PCINet: a Prototype- and Concept-based Interpretable Network for Mutli-scene Recognition

Yuansheng Hua[1]*, Jiasong Zhu[1], Qingquan Li[2]

[1] School of Civil and Traffic Engineering, Shenzhen University, Shenzhen, China - (yuansheng.hua, zjsong)@szu.edu.cn
[2] School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China - liqq@szu.edu.cn

Keywords: Aerial image interpretation, Multi-scene recognition, Network interpretability, Concept bottleneck.

Abstract

With the development of remote sensing techniques, a large number of high-resolution aerial images is now available and benefit many applications. Multi-scene recognition plays a key role in applying remote sensing images to these applications, which refers to predicting multiple scenes coexisted in an aerial image and has attracted an increasing attention. Recently, most researchers tend to invent deep learning-based recognition models and has gained great achievements. However, few efforts have been deployed to explaining the success of deep neural networks in multi-scene recognition. To address this, we introduce concept bottleneck model (CBM) to interpreting model performance and propose a novel network, namely Prototype- and Concept-based Interpretable Network (PCINet), that projects aerial imagery into a prototype-concept memory bank and encode their correlations for explaining how a network can identify coexisting scenes in an aerial image. Specifically, the proposed network mainly consists of two branches: prototype matching that measures similarity scores between image features and scene prototypes, and concept bottleneck branches that aligned image features to textual embeddings and compute their relations with concept embeddings. Afterwards, Outputs are integrated for inferring scene categories. Experimental results show that the model enhances interpretability, providing valuable insights for urban planning and resource management, thereby bridging the gap between deep learning models and practical applications.

1. Introduction

With the development of remote sensing techniques, a large number of high-resolution aerial images is now available and beneficial to many applications, e.g., urban planning (Marmanis et al., 2018, Fang et al., 2023), traffic monitoring (Mou and Zhu, 2018, Mou and Zhu, 2016) and natural resource management (Du et al., 2022, Qiu et al., 2019, Weng et al., 2018). As a bridge between imagery and applications, multi-scene recognition that refers to inferring multiple scenes coexisted in an aerial image has now attracted an increasing attention. Recently, most researchers tend to invent deep learning-based recognition models and has gained great achievements (Long et al., 2021, Zheng et al., 2022). However, few efforts have been deployed to explaining the success of deep neural networks in multi-scene recognition. To address this, we introduce concept bottleneck model (CBM) (Koh et al., 2020) to interpreting model performance and propose a novel network that projects aerial imagery into a prototype-concept memory bank and encode their correlations for explaining how a network can identify coexisting scenes in an aerial image. Afterwards, these correlations are fed to a decision layer for scene classification.

2. Methodology

Our proposed model, called Prototype- and Concept-based Interpretable Network (PCINet), mainly consists of two branches: prototype matching that measures similarity scores between image features and scene prototypes, and concept bottleneck branches that aligned image features to textual embeddings and compute their relations with concept embeddings (cf. Figure 2). Afterwards, Outputs are integrated for inferring scene categories.

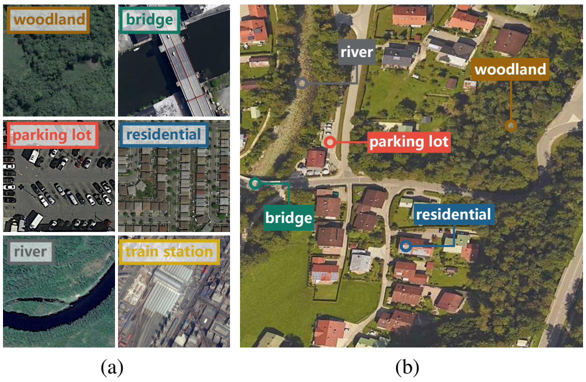


Figure 1. Comparisons between (a) single- and (b) multi-scene recogntion. In (a), each aerial image contains one dominant scene, and the task is to classify each image into one scene category. In (b), multiple scenes are present simultaneously in one single image, and they are required to be thoroughly identified. In our case, single-scene images, such as images in (a), are leveraged to learn scene prototypes for inferring scenes in multi-scene images.

