does not significantly impact efficiency. To further demonstrate that CAT can mine the auxiliary image context to combat intra-class variability and inter-class similarity, we compare the class-wise accu-

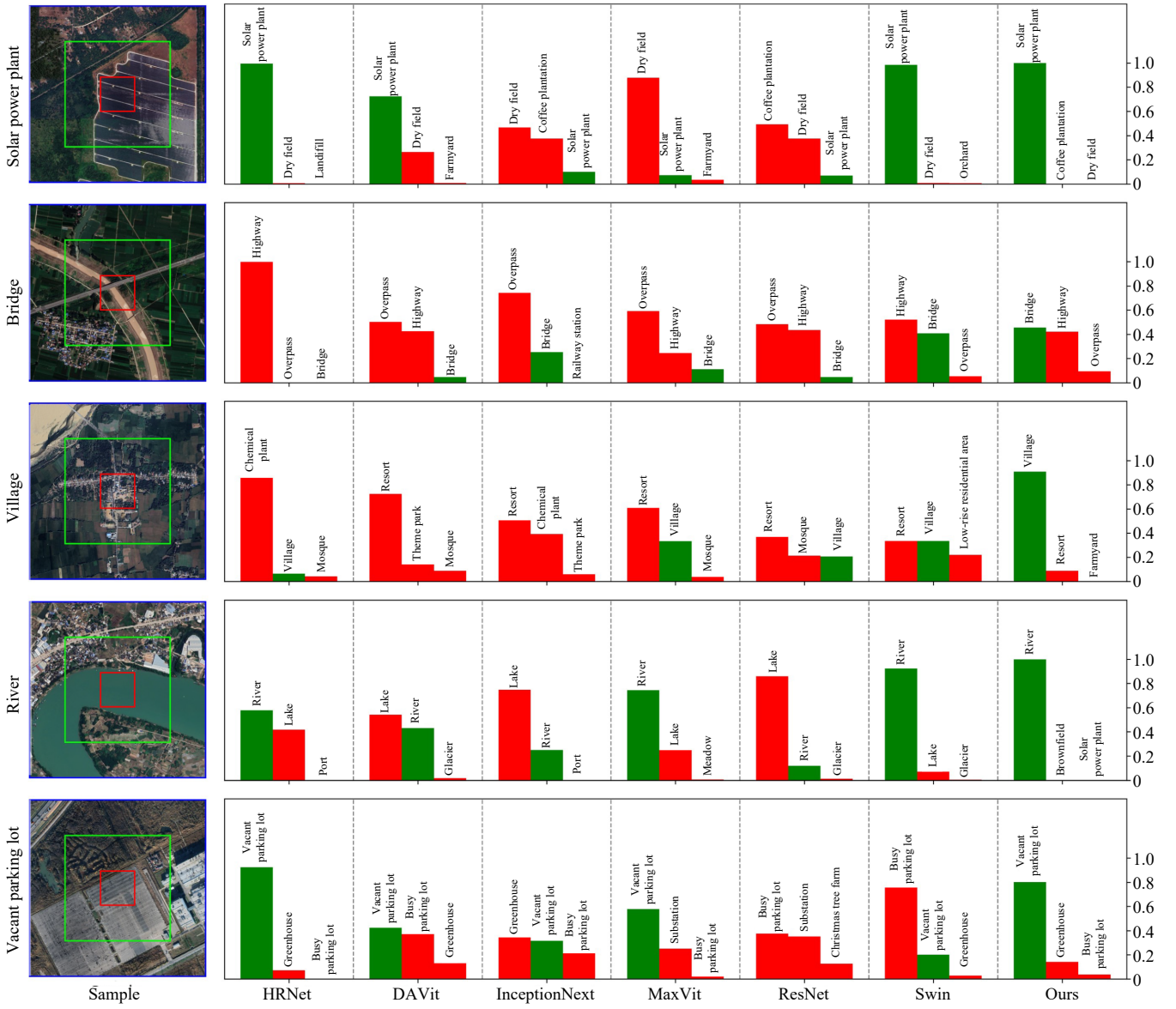


Fig. 9. Several samples on MEET. Top3 predictions are presented from various comparative methods, as well as those provided by our CAT. Correct prediction categories are displayed in green, and incorrect prediction categories are displayed in red.

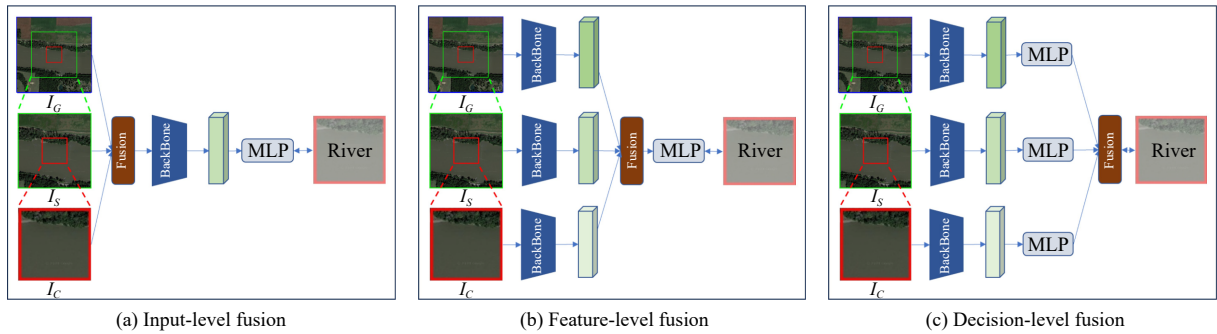


Fig. 10. Illustration of three baseline fusion strategies.

racy across all 80 categories on the MEET dataset before and after applying CAT, as illustrated in Fig. 11. The results show a significant improvement in accuracy for the majority of categories. Specifically, the accuracy for the River category improved by 3%, and the accuracy for the Lake category

improved by 5%.

