# Multi-View Urban Scene Classification with a Complementary-Information Learning Model

Wanxuan Geng, Weixun Zhou, and Shuanggen Jin


## Abstract

*Traditional urban scene-classification approaches focus on images taken either by satellite or in aerial view. Although single-view images are able to achieve satisfactory results for scene classification in most situations, the complementary information provided by other image views is needed to further improve performance. Therefore, we present a complementary information-learning model (CILM) to perform multi-view scene classification of aerial and ground-level images. Specifically, the proposed CILM takes aerial and ground-level image pairs as input to learn view-specific features for later fusion to integrate the complementary information. To train CILM, a unified loss consisting of cross entropy and contrastive losses is exploited to force the network to be more robust. Once CILM is trained, the features of each view are extracted via the two proposed feature-extraction scenarios and then fused to train the support vector machine classifier for classification. The experimental results on two publicly available benchmark data sets demonstrate that CILM achieves remarkable performance, indicating that it is an effective model for learning complementary information and thus improving urban scene classification.*


## Introduction

With the rapid development of remote sensing technology, traditional pixel-level image analysis has been unable to meet the needs of high-level image-content interpretation due to increasing spatial resolution, and urban scene classification has therefore been a hot topic in the remote sensing field (Zhou *et al.* 2018). Scene classification is assigning a specific label to each image according to its content (Kang *et al.* 2020), providing relatively high-level interpretation of a remote sensing image compared with pixel- and object-based classification (Xia *et al.* 2017). It is a practical application of high-resolution remote sensing image processing, which can provide data support for land planning and utilization (K. Xu *et al.* forthcoming), and is widely used in urban functional zoning planning (Huang *et al.* 2018), natural-disaster monitoring (Attari *et al.* 2018), and object detection (Schilling *et al.* 2018). Though the literature has developed a large number of scene-classification approaches—including handcrafted methods and ones based on deep learning—which can achieve remarkable performance, there are still problems to be solved.


Wanxuan Geng and Weixun Zhou are with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China (zhouwx@nuist.edu.cn).

Shuanggen Jin is with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China; and the China and Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai, China.


*Contributed by Alper Yilmaz, August 30, 2021 (sent for review August 30, 2021.*

On one hand, a high-resolution remote sensing image has rich spatial information and a complex background, making it difficult to extract powerful features for scene classification (T. Tian *et al.* 2021), and accordingly results in worse performance. On the other hand, most of the existing scene-classification approaches focus on images taken from a single view, such as satellite or aerial, but it has been demonstrated that the complementary information provided by other views is able to further improve classification performance (Machado *et al.* 2021), as shown in Figure 1. It is notable that scene classification of an aerial image can benefit from the complementary information provided by a ground-level image, and vice versa. For instance, we cannot obtain the correct classification result of an airport unless both aerial and ground-view images are exploited. In recent work by Machado *et al.* (2021), early and late fusion based on a convolutional neural network (CNN) are exploited to perform multi-view scene classification. More specifically, the early fusion is conducted by fusing the convolutional features of each view via a concatenation layer, whereas the late fusion is conducted by combining the prediction result of each view achieved by an individual CNN. Both early and late fusion have been proven effective for scene classification, but for early fusion, the concatenation layer is inserted in the first several convolutional layers, which cannot integrate the high-level features of each view image. For late fusion, an individual CNN must be trained for the prediction of each view image, and the training process is time-consuming and totally separated. We therefore raise the question: *Is it possible to learn complementary information via feature-level fusion and perform multi-view classification using a single CNN framework*?

Inspired by cross-view geo-localization (Vo and Hays 2016; T. Tian *et al.* 2021), in this article we extend our previous work (Geng *et al.* 2021) and propose a complementary information-learning model (CILM) for multi-view urban scene classification of aerial and ground-level images. The proposed CILM is a two-branch network trained using a unified loss to enhance the performance. Once CILM is trained, the high-level features of each view image are extracted and then combined to train a support vector machine (SVM) classifier to perform the final prediction. It should be noted that our work is different from that of Machado *et al.* (2021) in that, although both approaches take aerial and ground-level image pairs as input, for Machado *et al.* aerial and ground-level images in each pair are from the same location and the same class, whereas we ignore the location and the class of image pairs. Therefore, we explored how the information provided by pairs of images from different locations can benefit urban scene classification. Also, in our work, CILM is regarded as a feature extractor for extracting high-level features of each view image, which is not exploited for prediction. And we train an SVM classifier


Photogrammetric Engineering & Remote Sensing
Vol. 88, No. 1, January 2022, pp. 65–72.
0099-1112/22/65–72