2.1 Prototype Matching Branch

Given an aerial image, the prototype matching branch first extracts the feature map $\boldsymbol{X}$ using a convolutional neural network (CNN), denoted as f_ϕ. The feature map $\boldsymbol{X}$ is then compared with a set of predefined scene prototypes $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n]^T$, where N is the number of scene categories and $\boldsymbol{p}_i$ denotes the prototype of the i-th scene. In this work, we follow (Hua et al., 2021a) and generate scene prototypes by first training f_ϕ on a single-scene aerial image dataset and then summarizing features of samples belonging to the i-th scene as its prototype $\boldsymbol{p}_i$. Thus, $\boldsymbol{p}_i$ is expected to be representative of its correspond-

[*] Corresponding author

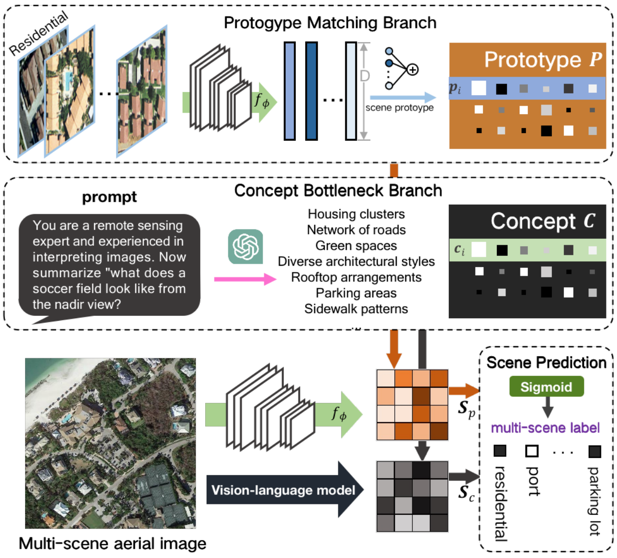


Figure 2. Architecture of the proposed PCINet. It mainly consists of two branches: a prototype matching branch that measures similarity scores between image features and scene prototypes, and a concept bottleneck branch that aligned image features to textual embeddings and compute their relations with concept embeddings. Afterwards, outputs are integrated and fed to the final classification layer for scene prediction.

ing scene (see Figure 5). Afterwards, the similarity score s_i between the feature map and the i-th prototype $\boldsymbol{p}_i$ is computed through a dot product and a softmax function as follows:

$$\boldsymbol{S}_p = \mathrm{softmax}(\mathrm{Q}(\boldsymbol{X}) \cdot \mathrm{K}(\boldsymbol{P})^T), \tag{1}$$

where Q and K are query and key mapping functions stemming from the Transformer (Vaswani et al., 2017).

2.2 Concept Bottleneck Branch

The concept bottleneck branch aims to align the image features with textual embeddings and compute their relations with concept embeddings. One of the crucial steps is to construct the concept bank. To this end, we employ GPT-3.5, know as a powerful large-scale language model, to distill keywords of textual descriptions related to each scene. For example, we send a prompt *You are a remote sensing expert and experienced in interpreting images. Now summarize ”what does a soccer field look like from the nadir view? ” with 10 keywords or phrases.* to GPT-3.5, and it will respond with *Rectangular shape, Green playing surface, Goalposts, White boundary lines, Central circle, Corner flags, Goalkeeper boxes, Spectator stands, Surrounding facilities, Team markings*. Then we compute word embeddings of these concepts and generate concept embeddings $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_m]^T$, where m is the number of concepts. To align image and language features, we employ a vision-language model pretrained with Contrastive Language Image Pretraining (CLIP) techniques (Radford et al., 2021). Specifically, an aerial image is first transformed into textual space, yielding a concept feature map $\boldsymbol{X}_{text}$. The concept feature map $\boldsymbol{X}_{text}$ is then compared with predefined concept embeddings. The output of the concept bottleneck branch, $\boldsymbol{S}_c$, is computed with Eq. 1 but replacing $\boldsymbol{X}$ and $\boldsymbol{P}$ with $\boldsymbol{X}_{text}$ and $\boldsymbol{C}$.