The effectiveness of our method has been quantitatively evaluated in Tables II and IV. To further illustrate its capability in contextual feature extraction, we visualize the t-SNE feature map on each branch. We employed t-SNE method to

TABLE IV
PERFORMANCE (%) COMPARISON OF THE ABLATION STUDY (RUNNING TIME REFERS TO RUNNING TIME PER ONE SAMPLE)

ACF	MLS	AFT	Running time (s)	Parameters (M)	OA	BA_{many}	BA_{med}	BA_{few}	BA
×	×	×	0.0139	195.12	92.23	94.89	73.95	71.24	75.78
✓	×	×	0.0168	226.05	94.29	97.94	79.00	67.80	77.26
✓	✓	×	0.0168	226.05	<u>95.22</u>	97.01	<u>80.64</u>	<u>76.65</u>	<u>81.35</u>
✓	✓	✓	0.0174	233.15	95.87	<u>97.04</u>	82.14	80.05	83.38



Fig. 11. Comparison of class-wise accuracy across all 80 categories on the MEET dataset before and after applying ACF + MLS + AFT.

visualize the learned representative features of each ablation study setting. The perplexity for all four cases is 10 and 50 samples from all 80 geospatial scene categories are randomly selected to create a t-SNE plot. As shown in Fig. 12, after

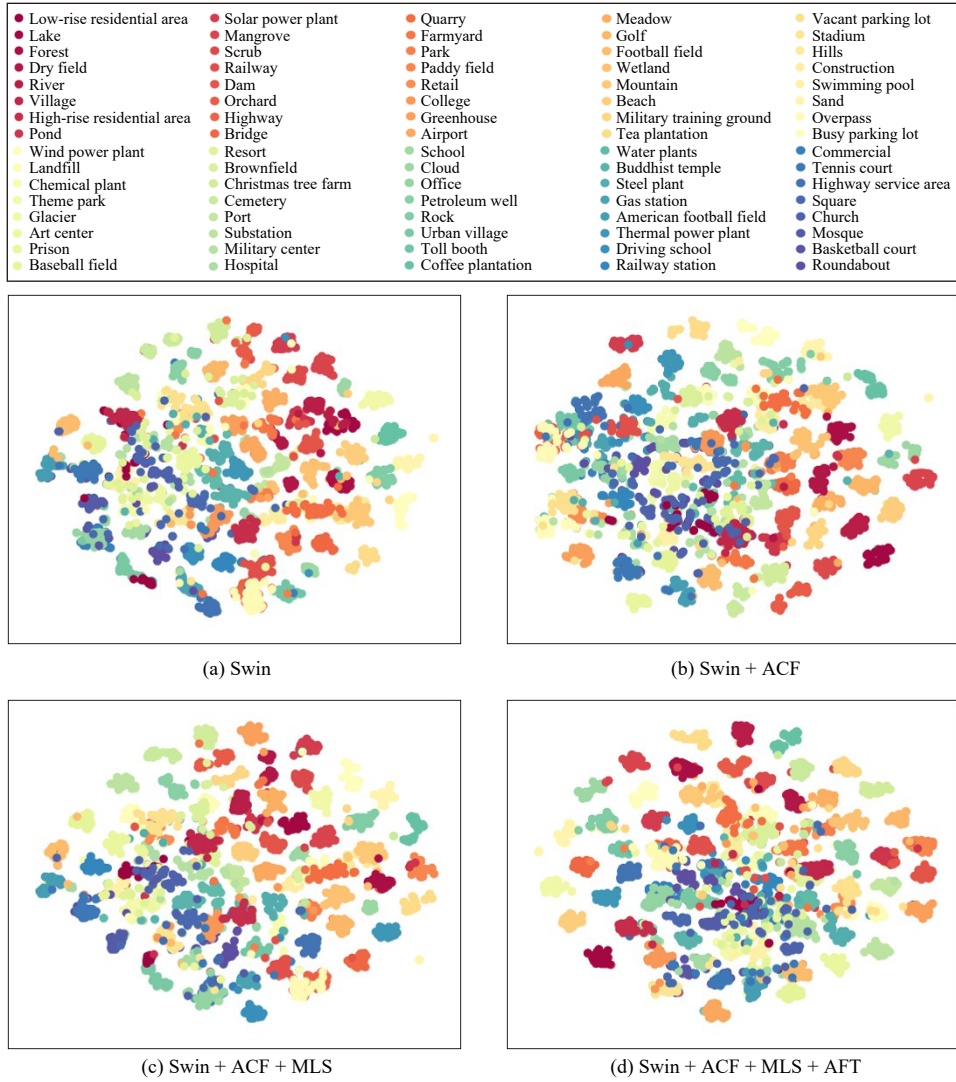


Fig. 12. Visualization of the features for different ablation study settings of our CAT.

applying ACF, the embeddings do not become significantly more separable in the feature space. This may be due to the model overfitting to some extent on the auxiliary scene. However, after applying MLS to enhance the feature extraction capabilities for both the center scene and surrounding scenes, and applying the AFT module to enhance the feature capabilities of the three branches, the embeddings become significantly more separable. These results indicate that our CAT achieves better class separability at the feature level.

