

RESEARCH ARTICLE

Optimizing Remote Sensing Image Scene Classification Through Brain-Inspired Feature Bias Estimation and Semantic Representation Analysis

ZHONG DONG^{1,2}, BAOJUN LIN^{3,4,5,6}, AND FANG XIE^{2,3,4,5,6}¹Department of Automation, Tsinghua University, Beijing 100080, China²Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China³University of Chinese Academy of Sciences, Beijing 100094, China⁴Innovation Academy for Microsatellites, Chinese Academy of Sciences, Shanghai 201210, China⁵Shanghai Engineering Center for Microsatellites, Shanghai 201304, China⁶School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

Corresponding author: Zhong Dong (dongzhong1987@126.com)

ABSTRACT In the realm of remote sensing image classification and detection, deep learning has emerged as a highly effective approach, owing to the remarkable advancements in object perception models and the availability of abundant annotated data. Nevertheless, for specific remote sensing image scene classification tasks, obtaining diverse and large amounts of data remains a daunting challenge, leading to limitations in the applicability of trained models. Consequently, researchers are increasingly focusing on optimal data utilization and interpretability of learning. Drawing inspiration from brain neural perception research, researchers have proposed novel approaches for deeper interpretation and optimization of deep learning models from diverse perspectives. In this paper, we present a brain-inspired network optimization model for remote sensing image scene classification, which considers both shape and texture features and reconstructs feature scaling of data through feature bias estimation. The model achieves greater robustness through complementary training. We evaluate our optimized model on general datasets by integrating it into an existing benchmark method and compare its performance with the original approach. Our results demonstrate that the proposed model is highly effective, with dynamically reconstructed data leading to a significant enhancement of model learning.


INDEX TERMS Remote sensing image, scene classification, brain-inspired learning, feature bias, data enhancement.

I. INTRODUCTION

Remote sensing (RS) images offer a plethora of feature data and high-quality image datasets have seen a rapid growth in recent years. These datasets have become increasingly important in urban planning [1], environmental monitoring [2], and natural disaster monitoring [3]. Effective classification of RS image data is crucial to better utilize the data, and it is an active research area. However, classification of RS images poses great difficulties due to the inclusion of multiple

types of specific targets, different edge information, texture information, and more within the same type of image.

RS image classification can be performed at the pixel-level, object-level, or scene-level. Scene-level classification has been a particularly challenging task due to the difficulty of augmentation and the serious deformation of annotation data. Early RS image classification methods relied on manually designed features, such as scale invariant feature transform (SIFT) [4], histogram of oriented gradients (HOG) [5], histogram of colors (CH) [6]. However, manual features have limited representation capability and weak model migration capability, resulting in low classification efficiency.

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino .

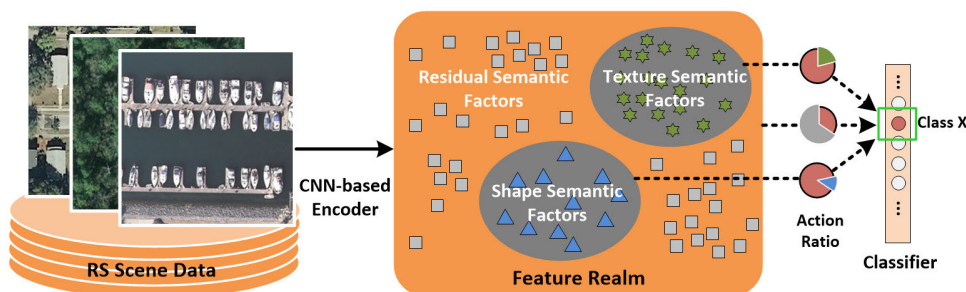


FIGURE 1. Visual representation of semantic representation factors of different types of features in the feature realm. The factors of different semantic types are encoded in feature realm, and the action ratio they play in the classifier differs due to the quantity variance between factor types (without considering the representational capacity). Here we define “Residual” as all the feature types except for texture and shape (e.g. color).