2.3 Scene Prediction

Afterwards, the outputs from the prototype matching branch and the concept bottleneck branch are integrated and fed to the final classification layer to make the final prediction. with the following equation:

$$y = g([S_p \cdot V_p(\boldsymbol{P}), S_c \cdot V_c(\boldsymbol{C})]), \tag{2}$$

where g is the classification layer, and V_p and V_c are value mapping functions for prototype matching and concept bottleneck branches, respectively. By doing so, we can interpret network decisions by figuring out prototypes and concepts with the highest scores.

3. Experimental Results

We generate scene prototypes by training CNNs on single-scene aerial image datasets, i.e., UCM (Yang and Newsam, 2010) and AID (Xia et al., 2017) datasets, and evaluate the performance of our model on the MAI dataset (Hua et al., 2021b), which is specifically designed for multi-scene recognition. Quantitative and qualitative results are reported for analysis and discussion.

3.1 Dataset Description and Configuration

MAI dataset. The MAI dataset includes 3923 large-scale images collected from Google Earth Imagery that covers the United States, Germany, and France. Each is manually assigned one or more of 24 predefiend scene categories: runway, apron, baseball, beach, commercial, farmland, woodland, parking lot, port, residential, river, sea, bridge, lake, park, roundabout, golf course, stadium, train station, works, soccer field, sparse shrub, storage tanks, and tennis court. The size of each image is 512 × 512 pixels, and the spatial resolution ranges from 0.3 to 0.6 m/pixel.

UCM dataset. The UCM dataset is a widely used collection of single-scene aerial images developed by Yang and Newsam at the University of California Merced. These images, totaling 2100 in number, are extracted from aerial ortho imagery provided by the United States Geological Survey (USGS) National Map. They have a spatial resolution of one foot and each image measures 256 × 256 pixels. The dataset covers a diverse range of scenes, with 21 scene-level classes including overpass, forest, beach, baseball diamond, building, airplane, freeway, intersection, harbor, golf course, runway, agricultural land, storage tank, mobile home park, medium residential area, sparse residential area, chaparral, river, tennis courts, dense residential area, and parking lot. Each scene category consists of 100 aerial images. For the purpose of training and validating the embedding function to derive scene prototypes from these images, we randomly allocate 80% of the samples from each scene category for training and validation, reserving the remaining 20% for testing.

AID dataset. The AID dataset is another widely used collection of single-scene aerial images, comprising 10,000 images with dimensions of 600 × 600 pixels. These images are sourced from Google Earth imagery covering various regions including China, the United States, England, France, Italy, Japan, and Germany. The spatial resolutions of the images range from 0.5 m/pixel to 8 m/pixel. The dataset encompasses a total of 30 scene categories, such as viaduct, river, baseball field, city center, farmland, railway station, meadow, bare land, storage tanks, beach, mountain, park, bridge, playground, church, commercial