F. Superiority Analysis of Our CAT

In Fig. 13, we present some examples with scores and class activation map (CAM) on each branch. The scores represent the prediction on the ground truth category (shown on the left side) after applying softmax from classification head on that branch. Changes in the scores reflect the gain in performance due to the accumulation of multi-level context. It can be observed that the predictions consistently improve with the introduction of multi-level auxiliary scenes, which is reasonable for cases that require auxiliary scenes. However, for cases that can achieve sufficient saliency without auxiliary scenes, such as airport example in Fig. 13, the model perfor-

mance has not degraded with the introduction of redundant information. This demonstrates that our model has an adaptive context fusion capability, showing strong generalization. With the introduction of auxiliary scenes, the model can extract more visual features from the context to interpret the center scene. Specifically, from the examples of river and lake, it can be observed that using only the center scene as input is not sufficient. After introducing auxiliary scenes, the input data includes contextual information, such as riverbanks, enabling the model to correctly differentiate the water body into a river or lake. For the village example illustrated, the visual features of fields included in the auxiliary scenes contribute to distinguish the village category from other similar categories in the center scene, such as low-rise residential area category.

G. Application Evaluation of Our MEET on Urban Functional Zone Mapping

To validate the setting superiority of the zoom-free characteristic and scene-in-scene sample layout of our MEET dataset, we conduct experiments on urban functional zone mapping (UFZ). In the pilot application, UFZ aims to predict

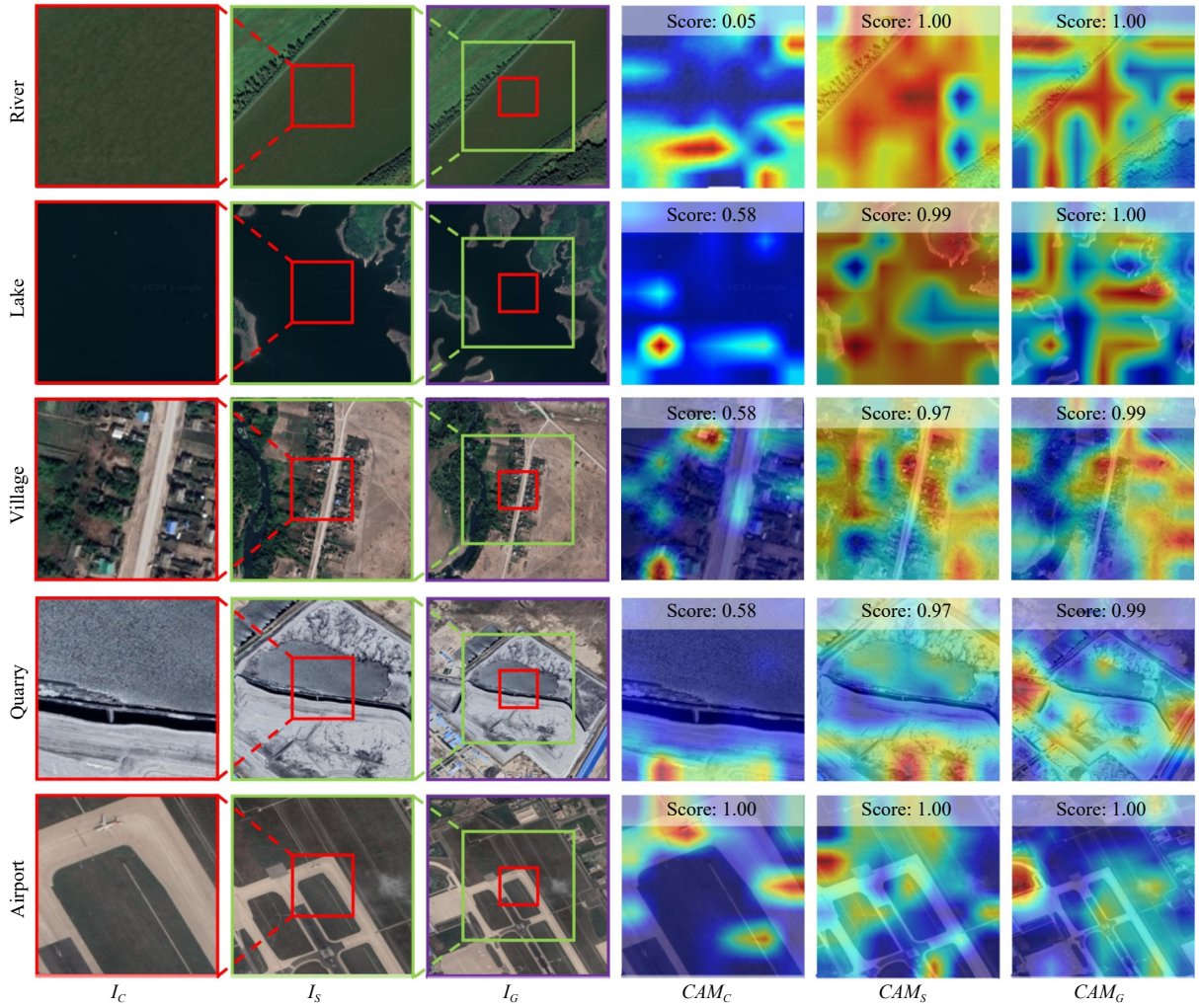


Fig. 13. Visualization of the samples using CAM with our CAT. The scores refers to the prediction probability on the ground truth category after applying different amounts of contextual features. Changes in the score reflect the gain in performance due to the accumulation of multi-level context.