To overcome these limitations, convolutional neural network (CNN) based methods have been proposed and have shown to be more efficient than using low-level features [7]. Since then, a large number of RS image scene classification methods based on deep CNNs have been derived [8], [9], [10]. However, deep learning-based approaches still face some core problems when dealing with complex RS image scenes. First, training a model with satisfactory accuracy requires a large amount of labeled data. Second, it is difficult to know what data features are more effective for the generalization performance of pre-trained models.

In recent years, researchers have proposed various methods to optimize the performance of the CNN model for coping with complex RS image scenes. Chen and Zhu [11] proposed a CNN structure based on contextual spatial attention and dense connectivity, and a method to improve data utilization. The model achieved a competitive classification accuracy in the mountain scene dataset. Wang et al. [12] proposed an adaptive learning strategy for training CNN-based models, which mainly consists of two parts: adaptive training and adaptive labeling. The method optimizes the pre-trained model by adjusting the data side and achieves an effective improvement in classification performance. Shen et al. [13] proposed a method to combine dual model features by bilinear fusion, which improves the scale adaptation of the model and improves the ability of the model to resist the effect of complex background redundancy. These innovative classification efforts for complex scenes do optimize the performance of the models, but there are still some potential problems. First, the training data has difficulty in fully covering situations with interference from factors such as lighting, season, and visibility, which obscures features in the images that would otherwise contribute significantly to the classification, making it difficult for the models to achieve accurate classification of images with drastic changes in certain features (e.g. color). Secondly, the traditional data augmentation methods (e.g. color transformation and symmetry flip) have limited enhancement of texture features, and we can hardly evaluate whether these enhancements have made the maximum contribution to the generalization ability of the model.

To address these issues, we try to explore a data feature distribution that is closer to the human brain in this paper. This research is mainly based on the premise that different types of features do not play the same role in the classification (or recognition) process of different scenarios (or targets), as shown in Fig. 1. We propose the idea that by adjusting the degree of involvement of different feature types in learning, the model can achieve an ideal state for optimal learning. To verify our idea concisely and precisely, shape and texture will be the two feature types focused on in this paper. In summary, the contributions of this paper are as follows.

(1) Proposing an optimization model that enables dynamic reconstruction of classification data through the estimation of feature bias, enhancing the robustness of the classification model while mitigating potential accuracy loss for the classifier due to the inherent feature bias of the encoder and dataset.

(2) Studying the impact of potential semantic representations of shape and texture features in the encoded realm on classifier training and the relationship between the number of different semantic representation neurons (or factors) and classifier performance. Our research provides a new optimization approach to the CNN-based RS scene classification problem and a reference idea in the interpretability problem of deep classification model learning.

II. RELATED WORK

A. BRAIN-INSPIRED LEARNING

Brain-inspired learning emphasizes the crucial importance of identifying the precise image features learned by convolutional neural networks (CNNs) during training. Despite fundamental differences between the human brain and CNNs in vision representation, it is widely accepted that CNNs are the most predictive models for object recognition in the human ventral stream [14], [15]. As a result, the brain-like learning approach for artificial neural networks has gained increasing attention among researchers.

Recent studies have highlighted the categorical topological correspondence between deep CNNs and the brain [16]. Specifically, it has been demonstrated that the early visual

cortex (V1) and early CNN layers encode shape information, while the anterior ventral temporal cortex encodes class information that is best correlated with the final layer of CNNs [17]. Additionally, it has been shown that CNNs rely on local texture and shape features rather than global shape contours, which may explain why lower CNN layers are able to fully capture the representation structure of real-world object images in lower visual regions with smaller receptive domains [18].

Pioneering research has focused on incorporating human brain learning mechanisms into deep learning [19], [20], [21], and on mechanistic and functional comparisons between the two [22], [23], [24]. These studies have contributed to a deeper understanding of the similarities and differences between the human brain and CNNs in representing vision, and have provided valuable insights into the development of more effective brain-inspired learning approaches for artificial neural networks.