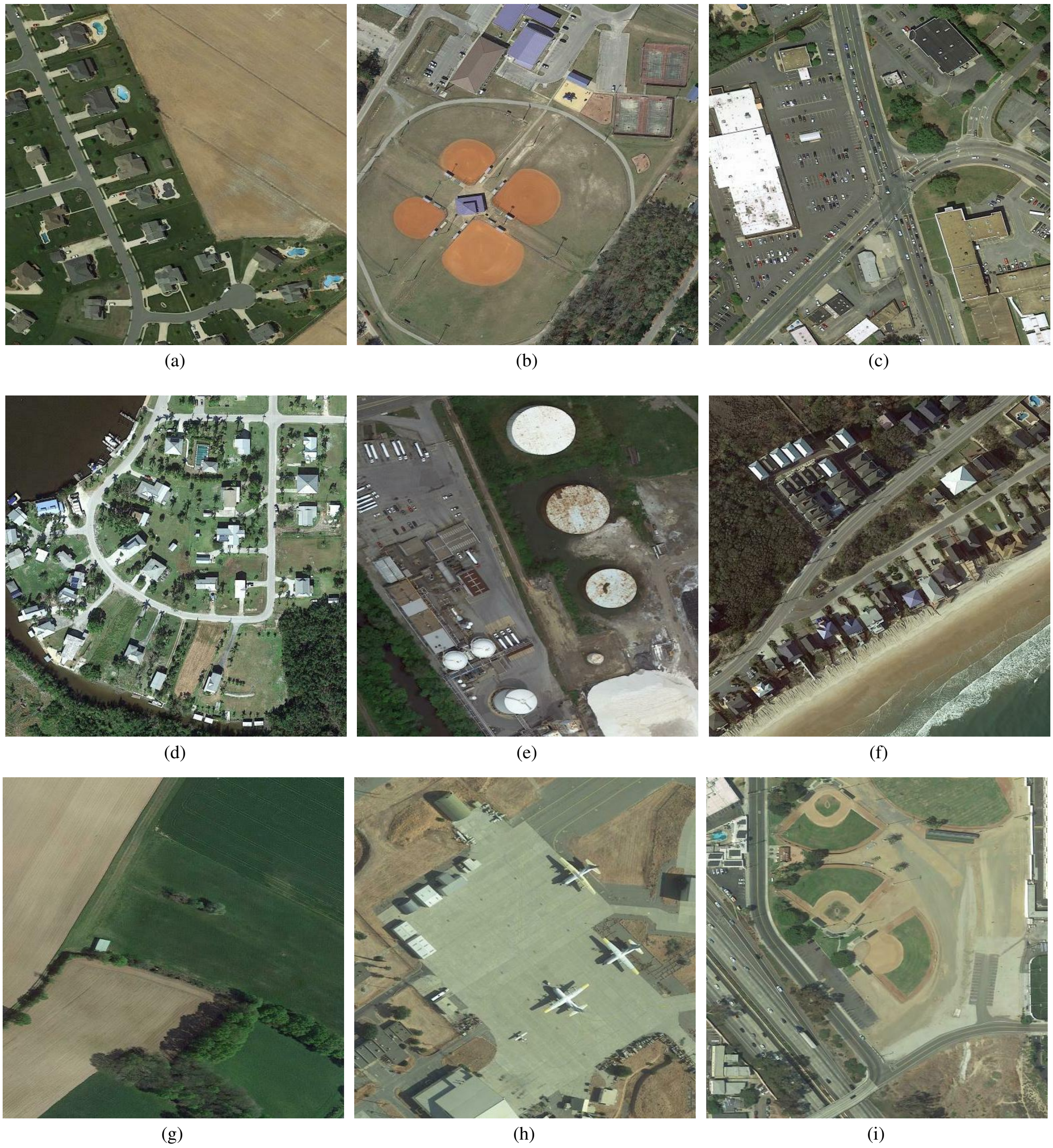

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 3. Example images in our MAI dataset. Each image is 512×512 pixels, and their spatial resolutions range from 0.3 m/pixel to 0.6 m/pixel. We list their scene-level labels here: (a) farmland and residential; (b) baseball, woodland, parking lot, and tennis court; (c) commercial, parking lot, and residential; (d) woodland, residential, river, and runway; (e) river and storage tanks; (f) beach, woodland, residential, and sea; (g) farmland, woodland, and residential; (h) apron and runway; (i) baseball field, parking lot, residential, bridge, and soccer field.

area, desert, forest, parking lot, industrial area, town square, sparse residential area, pond, medium residential area, port, resort, airport, school, stadium, and dense residential area. The number of images varies across categories, ranging from 220 to 420. Similar to the UCM dataset, we adopt a data split approach where 20% of images from each scene category are allocated as test samples, while the remaining images are used for training and validation of the embedding function.

Dataset configuration. In order to widely evaluate the performance of our method, we utilize two variant dataset configurations, MAI-UCM and MAI-AID, based on common scene categories shared by UCM/AID and MAI. Specifically, the MAI-UCM configuration consists of 1600 single-scene aerial images from the UCM dataset and 1649 multi-scene images from our MAI dataset. 16 aerial scenes that are commonly included in both two datasets are considered in UCM2MAI, and numbers of their associated images are listed in Table 1. Besides, the MAI-AID configuration is composed of 7050 and 3239 aerial images from the AID and MAI datasets, respectively. 20 common scene categories are taken into consideration, and the number of images related to each scene is present in Table 1. Although such configurations might limit the number of recognizable scene classes, we believe this limitation can be addressed by collecting more single-scene images by crawling OSM data and producing large-scale multi-scene aerial image datasets. We select only 90 and 120 multi-scene aerial images