the land-use category of each fixed-resolution RSI block and considers 8 land-use categories: the residential (Res.) category denotes various types of residential buildings of different heights; the commercial (Com.) category indicates commercial and business activities including offices, retail and malls; the industrial (Ind.) category denotes the land with industrial purposes; the transportation (Tra.) category include various transportation facilities; the educational (Edu.) category denotes educational institutions including universities, colleges and primary and secondary schools; the medical (Med.) category is primarily dedicated to healthcare facilities; the sport and cultural (Spo.) category indicates sports and cultural activities including sports fields and art centers; the park and greenspace (Par.) category consists of parks, forests and other public green spaces. Across five cities (e.g., Shanghai, Lanzhou, Wuhan, Guangzhou and Yulin), we create a UFZ evaluation dataset where one large region in each city is selected and its corresponding 1-meter spatial resolution RSI is split into blocks with 256×256 pixels. A total of 4323 blocks are manually annotated by experts in remote sensing with the above 8-class land-use classification system. To avoid ambiguous annotation, we only choose semantically clear blocks for manual labeling, which results in a relatively

sparse annotation distribution. The specific annotation information is summarized in Table V.

To verify the practicability and superiority of the given MEET dataset, we train models on MEET and the other widely adopted datasets such as AID and NWPU to address UFZ. As far as models, RVSA [70] and MTP [71] are selected as they are the state-of-the-art models on AID and NWPU, respectively. During inference, both the RVSA model trained on AID and the MTP model trained on NWPU adopt the center block as input for classification. To facilitate the unified evaluation, we map the classification results of models trained on AID, NWPU and MEET into the 8-class land-use classification system.

The evaluation results of different models trained on AID, NWPU, and MEET are shown in Table VI. The experimental results indicate that the combination of MEET and CAT achieves the best performance in most categories, with a significant improvement over the other combinations. This improvement comes from both the dataset and the method. From the dataset perspective, MEET has more fine-grained categories, enabling the model to extract more detailed semantic information from complex urban environments, leading to a more comprehensive understanding of different UFZ cate-

TABLE V
NUMBER OF ANNOTATED BLOCKS IN THE UFZ EVALUATION DATASET

Location	Res.	Com.	Ind.	Tra.	Edu.	Med.	Spo.	Par.	All
Wuhan	710	6	182	10	391	4	29	241	1573
Shanghai	784	34	89	119	69	4	38	110	1247
Guangzhou	196	43	107	18	155	4	165	393	1081
Lanzhou	95	9	5	0	26	2	11	78	226
Yulin	129	2	2	6	49	4	7	7	196
Total	1914	94	385	143	690	18	250	829	4323

TABLE VI
PERFORMANCE (%) COMPARISON OF DIFFERENT EXPERIMENTAL SETTINGS ON UFZ

Dataset	Method	Res.	Com.	Ind.	Tra.	Edu.	Med.	Spo.	Par.	OA	BA
AID	RVSA	52.87	51.06	71.43	42.66	26.67	0.00	10.40	91.07	54.61	43.27
NWPU	MTP	11.08	63.83	70.91	81.82	0.00	0.00	83.60	52.47	30.21	45.46
MEET	CAT	93.73	59.57	66.49	88.11	58.26	50.00	90.00	90.83	83.76	74.63