doi: 10.14358/PERS.21-00062R2

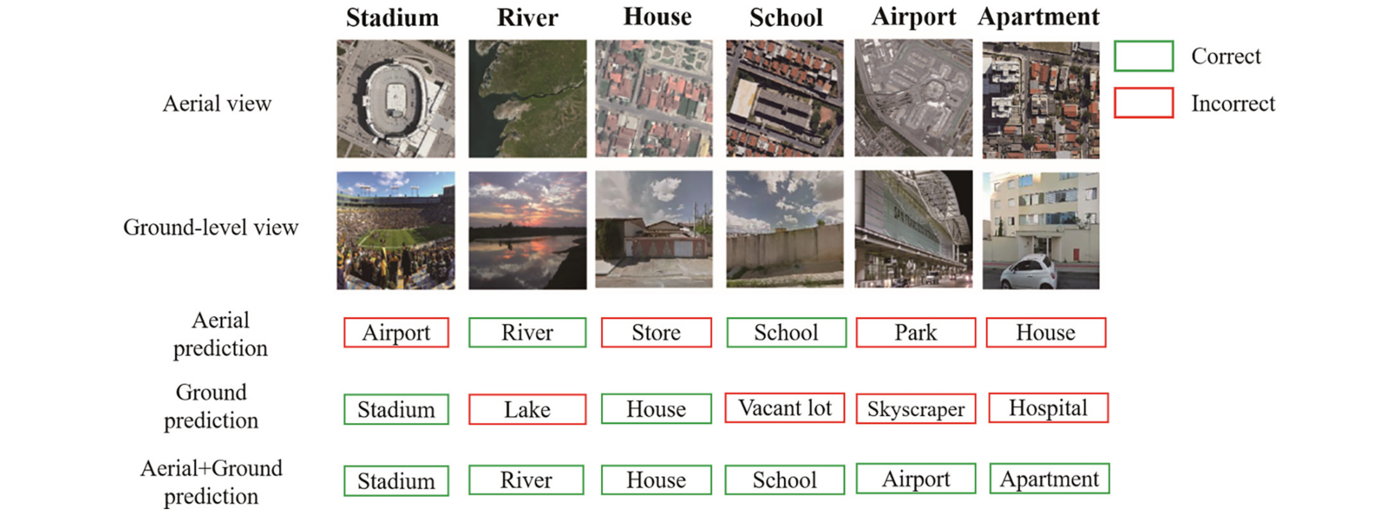

Figure 1. Visual examples of single- and multi-view scene classification.

using the fused high-level features to integrate complementary information for classification, which has been demonstrated to outperform the softmax classifier for scene classification (Xia *et al.* 2017).

In summary, the main contributions of this article are as follows.

- We propose a complementary information-learning model trained with a unified loss to integrate complementary information for multi-view scene classification of aerial and ground-level images. The unified loss is composed of cross entropy and contrastive losses, where the cross-entropy loss is to distinguish the class of each view image in the pair and to identify whether the input is a matched pair (i.e., aerial and ground-level images belonging to the same class) and the contrastive loss is to pull matched pairs closer and push unmatched pairs away in the feature space.
- We explore two pretrained CNNs as the basic network to construct CILM, which is then evaluated on two publicly available benchmark data sets with various experimental configurations, thus providing baseline results for future research.

The remainder of this paper is organized as follows. The next section reviews related work on urban scene classification. The proposed CILM is introduced in detail in the section after that, and then the experimental setup and results presented. Finally, we give a brief conclusion.

## Related Work

In this section, we briefly review the work on scene classification and cross-modal methods for the processing of multi-view images.

### Scene Classification

Traditional remote sensing scene classification is based on handcrafted low- and middle-level features. The low-level features are either global features, such as the color histogram (Swain and Ballard 1991), texture features (Haralick *et al.* 1973), and gist (Oliva and Torralba 2001), or local features, such as the famous scale-invariant feature transformation (Lowe 2004). In contrast, middle-level features establish the relationship with semantics through statistical-distribution analysis of low-level features; bag of visual words (Mansoori *et al.* 2013) is one of the representative methods, commonly used for classification tasks (Okumura *et al.* 2011). In recent years, methods based on deep learning have been widely exploited for scene classification, since CNNs outperform their counterpart traditional approaches on ImageNet (Krizhevsky *et al.* 2012), and have become the most popular approaches for image recognition since then. Zhou *et al.* (2017) proposed using a three-layer perceptron and a couple of convolutional layers to construct a low-dimensional CNN for remote sensing image retrieval. Han *et al.* (2017) integrated the pretrained AlexNet with spatial pyramid pooling and side supervision to improve scene-classification performance. Bian *et al.* (2017) proposed a simple yet effective saliency-patch sampling method to extract image regions that are the most informative.

Since effective and discriminative feature representation plays an important role in classification results (Zhang *et al.* 2019), some works focus on how to extract powerful features. Liu *et al.* (2018) rearranged deep features and used discriminative convolution filters with different kernel sizes for scene classification. Xu *et al.* (2020) used the transferred VGG16 to extract the multi-layer convolutional features and added several layers to process hierarchical features in different branches, which can improve performance; whereas Liu *et al.* (2018) combined spatial pyramid pooling with deep CNNs and designed a multiple-kernel learning strategy to fuse multi-scale features.

Though these handcrafted and particularly CNN feature-based methods have achieved significant success for scene classification, their data sources are single-view satellite or aerial images; whether the complementary information provided by other view images can benefit scene classification has not been explored.

### Cross-Modal Approaches for Multi-View Images

A cross-modal network, as its name implies, is trained using more than one kind of data, and is a commonly used approach to process images of different views simultaneously. In work by X. Xu *et al.* (2015), the earliest cross-modal network was presented for image and text retrieval, which supports searching across multi-modal data and thus is suitable for remote sensing data (X. Xu *et al.* 2017). T. Tian *et al.* (2021) proposed an effective framework of cross-view matching for geo-localization in urban environments. Khokhlova *et al.* (2020) introduced a multi-modal network across time that learns to retrieve by content vertical aerial images of French urban and rural territories taken about 15 years apart. Xiong *et al.* (2020) proposed a novel deep cross-modality hashing network for cross-modal content-based remote sensing image retrieval between synthetic aperture radar and optical sensors. Feng *et al.* (forthcoming) proposed a framework for multi-view spectral–spatial feature extraction and fusion for analysis and classification of hyperspectral images. Xu *et al.* (2020) used hand-drawn sketches describing mental pictures to retrieve the desired targets in large-scale remote sensing images.

Differentiating our work here, most of the existing cross-modal works are essentially image matching to determine whether the input pairs are matched, such as the problem of image retrieval and geo-localization. The function of CILM, on the other hand, is to integrate the complementary information provided by each view image and then perform scene classification of multi-view images, which is a more difficult task than image matching.

## Methodology

This section presents our methodology. We first introduce the architecture of the proposed CILM, then describe the unified loss used to train the network.