B. TEXTURE AND SHAPE FEATURE BIAS

From a bio-visual perspective, shape and texture are the two most crucial factors in determining the object class, and they represent highly complementary forms of feature expression for classification tasks. However, the contribution of each feature to CNN classification results can differ significantly across different tasks [25], a distinction that is often overlooked during the traditional construction of CNNs. To investigate this issue, Geirhos et al. [26] examined the response of specific layers in ImageNet-trained CNNs to shape and texture and found that these networks exhibit a bias towards texture. Nevertheless, increasing shape bias can improve the network's accuracy and robustness, as demonstrated by Resnet-50 training using stylized ImageNet images where object classification performance improved significantly.

While a suitable training dataset can overcome the texture bias in standard CNNs and allow them to utilize more shape cues, a recent study by Shi et al. [27] modeled a novel dropout method for mitigating CNN texture bias and improving the model's robustness. It is worth noting that CNNs with added shape bias tend to be more similar to the human visual learning process. Furthermore, recent studies have shown that the degree of CNN bias towards texture depends on the training dataset and the specific learning task [28].

Similar to the mechanism of brain neural encoding of features, a pre-trained CNN encoder has a portion of its neurons strongly associated with shape features, another portion strongly associated with texture features, and a significant portion for encoding other features. However, unlike human visual perception, the distribution of these features does not change after the CNN learning process is completed, and it is difficult to intervene and adjust manually. When the difference in the number of neurons used to encode shape and those used to encode texture

features is too large, the texture bias of CNNs can become a significant vulnerability to attack, and adversaries can alter the classification results by manipulating texture [29]. This can lead to the lack of robustness of CNNs in real-world applications, including in critical domains such as medical imaging, where fatal errors can occur [30], [31].

C. FEATURE FOR RS SCENE CLASSIFICATION

Texture features and shape features are crucial for RS image classification, and can be considered as high-frequency and low-frequency features respectively. During the CNN forward process, high-frequency features are gradually blurred through layered downsampling and convolution operations. These high-frequency features are important for capturing diversity within classes and similarity between classes, and can effectively distinguish between different categories. However, the impact of bias in feature frequency is ubiquitous in classification, as scenes within the same category often exhibit analogous feature dependencies [32]. For example, recognizing mountains and forests may benefit from texture bias, while recognizing categories such as roads and villages may require more support from shape features.

To explore the solutions, researchers have conducted innovative studies. For example, Chen et al. [33] proposed a texture-enhanced scene classification method for texture-rich images that improved the classification accuracy by over 6%. Chen et al. [34] developed a visual bag-of-words scene classifier based on regional covariance that can fuse multiple relevant features and reduce the dimensionality of the features. Fei et al. [35] explored the contribution of texture features to the classification accuracy of cotton farmland, and improved the classification efficiency using a random forest feature filtering method. However, most of these studies lacked quantitative estimates of features and did not progress in interpretability.

III. PROPOSED METHOD

Researches in the field of neuroscience have demonstrated that the human visual processing system comprises multiple distinct perceptual pathways for different types of features [36]. In particular, the low-level visual feature-processing regions of the human brain can automatically form feature biases based on a priori knowledge when recognizing objects, allowing for the quick and efficient identification of relevant features. However, CNN-based models do not acquire this feature bias dynamically and adaptively. To better avoid the influence of feature bias on classification tasks, we propose a novel classification model based on feature bias analysis and data fusion. The model comprises several components, including data preprocessing, a CNN-based classifier, feature bias estimation, and data re-enhancement, as shown in Fig. 2.

During the data preprocessing stage, the dataset is first enhanced to bring out texture or shape features through