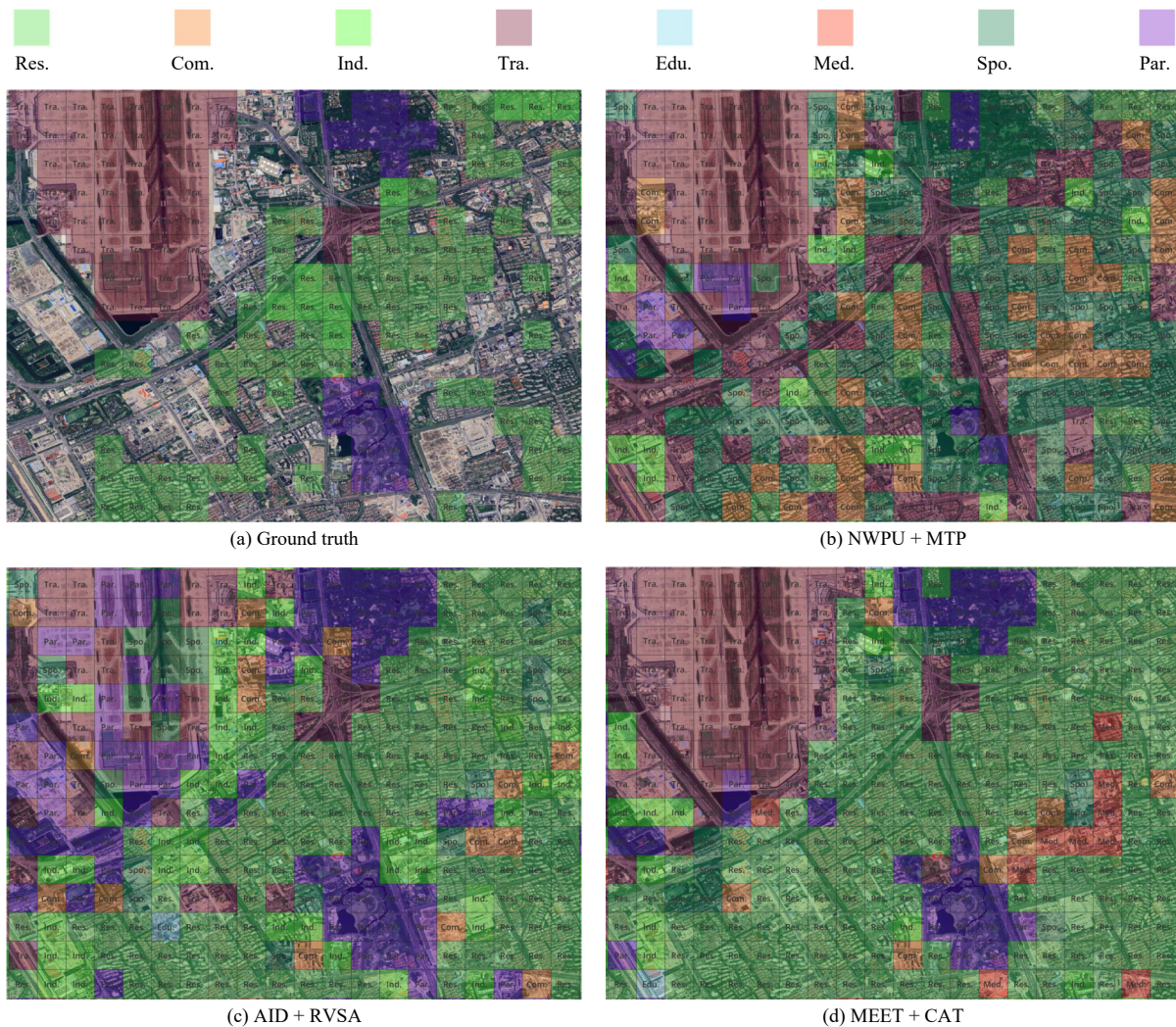


Fig. 14. Illustration of the mapping results of different combinations of dataset and model on the pilot area of Shanghai. The displayed image is a sub-region within the study area of Shanghai.

gories. From the method perspective, CAT can adaptively integrate auxiliary scenes, providing more stable classification performance for cases where the center scene is not very

salient. Furthermore, Figs. 14 and 15 illustrate the mapping results under various settings from two different cities. It is evident that our CAT, when trained on the MEET dataset,

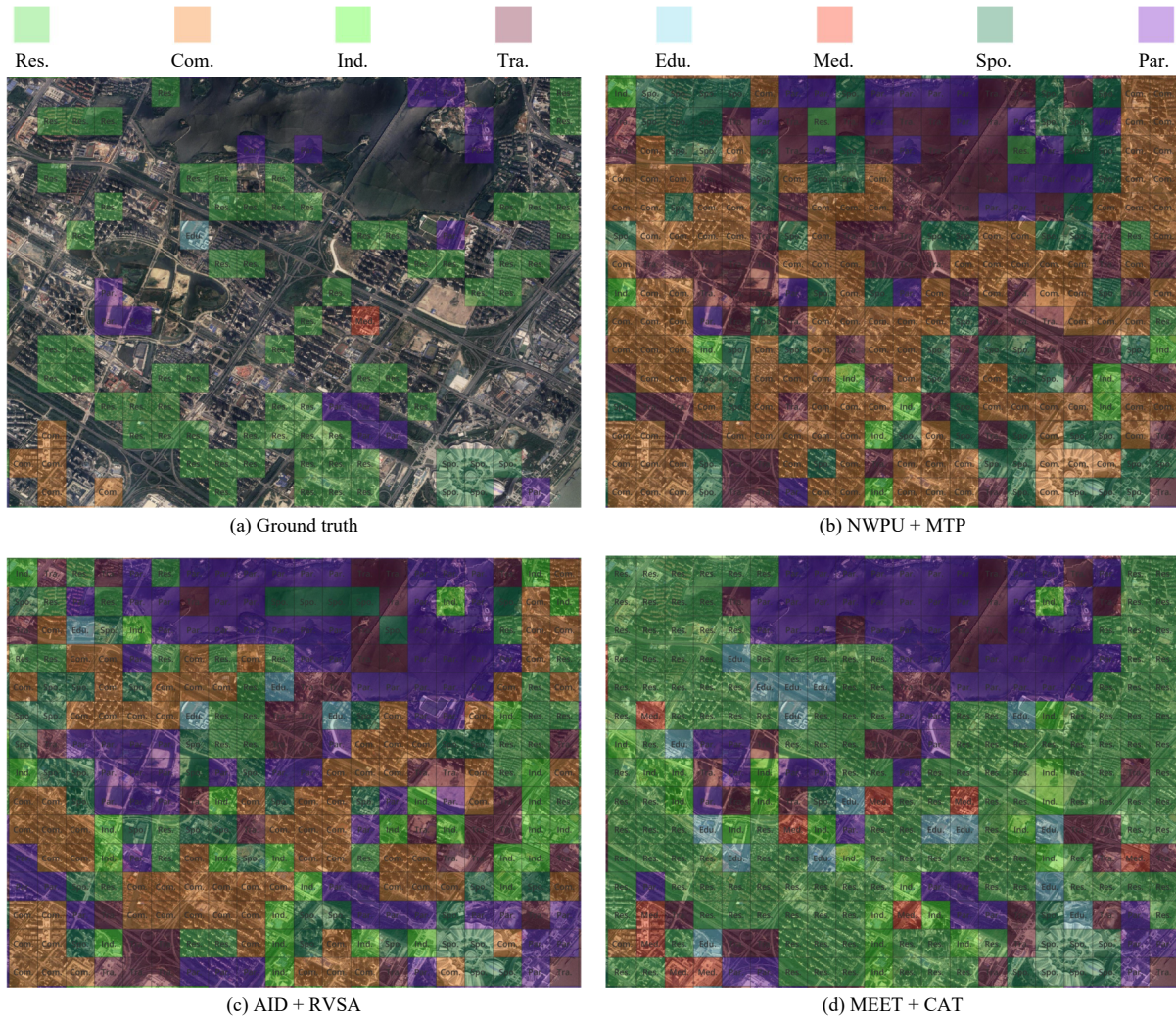


Fig. 15. Illustration of the mapping results of different combinations of dataset and model on the pilot area of Wuhan. The displayed image is a sub-region within the study area of Wuhan.