### The Architecture of CILM

Our CILM consists of two identical subnetworks and three additional fully connected (FC) layers, as shown in Figure 2. The subnetwork is a CNN pretrained on ImageNet and contains convolution, pooling, and FC layers. CILM takes positive and negative image pairs as input, where a positive image pair is assigned the label 1 and a negative image pair is assigned the label 0. For positive image pairs, the aerial and ground-level images are from the same class, whereas for negative image pairs, they are from different classes.

During training, the aerial and ground-level images in a pair are each fed into one of the two subnetworks. The output feature vectors from each subnetwork are combined through a subtraction operation and the result is passed through the additional FC layer $FC_{ag}$, with a single output. We use a sigmoid function to convert this output value to a probability between 0 and 1, indicating the prediction of whether the input pairs are matched or unmatched. The first loss L1 is used for this task during training.

Relating to the other two additional FC layers, both $FC_a$ and $FC_g$ convert the 4096-D feature vectors from the subnetworks to $N$-D feature vectors, where $N$ is the number of scene categories. Therefore, $FC_a$ is used for aerial scene classification, whereas $FC_g$ is used for ground-level scene classification. The motivation here is to force CILM to be more robust by using single-view image classification, which has been proven effective for scene classification (X. Liu *et al.* 2019). The second loss $L_{2a}$ and $L_{2g}$ are used for aerial and ground-level classification, respectively, during training.

The discriminative feature representation is significant for scene classification (Cheng *et al.* 2018); we therefore use the third loss L3 to learn powerful features. This is a ranking loss that can pull matched pairs closer and push unmatched pairs away in the feature space.

Once CILM is trained, we propose two scenarios to extract feature vectors to train the SVM classifier for classification, since SVM has been demonstrated to be more effective than the softmax classifier. More specifically, for the first scenario we extract features (i.e., $f_a$ and $f_g$) from the last FC layers of the subnetworks, whereas for the second scenario we extract features (i.e., $f_a'$ and $f_g'$) from $FC_a$ and $FC_g$. The extracted features are then fused to a feature vector through an addition operation.

### Loss for CILM

The unified loss is exploited to update CILM during training. The unified loss $L_U$ is defined as

$$L_U = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 \quad (1)$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are three trade-off parameters that control the importance of these three losses.

$L_1$ is a binary cross-entropy loss defined as

$$L_1 = -q\log(p) - (1 - q)\log(1 - p) \quad (2)$$

$$p = \text{sigmoid}(f_{ag}) \quad (3)$$

where $q$ and $p$ are the ground truth and the predicted label of the input pair, respectively, and $f_{ag}$ is the output value of the $FC_{ag}$ layer.

$L_2$ is a softmax cross-entropy loss consisting of two parts, $L_{2a}$ for aerial-view classification and $L_{2g}$ for ground-level view classification:

$$L_2 = L_{2a} + L_{2g} \quad (4)$$

$$L_{2a} = -\sum_{i=1}^{N} q_i^a \log\left(p_i^a\right) \quad (5)$$

$$L_{2g} = -\sum_{i=1}^{N} q_i^g \log\left(p_i^g\right) \quad (6)$$

where $N$ is the number of scene categories, $q_i^a$ and $p_i^a$ are the ground truth and predicted label of the aerial image, and $q_i^g$ and $p_i^g$ are the ground truth and predicted label of the ground-level image.

$L_3$ is a contrastive loss aiming to compare the similarity between aerial and ground-level images in the pairs:

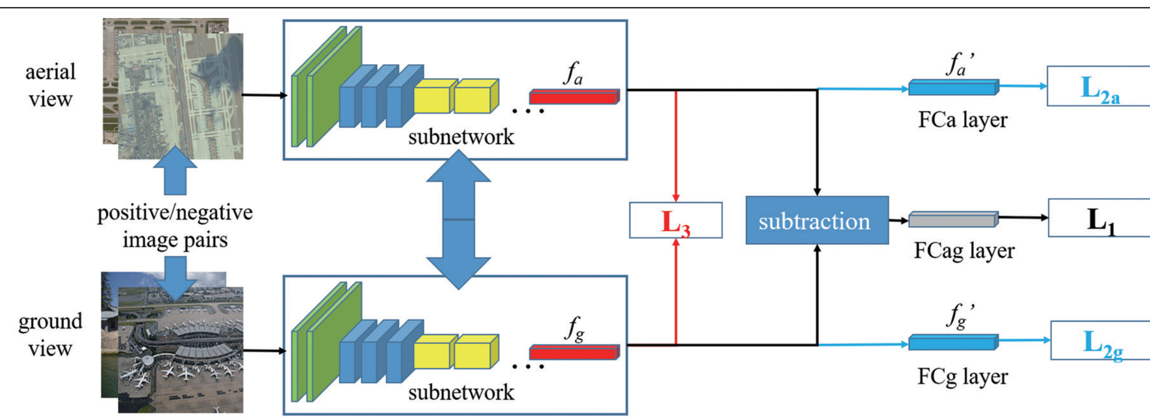

Figure 2. The network architecture of the proposed complementary information-learning model (CILM).

$$L_3 = \frac{1}{2} y d^2 + (1-y)\max(m-d,0)^2 \tag{7}$$

$$d = \|f_a - f_g\|_2 \tag{8}$$

where $y$ is the label of the input pair, $d$ is the Euclidean distance between $f_a$ and $f_g$, and $m$ is the margin parameter used for constraint. If aerial and ground-level images in a pair are similar (i.e., the two images are from the same class), then $d$ should be smaller than $m$; otherwise it is larger.

## Experiments

In this section, we first describe two publicly available benchmark multi-view data sets, and then we introduce the experimental settings for our experiments. Finally, the experimental results and discussions are given.