	UCM2MAI		AID2MAI	
Scene Category	UCM	MAI	AID	MAI
apron	100	194	360	54
baseball field	100	75	220	235
beach	100	94	400	130
commercial	100	607	350	1391
farmland	100	680	370	983
woodland	100	762	250	1312
parking lot	100	708	390	1777
port	100	3	380	9
residential	200	958	700	2082
river	100	209	410	686
storage tanks	100	89	360	193
sea	100*	51	400*	59
golf course	100	75	-	-
runway	100	230	-	-
sparse shrub	100	336	-	-
tennis court	100	114	-	-
bridge	-	-	360	878
lake	-	-	420	756
park	-	-	350	638
roundabout	-	-	420	281
soccer field	-	-	370	302
stadium	-	-	290	136
train station	-	-	260	9
works	-	-	390	186
All	1600	1649	7050	3239

* indicates that the number of images is not counted in total amounts, as the scene prototype of `beach` and `sea` are learned from the same images.

Table 1. The Number of Images Associated with Each Scene.

from MAI-UCM and MAI-AID as training instances, respectively, and test networks on the remaining multi-scene images. For rare scenes (e.g., port and train station), we select all associated training images, while for common scenes, we randomly select several of their training samples. It is noteworthy that we yield the scene prototype of `residential` by taking an average of high-level representations of aerial images belonging to scene `medium residential` and `dense residential`. Besides, although the UCM and AID datasets do not contain images for `sea`, their images for `beach` often comprise both sea and beach. Therefore, we make use of training samples labeled as `beach` to yield the prototype representation of `sea`.

3.2 Concept Generation

To construct a comprehensive and precise initial set of scene concepts, this project proposes to adopt a concept generation approach based on large language models (LLMs) and prompt engineering. By designing prompt paradigms, we aim to guide LLMs to simultaneously retrieve training sample corpora covering a wide range of contexts and online expert knowledge bases with strong timeliness, thus generating an initial set of concepts describing scene appearance, compositional structure, functional purposes, adjacent features, and more. In the process of prompt design, we first specify the system roles undertaken by the LLM and clarify the purposes, contents, and formats of the questions and answers. Next, we engage in multi-round question-and-answer sessions to establish a model thinking chain. To enhance the accuracy and robustness of model outputs, we will employ active-prompt techniques, where we calculate the uncertainty of the model's multiple responses to the same prompt and supplement the model's thinking chain by manually retrieving relevant corpora for prompts with low confidence in the answers, repeating this process until the model produces highly confident results. Finally, we summarize the LLM's responses to prompts from different angles on the same scene, manually filtering out highly irrelevant concepts to construct an initial set of scene concepts with rich descriptive dimensions and high semantic confidence.

3.3 Training Details

The training process involves two phases: 1) learning the embedding function f_ϕ using a large dataset of single-scene aerial images, and 2) training the entire PCINet using a limited number of multi-scene images in an end-to-end fashion. Different training strategies are applied to each phase, detailed as follows.

During the initial training phase, we initialize the feature extraction modules with CNNs pre-trained on ImageNet (Deng et al., 2009). We utilize crossentropy as the loss function and employ Nesterov Adam (Dozat, n.d.) as the optimizer, with recommended parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 08$. The initial learning rate is set to $2e - 04$ and decayed by $\sqrt{0.1}$ if the validation loss does not decrease for two consecutive epochs.

In the subsequent training phase, we initialize f_ϕ with the parameters learned in the previous phase and use a Glorot uniform initializer to initialize all weights in Q_h, V_h, K_h, and the final fully-connected layer. We set L and U to 256, and the number of heads to 20. All weights are trainable, and the embedding function is fine-tuned during this phase as well. Scene-level labels are encoded as multi-hot vectors, where 0 indicates the absence of a scene and 1 indicates its presence. The loss function is defined as binary cross-entropy. The optimizer remains the same as in the initial phase, but we use a relatively larger learning rate of $5e - 4$. The network is implemented using TensorFlow and trained on a single NVIDIA Tesla P100 16GB GPU for 100 epochs. We set the training batch size to 32 for both phases.