demonstrates a substantial performance improvement in mapping accuracy and geographical coherence. It is noted that our CAT can effectively utilize contextual information to improve classification performance even in areas with low saliency. In Fig. 14, the CAT with MEET shows better classification capabilities for objects such as parks, airports, and residential buildings, thanks to the effective use of contextual information. In Fig. 15, the CAT with MEET demonstrates significantly better classification performance for the low-salient categories of Educational and Medical, thanks to the MEET dataset's more comprehensive and rich subclass samples for UFZ categories. As a whole, these comparisons underscore the effectiveness of the zoom-free scene-in-scene sample layout in MEET.

VI. CONCLUSION

In this paper, we introduce a large dataset named MEET for FGSC with zoom-free RSI. MEET is comprised of over 1.03 million samples with scene-in-scene layout, encompassing 80 fine-grained geospatial scene categories. Samples are collected globally and include multi-level spatial context information. The large sample size, the granularity of categories, and the inclusion of spatial context imagery make MEET a

valuable dataset. It provides essential data conditions for advancing a challenging yet meaningful new task, FGSC with zoom-free RSI. Additionally, we propose a CAT for FGSC, which effectively integrates contextual information and achieves progressive visual feature extraction. Compared with existing state-of-the-art algorithms, our CAT performs excellently on the MEET dataset, demonstrating the CAT's value from both quantitative and qualitative perspectives.

In future work, we plan to further enrich the MEET dataset in terms of sample volume and category diversity and propose global-scale scene classification mapping products. These will be made available to a broader scientific community in need of related analytical data, thereby continuously driving progress in this research area.

REFERENCES

- [1] 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, Sept. 2019.
- [2] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2030–2045, 2021.
- [3] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene

- classification by gated bidirectional network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
- [4] 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, Nov. 2015.
 - [5] S. R. Phinn, C. M. Roelfsema, and P. J. Mumby, “Multi-scale, object-based image analysis for mapping geomorphic and ecological zones on coral reefs,” *Int. J. Remote Sens.*, vol. 33, no. 12, pp. 3768–3797, Jun. 2012.
 - [6] N. B. Mishra and K. A. Crews, “Mapping vegetation morphology types in a dry savanna ecosystem: Integrating hierarchical object-based image analysis with random forest,” *Int. J. Remote Sens.*, vol. 35, no. 3, pp. 1175–1198, Feb. 2014.
 - [7] Z. Yang, H. Yu, M. Feng, W. Sun, X. Lin, M. Sun, Z.-H. Mao, and A. Mian, “Small object augmentation of urban scenes for real-time semantic segmentation,” *IEEE Trans. Image Process.*, vol. 29, pp. 5175–5190, 2020.
 - [8] P. Gamba, “Human settlements: A global challenge for EO data processing and interpretation,” *Proc. IEEE*, vol. 101, no. 3, pp. 570–581, Mar. 2013.
 - [9] 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, Dec. 2011.
 - [10] G. Cheng, L. Guo, T. Zhao, J. Han, H. Li, and J. Fang, “Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA,” *Int. J. Remote Sens.*, vol. 34, no. 1, pp. 45–59, Jan. 2013.
 - [11] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, “Classification for high resolution remote sensing imagery using a fully convolutional network,” *Remote Sens.*, vol. 9, no. 5, p. 498, May 2017.
 - [12] 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, Feb. 2020.
 - [13] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, “Swin transformer V2: Scaling up capacity and resolution,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, USA, 2022, pp. 11999–12009.
 - [14] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu, H. He, J. Wang, J. Chen, M. Yang, Y. Zhang, and Y. Li, “SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, USA, 2024, pp. 27662–27673.
 - [15] 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, 2021.
 - [16] 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, Oct. 2017.
 - [17] 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.
 - [18] F. Hu, W. Yang, J. Chen, and H. Sun, “Tile-level annotation of satellite images using multi-level max-margin discriminative random field,” *Remote Sens.*, vol. 5, no. 5, pp. 2275–2291, May 2013.
 - [19] Y. Li, X. Huang, and H. Liu, “Unsupervised deep feature learning for urban village detection from high-resolution remote sensing images,” *Photogramm. Eng. Remote Sens.*, vol. 83, no. 8, pp. 567–579, Aug. 2017.
 - [20] Y. Huang, F. Zhang, Y. Gao, W. Tu, F. Duarte, C. Ratti, D. Guo, and Y. Liu, “Comprehensive urban space representation with varying numbers of street-level images,” *Comput. Environ. Urban Syst.*, vol. 106, p. 102043, Dec. 2023.
 - [21] C. Xiao, J. Zhou, J. Huang, H. Zhu, T. Xu, D. Dou, and H. Xiong, “A contextual master-slave framework on urban region graph for urban village detection,” in *Proc. IEEE 39th Int. Conf. Data Engineering*, Anaheim, USA, 2023, pp. 736–748.
 - [22] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, “A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data,” *Remote Sens. Environ.*, vol. 270, p. 112830, Mar. 2022.
 - [23] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th SIGSPATIAL Int. Conf. Advances in Geographic Information Systems*, San Jose, USA, 2010, pp. 270–279.
 - [24] 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 Symp.—100 Years ISPRS*, Vienna, Austria, 2010, pp. 298–303.
 - [25] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, “DeepSat: A learning framework for satellite imagery,” in *Proc. 23rd SIGSPATIAL Int. Conf. Advances in Geographic Information Systems*, Seattle, USA, 2015, pp. 37.
 - [26] O. A. B. Penatti, K. Nogueira, and J. A. Dos Santos, “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, Boston, USA, 2015, pp. 44–51.
 - [27] L. Zhao, P. Tang, and L. Huo, “Feature significance-based multibag-of-visual-words model for remote sensing image scene classification,” *J. Appl. Remote Sens.*, vol. 10, no. 3, p. 035004, Jul. 2016.
 - [28] Q. Zhu, Y. Zhong, B. Zhao, G.-S. 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, Jun. 2016.
 - [29] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu, “High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective,” *Remote Sens.*, vol. 9, no. 7, p. 725, Jul. 2017.
 - [30] P. Helber, B. Bischke, A. Dengel, and D. Borth, “EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.
 - [31] W. Zhou, S. Newsam, C. Li, and Z. Shao, “PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
 - [32] 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, Feb. 2019.
 - [33] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, “BigearthNet: A large-scale benchmark archive for remote sensing image understanding,” in *Proc. IEEE Int. Geoscience and Remote Sensing Symp.*, Yokohama, Japan, 2019, pp. 5901–5904.
 - [34] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao, “RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data,” *Sensors*, vol. 20, no. 6, p. 1594, Mar. 2020.
 - [35] 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. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.
 - [36] H. Li, H. Jiang, X. Gu, J. Peng, W. Li, L. Hong, and C. Tao, “CLRS: Continual learning benchmark for remote sensing image scene classification,” *Sensors*, vol. 20, no. 4, p. 1226, Feb. 2020.
 - [37] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, “Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, Sept. 2021.
 - [38] Y. Hua, L. Mou, P. Jin, and X. X. Zhu, “MultiScene: A large-scale dataset and benchmark for multiscale recognition in single aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5610213, 2022.
 - [39] J. Yuan, L. Ru, S. Wang, and C. Wu, “WH-MAVS: A novel dataset and deep learning benchmark for multiple land use and land cover applications,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 1575–1590, 2022.
 - [40] F. Fang, L. Zeng, S. Li, D. Zheng, J. Zhang, Y. Liu, and B. Wan, “Spatial context-aware method for urban land use classification using street view images,” *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 1–12, Oct. 2022.