### Multi-View Data Sets

Our approach is evaluated using two benchmark data sets presented by Machado *et al.* (2021). The first, AiRound, is composed of 11 classes: airport, bridge, church, forest, lake, river, skyscraper, stadium, statue, tower, and urban park (Figure 3). Each class contains images in three distinct perspectives: satellite view, aerial view, and ground-level view. Therefore, each image in AiRound is composed of a triplet, with all three images acquired from the same place. Figure 4 shows some examples of image pairs; in our experiments, we use only the aerial and ground-level view images.

The second data set, CV-BrCT, is composed of approximately 24 000 pairs of images split into nine urban classes: apartment, hospital, house, industrial, parking lot, religious, school, store, and vacant lot (Figure 5). Each class has images in two distinct perspectives: aerial view and ground-level view. The two view images in each pair are also acquired from the same place. Figure 6 shows some examples of image pairs.

### Experimental Setting

As described before, we did not consider whether the aerial and ground-level images in each pair were from the same location or the same class. In our experiments, we construct image pairs by first randomly splitting the images in each

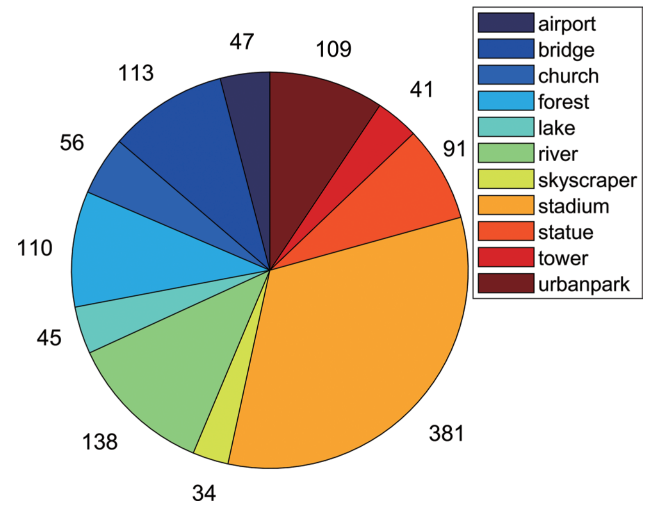


Figure 3. The number of images of each class in the AiRound data set.

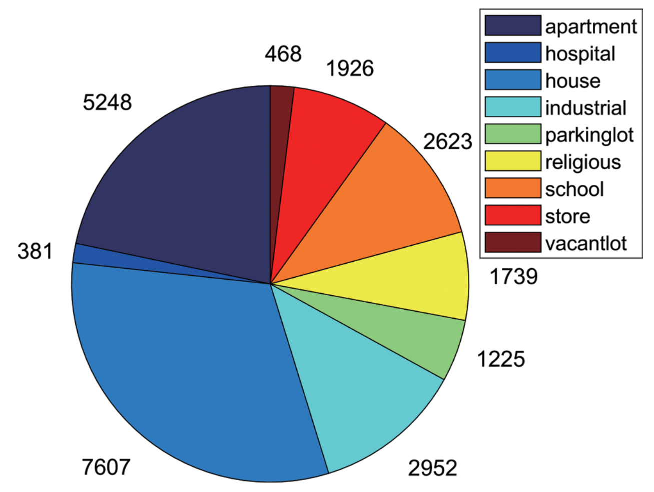


Figure 4. The number of images of each class in the CV-BrCT data set.

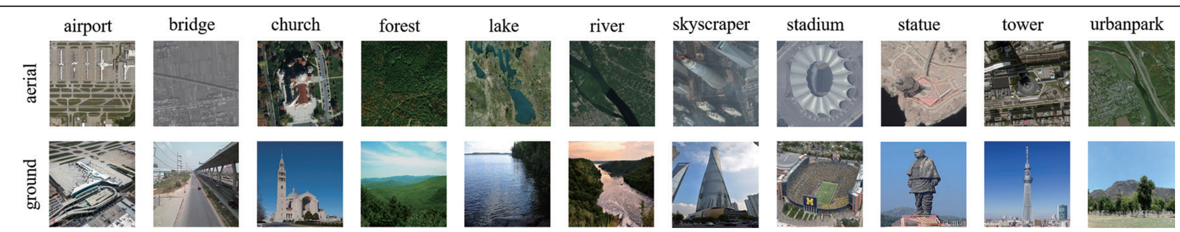


Figure 5. Examples of image pairs taken from the AiRound data set.

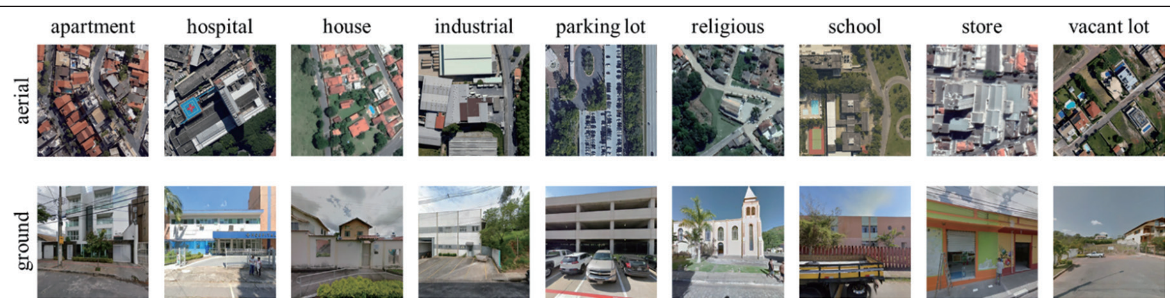


Figure 6. Examples of image pairs taken from the CV-BrCT data set.

Table 1. The training parameters of CILM on two data sets.

| Data Set | Basic Network | Batch Size | Learning Rate | Iteration |
|---|---|---|---|---|
| AiRound | AlexNet | 80 | 0.000 08 | 1000 |
| | VGG16 | 24 | 0.000 08 | 1500 |
| CV-BrCT | AlexNet | 80 | 0.000 08 | 3000 |
| | VGG16 | 24 | 0.000 08 | 5000 |

CILM = complementary information-learning model.