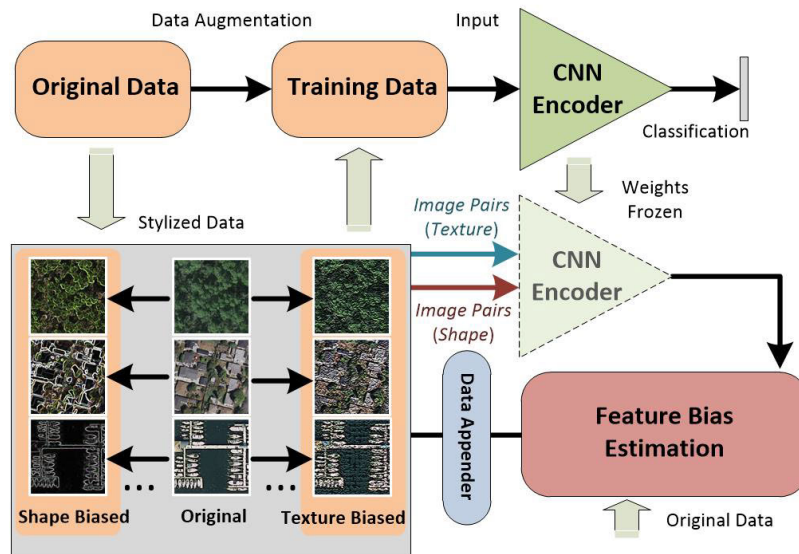


FIGURE 2. The overall frame of the proposed approach.

processing, in addition to conventional data augmentation techniques such as random flip and hue transformation. These data with feature-highlighting properties are then subjected to bias estimation and used as an alternate dataset to make adjustments to the composition of the original dataset and improve the performance of the classification model. Furthermore, we aim to investigate the impact of particular features in the augmented data on the model’s performance. To this end, we follow the approach in the literature [37] and generate two additional datasets with stylized features that exhibit some similarity to the original dataset.

The feature bias of a CNN is mainly determined by the dataset and the backbone network. The deeper the network layers, the more neurons are used to represent the shape [38]. To effectively balance texture bias and shape bias in the model, a quantitative tool is needed to accurately measure these biases. Following the approach proposed in [38], image pairs with similar texture and image pairs with similar shape were respectively fed into the pre-trained CNN with weights frozen. Then we estimate the dimensionality of semantic concepts in a pre-trained CNN encoder $E(I) = z$, where $|z|$ is the number of neurons (i.e. semantic factor). The main idea is that mutual information between similar image pairs I^a and I^b will be preserved in a particular neuron only if the neuron encodes a particular semantic concept (i.e. feature type). The mutual information $MI(z_i^a, z_i^b)$ between neuron pairs $z_i^a = E(I^a)$ and $z_i^b = E(I^b)$ can be used to quantify the extent to which a neuron encodes a specific semantic concept. The correlation coefficient ρ_i is used to estimate the mutual information $MI(z_i^a, z_i^b)$ between neuron pairs, which provides a lower bound for MI [39], [40]. Equation (1) expresses the above relation. By statistically calculating the correlation coefficients of the corresponding features in pairs of images with similar texture or similar shape, we obtain the bias score,

i.e. the number of neurons characterizing a feature.

$$MI(z_i^a, z_i^b) \geq -\frac{1}{2} \log(1 - \rho_i^2),$$

$$\text{where } \rho_i = \frac{\text{Cov}(z_i^a, z_i^b)}{\sqrt{\text{Var}(z_i^a) \text{Var}(z_i^b)}} \quad (1)$$

The primary objective of estimating the bias score is to assist in selecting more suitable data for the model, thereby facilitating a gradual convergence towards an optimal balance of texture and shape features. This convergence occurs through numerous iterative tuning steps, where the dataset configuration structure is modified to provide a better ratio of the semantic factor that are better conducive to the model’s classification performance. For instance, if a pre-trained model exhibits a texture bias, the data appender can extract shape biased images from the database and append them to the training data for retraining. Our experience indicates that such dynamic training is best performed after pre-training to reduce computational costs.

In contrast to prior work, our focus is primarily on the continuous fine-tuning of the dataset rather than fine-tuning of the network. Our proposed approach can be applied to most CNN-based coding and decoding structures. The number of neurons in softmax in these methods is not necessarily consistent, so in this paper, we set the bias ratio $\gamma = |z_{texture}|/|z_{shape}|$, and if $\gamma > 1$, a batch of shape-biased data is randomly selected from the mixed database to add to the training data. To balance the added data feature types, 10% of texture-biased data is also included in the batch. Conversely, if $\gamma < 1$, the appended data for retraining is mainly texture-biased. The number of images k in each batch can be adjusted based on (2), where η is the decay factor (initial value is less than 0.5, decaying once every m training

epochs), N is the number of images in the original training set, and “ $\lfloor \cdot \rfloor$ ” denotes rounding down.