- [41] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sept. 2019.
- [42] H. C. Wittich, M. Seeland, J. Wäldchen, M. Rzanny, and P. Mäder, "Recommending plant taxa for supporting on-site species identification," *BMC Bioinformatics*, vol. 19, no. 4, p. 190, May 2018.
- [43] Y. Li, L. Wang, T. Wang, X. Yang, J. Luo, Q. Wang, Y. Deng, W. Wang, X. Sun, H. Li, B. Dang, Y. Zhang, Y. Yu, and J. Yan, "STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 1832–1849, Mar. 2025.
- [44] W. Yu, P. Zhou, S. Yan, and X. Wang, "InceptionNeXt: When inception meets ConvNeXt," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, USA, 2024, pp. 5672–5683.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, pp. 770–778.
- [46] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, pp. 5686–5696.
- [47] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Proc. 17th European Conf. Computer Vision*, Tel Aviv, Israel, 2022, pp. 459–479.
- [48] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "DaViT: Dual attention vision transformers," in *Proc. 17th European Conf. Computer Vision*, Tel Aviv, Israel, 2022, pp. 74–92.
- [49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 9992–10002.
- [50] Y. Liu, J. Zhou, W. Qi, X. Li, L. Gross, Q. Shao, Z. Zhao, L. Ni, X. Fan, and Z. Li, "ARC-Net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020.
- [51] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5620214, 2022.
- [52] W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, and L. Wang, "GCSANet: A global context spatial attention deep learning network for remote sensing scene classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 1150–1162, 2022.
- [53] H. Song, Y. Yuan, Z. Ouyang, Y. Yang, and H. Xiang, "Quantitative regularization in robust vision transformer for remote sensing image classification," *Photogramm. Rec.*, vol. 39, no. 186, pp. 340–372, Jun. 2024.
- [54] 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. 14680–14707, Nov. 2015.
- [55] 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, Oct. 2017.
- [56] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
- [57] K. Xu, H. Huang, P. Deng, and G. Shi, "Two-stream feature aggregation deep neural network for scene classification of remote sensing images," *Inf. Sci.*, vol. 539, pp. 250–268, Oct. 2020.
- [58] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space-frequency joint representation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7492–7502, Oct. 2019.
- [59] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu, "Neural plasticity-inspired foundation model for observing the earth crossing modalities," arXiv preprint arXiv: 2403.15356, 2024.
- [60] S. Srivastava, J. E. Vargas Muñoz, S. Lobry, and D. Tuia, "Fine-grained landuse characterization using ground-based pictures: A deep learning solution based on globally available data," *Int. J. Geogr. Inf. Sci.*, vol. 34, no. 6, pp. 1117–1136, Jun. 2020.
- [61] Y. Yao, X. Yan, P. Luo, Y. Liang, S. Ren, Y. Hu, J. Han, and Q. Guan, "Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 106, p. 102664, Feb. 2022.
- [62] C. Arbingner, M. Bullin, and A. Henrich, "Exploiting geodata to improve image recognition with deep learning," in *Proc. Web Conf.*, Lyon, France, 2022, pp. 648–655.
- [63] Y. Li, W. Chen, X. Huang, Z. Gao, S. Li, T. He, and Y. Zhang, "MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Sci. China Inf. Sci.*, vol. 66, no. 4, p. 140305, Mar. 2023.
- [64] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, pp. 8916–8925.
- [65] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang, "CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, USA, 2020, pp. 8887–8896.
- [66] Y. Li, J. Luo, Y. Zhang, Y. Tan, J.-G. Yu, and S. Bai, "Learning to holistically detect bridges from large-size VHR remote sensing imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 11507–11523, Dec. 2024.
- [67] Z. Bai, G. Li, and Z. Liu, "Global-local-global context-aware network for salient object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 184–196, Apr. 2023.
- [68] Y. Liu, S. Shi, J. Wang, and Y. Zhong, "Seeing beyond the patch: Scale-adaptive semantic segmentation of high-resolution remote sensing imagery based on reinforcement learning," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Paris, France, 2023, pp. 16822–16832.
- [69] L. Zhang, Z. Tan, G. Zhang, W. Zhang, and Z. Li, "Learn more and learn useful: Truncation compensation network for semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 4403814, 2024.
- [70] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 5607315, 2023.
- [71] D. Wang, J. Zhang, M. Xu, L. Liu, D. Wang, E. Gao, C. Han, H. Guo, B. Du, D. Tao, and L. Zhang, "MTP: Advancing remote sensing foundation model via multitask pretraining," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 11632–11654, 2024.