Table 2. The implementation details of single-view classification approaches.

| Method | Implementation Details |
|---|---|
| CILM_1_2 | CILM + $L_1$ and $L_2$ losses + subnetwork + softmax classifier |
| CILM_U | CILM + unified loss + subnetwork + softmax classifier |

CILM = complementary information-learning model.

class as 80% training samples and 20% test samples. Then we group aerial and ground-level images in each class through the method of exhaustion to obtain image pairs.

Regarding CILM, we select AlexNet (Krizhevsky *et al.* 2012) and VGG16 (Simonyan and Zisserman 2015) as the subnetworks, which are famous shallow and deep CNNs, respectively, that have been widely used for image classification. We remove the last FC layers in each to for the subnetworks to output 4096-D feature vectors. During training, the image pairs are resized to 227×227 pixels for AlexNet and 224×224 pixels for VGG16. The Adam optimizer is exploited to minimize the unified loss, where the gradient decay and the squared gradient decay factor are set to 0.9 and 0.99, respectively. The training details of CILM, such as batch size, learning rate, and number of iterations, are shown in Table 1. For the unified loss, we set $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 0.0001$, and $m = 0.3$.

In the following experiments, we conduct single- and multi-view classification to evaluate the performance of CILM using the AiRound and CV-BrCT data sets. The single-view classification is aerial or ground-level classification using the subnetworks in CILM and the pretrained CNNs. Specifically, we evaluate the performance achieved by two CILM-based methods CILM_1_2 and CILM_U. The implementation details are shown in Table 2. Regarding the multi-view classification, we explore CILM with different configurations shown in Table 3.

In addition, CILM is compared to feature fusion and six-channel methods. Unless particularly stated, we extract features from the penultimate FC layer of the pretrained CNN and use SVM for classification.

## Results on AiRound and CV-BrCT

*Single-View Classification Results*

The results of single-view classification obtained by CILM are presented to explore how the complementary information provided by other view images can benefit scene classification. All the results obtained by CILM are presented in Table 4.

For both the AiRound and CV-BrCT data sets, we can see that CILM_U configured with VGG16 (not shared weights) as the subnetworks achieves the best performance for both aerial and ground-level images. In addition, CILM trained other than with shared weights achieves slightly better performance than with shared weights, and VGG16 is a better subnetwork than AlexNet.

*Multi-View Classification Results*

Table 3. The implementation details of multi-view classification approaches.

| Method | | Implementation Details |
|---|---|---|
| CILM_1 | FS | CILM+$L_1$ loss + FS |
| CILM_1_3 | FS | CILM + $L_1$ and $L_3$ losses + FS |
| CILM_1_2 | FS | CILM + $L_1$ and $L_2$ losses + FS |
| | SS | CILM + $L_1$ and $L_2$ losses + SS |
| CILM_U | FS | CILM + unified loss + FS |
| | SS | CILM + unified loss + SS |

Table 4. Single-view classification results of CILM on two data sets.

| | | | Data Set | | | |
|---|---|---|---|---|---|---|
| | | | AiRound | | CV-BrCT | |
| Weights | Subnetworks | Method | Aerial | Ground-level | Aerial | Ground-level |
| Shared | AlexNet | CILM_1_2 | 82.15 | 80.52 | 78.04 | 61.81 |
| | | CILM_U | 82.83 | 81.82 | 78.39 | 62.39 |
| | VGG16 | CILM_1_2 | 83.69 | 81.97 | 79.46 | 62.64 |
| | | CILM_U | 84.98 | 82.40 | 79.52 | 63.30 |
| Not shared | AlexNet | CILM_1_2 | 84.55 | 82.40 | 78.09 | 63.16 |
| | | CILM_U | 84.78 | 82.83 | 79.66 | 63.50 |
| | VGG16 | CILM_1_2 | 84.12 | 82.83 | 79.91 | 63.61 |
| | | CILM_U | **85.83** | **83.26** | **80.37** | **63.72** |

CILM = complementary information-learning model.

Table 5. Multi-view classification results of CILM on the AiRound data set.

| | | Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CILM_1 | | CILM_1_3 | | CILM_1_2 | | CILM_U | |
| Weights | Subnetworks | FS | SS | FS | SS | FS | SS | FS | SS |
| Shared | AlexNet | 87.55 | — | 87.98 | — | 90.56 | 90.99 | 91.42 | 92.27 |
| | VGG16 | 88.41 | — | 88.84 | — | 91.20 | 91.55 | 91.85 | 92.70 |
| Not shared | AlexNet | 89.27 | — | 89.70 | — | 91.38 | 91.70 | 92.06 | 93.13 |
| | VGG16 | **89.70** | — | **90.10** | — | **90.92** | **91.83** | **92.49** | **93.56** |

CILM = complementary information-learning model; FS = first scenario; SS = second scenario.

Table 6. Multi-view classification results of CILM on the CV-BrCT data set.

| | | Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CILM_1 | | CILM_1_3 | | CILM_1_2 | | CILM_U | |
| Weights | Subnetworks | FS | SS | FS | SS | FS | SS | FS | SS |
| Shared | AlexNet | 80.24 | — | 80.50 | — | 80.70 | 81.62 | 81.58 | 82.18 |
| | VGG16 | 81.80 | — | 81.90 | — | 83.62 | 84.06 | 83.97 | 84.24 |
| Not shared | AlexNet | 80.52 | — | 80.60 | — | 81.66 | 82.11 | 82.38 | 82.42 |
| | VGG16 | **82.35** | — | **82.45** | — | **83.66** | **84.09** | **84.15** | **84.32** |

CILM = complementary information-learning model; FS = first scenario; SS = second scenario.