3.4 Evaluation Metrics

To quantitatively evaluate network performance, we employ example-based F_1(Wu and Zhou, 2016) and F_2(Van Rijsbergen, 1979) scores as evaluation metrics. These scores are calculated using the following equation:

$$F_\beta = (1 + \beta^2)\frac{p_e r_e}{\beta^2 p_e + r_e}, \quad \beta = 1, 2, \tag{3}$$

where p_e and r_e represent example-based precision and recall (Tsoumakas and Vlahavas, 2007), which are computed as:

$$p_e = \frac{TP_e}{TP_e + FP_e}, \tag{4}$$

$$r_e = \frac{TP_e}{TP_e + FN_e}, \tag{5}$$

where TP_e, FP_e, and FN_e indicate the numbers of true positives, false positives, and false negatives, respectively, within each example. Each example in our case corresponds to a multi-scene aerial image. By averaging scores across all examples in the test set, we can determine the mean example-based F scores, precision, and recall. Additionally, we calculate label-based precision p_l and recall r_l using Eq. 4 and Eq. 5, respectively, but substituting the counts of false negatives, false positives, and true positives specific to each scene category. The mean p_l and r_l are then computed. It's worth noting that the primary metrics of interest are the mean F_1 and F_2 scores.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-265-2024 | © Author(s) 2024. CC BY 4.0 License.

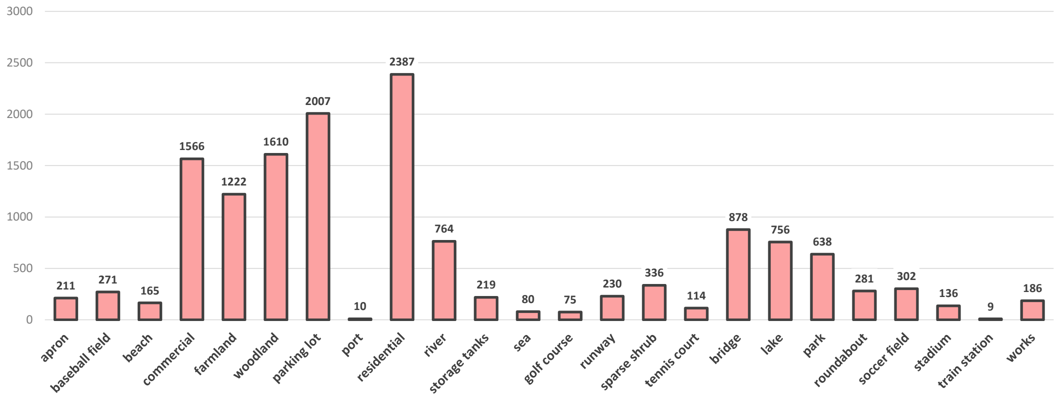
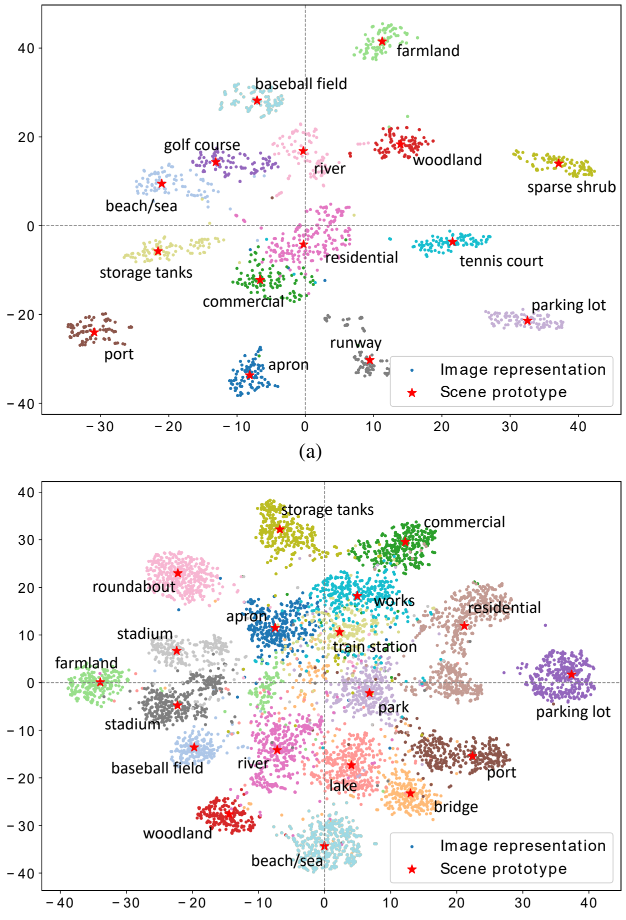