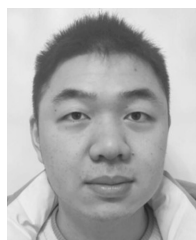
Yansheng Li (Senior Member, IEEE) received the B.S. degree in information and computing science from Shandong University in 2010, and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology in 2015. He is currently a Full Professor and Vice Dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has authored more than 100 peer-reviewed papers such as *IEEE TPAMI*, *RSE*, *IEEE TIP*, *CVPR*, *ECCV* and *AAAI*. His research interests include knowledge graph, deep learning and their applications in remote sensing big data mining. He is an Associate Editor of *IEEE TGRS* and a Junior Editorial Member of *The Innovation*.



Yuning Wu received the B.S. degree in computer science and technology from Wuhan University in 2022. He is currently pursuing the M.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include remote sensing scene classification and few-shot learning.



Gong Cheng (Member, IEEE) received the B.S. degree in biomedical engineering from Xidian University in 2007, and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from Northwestern Polytechnical University in 2010 and 2013, respectively. He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and pattern recognition.



Chuge Zhang is currently pursuing the B.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include remote sensing image segmentation and model compression.



Chao Tao received the B.S. degree from the School of Mathematics and Statistics, Huazhong University of Science and Technology in 2007, and the Ph.D. degree from the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology in 2012. He is currently an Associate Professor with the School of Geosciences and Info-Physics, Central South University. He has authored more than 30 peer-reviewed articles in international journals from multiple domains, such as

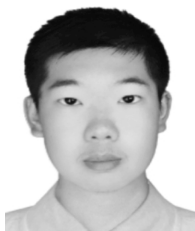
remote sensing and computer vision. His research interests include computer vision, machine learning, deep learning, and their applications in remote sensing. He has been frequently serving as a Reviewer for more than four international journals, including the *IEEE Transactions on Geoscience and Remote Sensing (IEEE-TGRS)*, *IEEE Geoscience and Remote Sensing Letters (IEEE-GRSL)*, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (IEEE-JSTAR)*, *PERS*, and *ISPRS*. He is also a Communication Evaluation Expert for the National Natural Science Foundation of China.



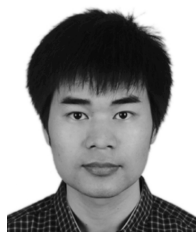
Yiting Liu is currently pursuing the B.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include remote sensing scene classification.



Xu Tang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic circuit and system from Xidian University in 2007, 2010, and 2017, respectively. From 2015 to 2016, he was a Joint Ph.D. Student along with Prof. W. J. Emery with the University of Colorado at Boulder, USA. He is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. His research interests include remote sensing image content-based retrieval and reranking, hyperspectral image processing, remote sensing scene classification, and object detection.



Bo Dang received the B.S. degree in remote sensing science and technology from Wuhan University in 2022. He is currently working toward the Ph.D. degree with the School of Remote Sensing and Information Engineering, Wuhan University. He has published several papers in *CVPR*, *AAAI*, *ISPRS Journal of Photogrammetry and Remote Sensing*, etc. His research interests include remote sensing semantic segmentation and remote sensing foundation model.



Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or coauthored more than 200 refereed journals and conference papers, including *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*/*IEEE Transactions on Image Processing (TIP)*, *International Journal of Computer Vision (IJCV)*, *Computer Vision and Pattern Recognition Conference (CVPR)*, *International Conference on Computer Vision*, and *European Conference on Computer Vision*. His research interests include computer vision, machine learning, and pattern recognition. Dr. Ma has been identified in the 2019–2022 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion* and an Associate Editor of *Neurocomputing*, *Sensors*, and *Entropy*.



Yu Wang received the B.S. degree at the School of Remote Sensing and Information Engineering, Wuhan University in 2023. He is currently pursuing his M.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include deep learning and knowledge graph.



Jiahao Zhang is currently pursuing the B.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include developing unified modeling frameworks from multi-source geospatial data.



Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University in 1997, 2000, and 2002, respectively. He is currently a Full Professor and Dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. His research interests include aerospace and low-altitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote sensing image interpretation, and 3-D city reconstruction.