Here we present the results of multi-view classification on the AiRound (Table 5) and CV-BrCT (Table 6) data sets obtained by the proposed CILM with different configurations. It can be observed that CILM_U configured with VGG16 (not shared weights) as the subnetworks and SS as the feature-extraction strategy achieves the best performance for both data sets. The results will be analyzed in detail.

It can be seen that CILMs not trained with shared weights achieve slightly better performance than those with shared weights, except for CILM_1_2 configured with VGG16 and the first scenario for the AiRound data set. The results make sense, since aerial and ground-level images are taken from different perspectives, and thus we can learn view-specific features when the subnetworks do not use shared weights. For the subnetworks, it seems that VGG16 is a better choice than AlexNet, but the performance difference is small. To explore how the proposed unified loss can improve the performance of CILM, we trained CILM using different losses. It is obvious that CILM_U outperforms the other approaches, indicating that the unified loss can benefit multi-view classification. We can also conclude that L2 is the most important among the three losses, according to the results obtained by CILM_1_2, CILM_1_3, and CILM_1. In addition, SS is a more appropriate feature-extraction scenario for CILM. This is because the first scenario extracts 4096-D features from the last FC layers of the subnetworks, whereas the second scenario extracts *N*-D features from the additional FC layers, where the features are class-specific high-level features, thus achieving better performance.

According to the results of multi- and single-view classification, we can conclude that multi-view scene classification can benefit from the complementary information provided by aerial or ground-level images. For AiRound, the best performance is 93.56, whereas the best single-view performance is 85.83 for the aerial view and 83.26 for the ground-level view. With respect to CV-BrCT, the best performance is 84.32, whereas the best single-view performance is 80.37 for the aerial view and 63.72 for the ground-level view. Therefore, multi-view classification improves the results of single-view classification by a significant margin, especially for the ground-level classification of CV-BrCT. This is possibly because the ground-level images in CV-BrCT are more challenging than the aerial images, as shown in Figure 6.

### Feature-Visualization Results

In addition to the single- and multi-view classification results, we also present the visualization results of features extracted by CILM to give a quantitative evaluation, as can be observed in Figures 7 and 8. For the AiRound data set, the features of multi-view images can be easily separated for different classes, whereas for single-view images, most of the image classes are clustered together—except for stadium. Regarding the CV-BrCT data set, we can observe similar results as with AiRound. But an interesting phenomenon is that the features of multi-view images and aerial images achieve similar clustering performance, both outperforming ground-level images by a significant margin. These results make sense, since

Table 7. Performance comparisons of CILM and counterpart approaches for single- and multi-view classification.

| | **Single-View Classification** | | | |
|---|---|---|---|---|
| | **AiRound** | | **CV-BrCT** | |
| **Method** | **Aerial** | **Ground** | **Aerial** | **Ground** |
| CNN-softmax (Simonyan and Zisserman 2015) | 82.84 | 81.55 | 79.18 | 62.12 |
| CNN-SVM (Simonyan and Zisserman 2015) | 80.52 | 80.09 | 69.87 | 54.95 |
| CILM | **85.83** | **83.26** | **80.37** | **63.72** |
| | **Multi-View Classification** | | | |
| **Method** | **AiRound** | | **CV-BrCT** | |
| Feature fusion (Simonyan and Zisserman 2015) | 90.4 | | 74.99 | |
| Six-channel (Vo and Hays 2016) | 70.39 | | 73.46 | |
| CILM | **93.56** | | **84.32** | |

CILM = complementary information-learning model; CNN = convolutional neural network.

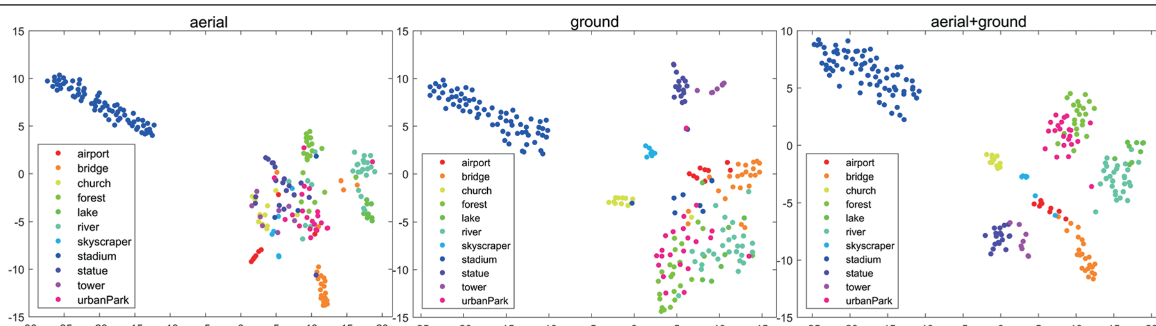


Figure 7. Feature-visualization results of single- and multi-view images in the AiRound data set.

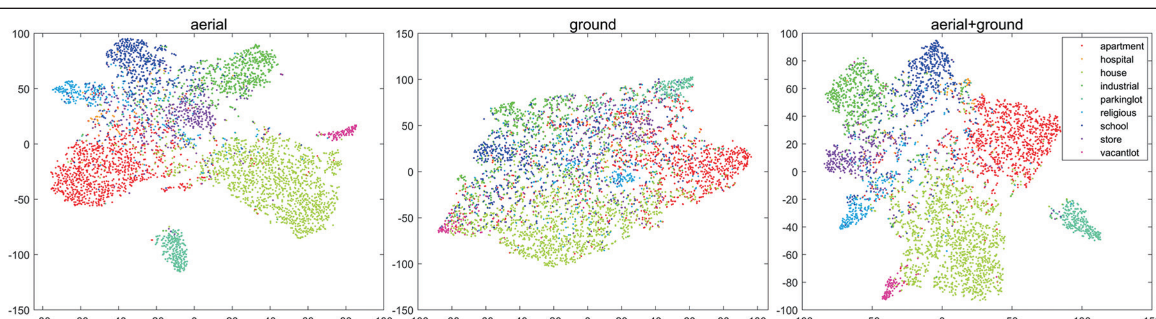


Figure 8. Feature-visualization results of single- and multi-view images in the CV-BrCT data set.