Figure 4. Sample distributions of all scene categories in the MAI dataset.



(a)

(b)

Figure 5. T-SNE visualization of image representations and scene prototypes learned by CNN (e.g., VGGNet) on (a) UCM and (b) AID datasets, respectively. Dots in the same color represent features of images belonging to the same scene, and stars denote scene prototypes.

3.5 Results

We report the results of our experiments in terms of accuracy, precision, recall, and F1-score. Our model achieves better performance, and correlations between images, prototypes and concepts are visualized to illustrate the decision process.

4. Conclusion

In conclusion, PCINet, with its dual branches integrating prototypes and concepts, achieves superior performance in unconstrained scene recognition for high-resolution aerial images. The model enhances interpretability, providing valuable insights for urban planning and resource management, thereby bridging the gap between deep learning models and practical applications.

References

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dozat, T., n.d. Incorporating Nesterov momentum into Adam. `http://cs229.stanford.edu/proj2015/054_report.pdf`. Online.

Du, S., Xing, J., Li, J., Du, S., Zhang, C., Sun, Y., 2022. Open-Pit Mine Extraction from Very High-Resolution Remote Sensing Images Using OM-DeepLab. *Natural Resources Research*, 31(6), 3173–3194.

Fang, H., Guo, S., Zhang, P., Zhang, W., Wang, X., Liu, S., Du, P., 2023. Scene Change Detection by Differential Aggregation Network and Class Probability-Based Fusion Strategy. *IEEE Transactions on Geoscience and Remote Sensing*.

Hua, Y., Mou, L., Lin, J., Heidler, K., Zhu, X. X., 2021a. Aerial Scene Understanding in The Wild: Multi-Scene Recognition via Prototype-based Memory Networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177, 89–102.

Hua, Y., Mou, L., Lin, J., Heidler, K., Zhu, X. X., 2021b. Aerial scene understanding in the wild: Multi-scene recognition via prototype-based memory networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177, 89–102.

Koh, P., Nguyen, T., Tang, Y., Mussmann, S., Pierson, E., Kim, B., Liang, P., 2020. Concept bottleneck models. *International Conference on Machine Learning*, 5338–5348.

Long, Y., Xia, G., Li, S., Yang, W., Yang, M., Zhu, X. X., Zhang, L., Li, D., 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 4205–4230.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135, 158–172.

Mou, L., Zhu, X. X., 2016. Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis. *2016 IEEE international geoscience and remote sensing symposium (IGARSS)*, IEEE, 1823–1826.

Mou, L., Zhu, X. X., 2018. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11), 6699–6711.

Qiu, C., Mou, L., Schmitt, M., Zhu, X. X., 2019. Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154, 151–162.

Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Tsoumakas, G., Vlahavas, I., 2007. Random K-labelsets: An ensemble method for multilabel classification. *European Conference on Machine Learning (ECML).*

Van Rijsbergen, C. J., 1979. *Information Retrieval*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Weng, Q., Mao, Z., Lin, J., Liao, X., 2018. Land-use scene classification based on a CNN using a constrained extreme learning machine. *International journal of remote sensing*, 39(19), 6281–6299.

Wu, X., Zhou, Z., 2016. A unified view of multi-label performance measures. *arXiv:1609.00288.*

Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. *International Conference on Advances in Geographic Information Systems (SIGSPATIAL).*

Zheng, Z., Zhong, Y., Su, Y., Ma, A., 2022. Domain adaptation via a task-specific classifier framework for remote sensing cross-scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13.