CV-BrCT is more challenging than AiRound, and the ground-level images in CV-BrCT have higher intraclass diversity.

Table 7 shows the comparison results of single- and multi-view classification achieved by CILM and other counterpart approaches on AiRound and CV-BrCT. For the multi-view classification, our method outperforms feature fusion and six-channel methods for both data sets. The six-channel method performs the worst among these approaches; is not as effective, as it was used for geo-localization (Vo and Hays 2016). This is because for image geo-localization, we only need to determine whether the two images are from the same location, whereas for multi-view classification we need to identify the classes of image pairs, which is definitely a more challenging problem. As for the single-view classification, our approach achieves better performance than the two pretrained CNN-based approaches for both data sets.

The confusion matrices of the multi-view results achieved by our approach on AiRound and CV-BrCT are shown in Figures 9 and 10, respectively. For AiRound, the classification accuracy of lake is below 0.8, and around 22% of lake samples are incorrectly classified to rivers due to the high similarity and the imbalanced number of samples between lake and river. Skyscraper also has a lower classification accuracy, due to the small number of samples, and some images are mistakenly classified in other building categories, such as airport and stadium. In addition, urban park is easily confused with forest. For CV-BrCT, the high similarity between different classes and the number of samples has a great influence on the classification accuracy. We can see that the classification accuracy of hospital is only 18%, because hospital is severely confused with apartment.

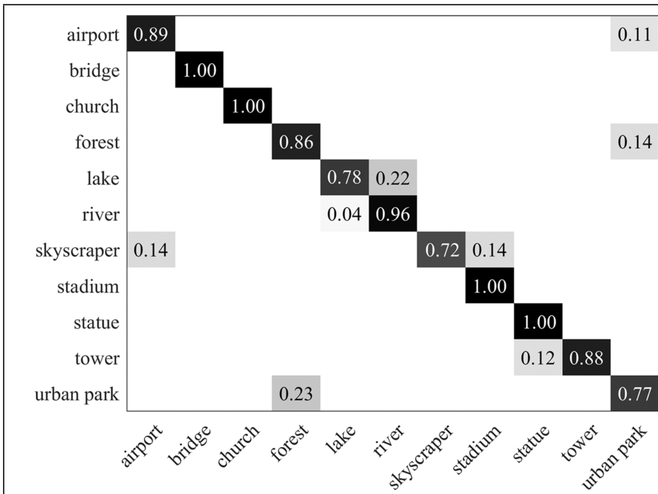


Figure 9. The confusion matrix for the multi-view classification results achieved by CILM on AiRound.

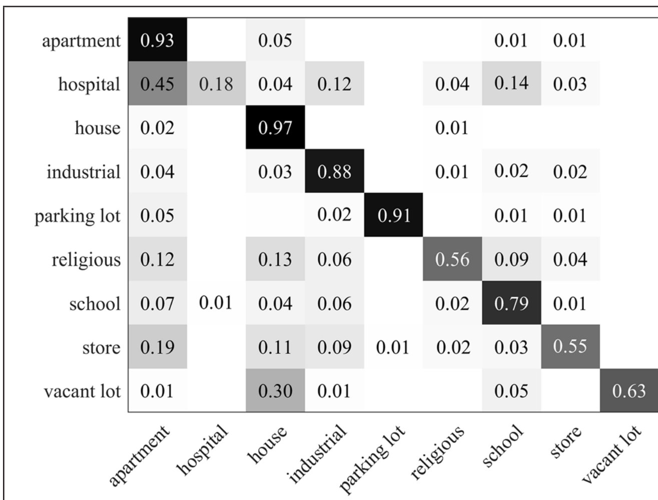


Figure 10. The confusion matrix for the multi-view classification results achieved by CILM on CV-BrCT.

## Conclusion

In this article, we proposed a complementary information-learning model (CILM) for multi-view urban scene classification. To enhance the training of CILM, we exploited a unified loss consisting of two cross-entropy losses and a contrastive loss. Unlike the existing works that use softmax for classification, we extract the high-level features of aerial and ground-level images via two feature-extraction scenarios, and then fuse the features to integrate complementary information to train an SVM for classification. We explored CILM with different configurations of subnetworks, losses, and feature-extraction scenarios to evaluate its performance. The experimental results show that CILM configured with VGG16 (weights not shared) as the subnetworks and the second scenario as the feature-extraction strategy achieves the best performance on both AiRound and CV-BrCT data sets. Further, the comparison results between multi- and single-view classification indicate that the complementary information provided by other view images can benefit scene classification.

## Acknowledgments

This work was supported by the Strategic Priority Research Program Project of the Chinese Academy of Sciences under grant XDA23040100; the National Natural Science Foundation of China under grant 42001285; the Natural Science Foundation of Jiangsu Province, China, under grant BK20200813; the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under grant 20KJB420002; and the Jiangsu Double Innovation Doctor Program under grant R2020SCB58.

The authors would like to thank the anonymous reviewers for their comments to improve the article, and to thank the researchers whose work presents the AiRound and CV-BrCT data sets.

## References

Attari, N., F. Ofli, M. Awad, J. Lucas and S. Chawla. 2017. Nazr-CNN: Fine-grained classification of UAV imagery for damage assessment. Pages 50–59 in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, held in Tokyo, Japan, 19–21 October 2017. Piscataway, NJ: IEEE.

Bian, X., C. Chen, L. Tian and Q. Du. 2017. Fusing local and global features for high-resolution scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10(6):2889–2901.

Cheng, G., C. Yang, X. Yao, L. Guo and J. Han. 2018. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing* 56(5):2811–2821.

Feng, J., J. Zhang and Y. Zhang. Forthcoming. A multiview spectral-spatial feature extraction and fusion framework for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*.

Geng. W, W. Zhou and S. Jin. 2021. Feature fusion for cross-modal scene classification of remote scene image. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIV-M-3-2021:63–66. https://doi.org/10.5194/isprs-archives-XLIV-M-3-2021-63-2021.

Han, X., Y. Zhong, L. Cao and L. Zhang. 2017. Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing* 9(8):848.

Haralick, R. M, K. Shanmugam and I. Dinstein. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics.* SMC-3(6):610–621.

Huang, B., B. Zhao and Y. Song. 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment* 214:73–86.

Kang, J., R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi and A. Plaza. 2020. Deep metric learning based on scalable neighborhood components for remote sensing scene characterization. *IEEE Transactions on Geoscience and Remote Sensing* 58(12):8905–8918.

Khokhlova, M., V. Gouet-Brunet, N. Abadie and L. Chen. 2020. Cross-year multi-modal image retrieval using Siamese networks. Pages 1–5 in *2020 IEEE International Conference on Image Processing (ICIP)*, held in Abhu Dhabi, United Arab Emirates, 25–28 October 2020. Piscataway, NJ: IEEE.

Krizhevsky, A., I. Sutskever and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems.* pp. 1097-1105.

Liu, N., X. Lu, L. Wan, H. Huo and T. Fang. 2018. Improving the separability of deep features with discriminative convolution filters for RSI classification. *ISPRS International Journal of Geo-Information* 7(3):95.

Liu, Q., R. Hang, H. Song and Z. Li. 2018. Learning multiscale deep features for high-resolution satellite image scene classification. *IEEE Transactions on Geoscience and Remote Sensing* 56(1):117–126.

Liu, X., Y. Zhou, J. Zhao, R. Yao, B. Liu and Y. Zheng. 2019. Siamese convolutional neural networks for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters* 16(8):1200–1204.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Machado, G., E. Ferreira, K. Nogueira, H. Oliveira, M. Brito, P. H. T. Gama and J. A. dos Santos. 2021. AiRound and CV-BrCT: Novel multiview datasets for scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14:488–503.

Mansoori, N. S, M. Nejati, P. Razzaghi and S. Samavi. 2013. Bag of visual words approach for image retrieval using color information. Pages 1–6 in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, held in Mashhad, Iran, 14–16 May 2013. Piscataway, NJ: IEEE.

Okumura, S., N. Maeda, K. Nakata, K. Saito, Y. Fukumizu and H. Yamauchi. 2011. Visual categorization method with a Bag of PCA packed Keypoints. Pages 950–953 in *2011 4th International Congress on Image and Signal Processing*, held in Shanghai, China, 15–17 October 2013. Piscataway, NJ: IEEE.

Oliva, A. and A. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Schilling, H., D. Bulatov, R. Niessner, W. Middelmann and U. Soergel. 2018. Detection of vehicles in multisensor data via multibranch convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11(11):4299–4316.

Simonyan, K. and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. Pages 1–14 in *International Conference on Learning Representations (ICLR)*, held in San Diego, California, US, 7-9 May 2015. New York: Elsevier.

Swain, M. J. and D. H. Ballard. 1991. Color indexing. *International Journal of Computer Vision* 7(1):11–32.

Tian, T., L. Li, W. Chen and H. Zhou. 2021. SEMSDNet: A multiscale dense network with attention for remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14:5501–5514.

Tian, Y., C. Chen and M. Shah. 2017. Cross-view image matching for geo-localization in urban environments. Pages 1998–2008 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, held in Honolulu, HI, USA, 21-26 July 2017. Piscataway, NJ: IEEE.

van der Maaten, L. and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Vo, N. N. and J. Hays. 2016. Localizing and orienting street views using overhead imagery. Pages 494–509 in *Computer Vision–ECCV 2016*. Lecture Notes in Computer Science vol. 9905. Cham, Switzerland: Springer.

Xia, G.-S., J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang and X. Lu. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* 55(7):3965–3981.

Xiong, W., Z. Xiong, Y. Zhang, Y. Cui and X. Gu. 2020. A deep cross-modality hashing network for SAR and optical remote sensing images retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:5284–5296.

Xu, F., W. Yang, T. Jiang, S. Lin, H. Luo and G.-S. Xia. 2020. Mental retrieval of remote sensing images via adversarial sketch-image feature learning. *IEEE Transactions on Geoscience and Remote Sensing* 58(11):7801–7814.

Xu, K., H. Huang, P. Deng and Y. Li. Forthcoming. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Transactions on Neural Networks and Learning Systems*.

Xu, K., H. Huang, Y. Li and G. Shi. 2020. Multilayer feature fusion network for scene classification in remote sensing. *IEEE Geoscience and Remote Sensing Letters* 17(11):1894–1898.

Xu, X., F. Shen, Y. Yang, H. T. Shen and X. Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing* 26(5):2494–2507.

Xu, X., Yang Y, Shimada A, Taniguchi R.I and He L. 2015. Semi-supervised coupled dictionary learning for cross-modal retrieval in Internet images and texts. Pages 847–850 in *MM' 15: Proceedings of the 23rd ACM International Conference on Multimedia*, held in Brisbane, Australia, 26–30 October 2015. Edited by Zhou, J. New York: Association for Computing Machinery.

Zhang, B., Y. Zhang and S. Wang. 2019. A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12(8):2636–2653.

Zhou, W., S. Newsam, C. Li and Z. Shao. 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing* 145(Part A):197–209.

Zhou, W., S. Newsam. C. Li and Z. Shao. 2017. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sensing* 9(5):489.