$$k = \left\lfloor \eta N \frac{|z_{\text{texture}} - z_{\text{shape}}|}{z_{\text{texture}} + z_{\text{shape}}} \right\rfloor \quad (2)$$

As retraining progresses, data with feature bias are gradually incorporated into the training set in a stepwise manner, as the number of images needed to achieve a relatively optimal state with respect to feature bias cannot be predetermined. Ultimately, this allows us to obtain a new dataset, thus mitigating the impact of feature bias on the classifier’s performance from the data side and significantly improving its robustness for complex scenes.

IV. EXPERIMENT

A. DATASET DESCRIPTION AND DATA AUGMENTATION

In this paper, we validate and test our method using the Aerial Image dataset (AID) [41] and the NWPU-RESISC45 dataset [42], both of which are specifically designed for RS image scene classification with a moderate amount of data, making them well-suited for our focus on feature bias. AID is a large-scale dataset collected by Google Earth, consisting of 10,000 images with a resolution between 0.5 and 8 m, each measuring 600×600 pixels. The NWPU-RESISC45 dataset includes 31,500 images covering 45 land scene categories, with each class containing 700 images ranging in size from 0.2m to 30m and with a resolution of 256×256 pixels. Unlike the data used for the detection task, each class of data in AID and NWPU-RESISC45 can be viewed as a separate set, all resulting in feature bias. To design the experiments in a concise and clear manner, we focus on the classification performance of three categories: forest, harbor and residential area in the experimental results. In the data preparation stage, we use three ways to augment: symmetric flip, hue transformation, and mixup.

B. EXPERIMENTAL SETUP

The experimental section of this study does not aim to present the latest or most advanced CNN encoder scheme. Instead, the focus is primarily on verifying the data appending and robust enhancement strategy. Therefore, the D-CNN framework [43], which handles intra-class diversity and inter-class similarity problems effectively, is utilized as the CNN-based classifier in the full-class experiments. In the single-class experiments, VGG-16, ResNet50, D-CNNs, and SCCov [44] are selected as the CNN-based classifier, with the latter two backbone networks uniformly using VGG-16. In the AID dataset, the data is divided into 20% and 50% for training and 80% and 50% for testing, respectively. In the NWPU-RESISC45 dataset, the training set is set to 10% and 20%, while the remaining 90% and 80% are used for testing (the training rate is denoted as Tr). The experiments are conducted using Pytorch with an Intel(R) Core(TM) i7-10700K CPU, 64GB RAM, and Nvidia GeForce RTX 3080 GPU. For the training parameters, the

batch size is set to 64, the learning rate is 0.001, and the learning rate decays to half of the original value every 10 epochs. The decay factor η is set to 0.1, and the decay period of retraining m is set to 10.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) SINGLE CLASS CLASSIFICATION EXPERIMENT

When distinguishing between different RS scenes, our reliance on texture versus shape information varies. To investigate this, we conducted experiments using three single-class scenes, specifically forests, ports, and residential areas, which were trained separately for classification. Since the data feed for each training are the same class data, the similarity between classes in the test data may lead to a loss of accuracy in this part of the experiments. Therefore, we focused on the change in model performance during retraining and did not consider the accuracy loss in the individual class experiments.

Before conducting the experiments, we trained the classifier using the original data from a single class, and after training until convergence, we froze all parameters and examined the proportion of texture features and shape features among the high-level features extracted by CNN. The estimated results of the potential representational semantic factor $|z_i|$ in the fifth stage of the CNN for the AID and NWPU-RESISC45 datasets are shown in Tables 1 and 2, respectively, where Tr is 50% and 20%. The total dimensionality of most potential representation $|z|$ is 2048 (2048×3 for SCCov). In addition to the estimated representations of texture and shape, the remaining dimensions may serve as potential representations of other features. The tables demonstrate that the CNN encoder in all four methods exhibits some feature bias, with texture bias being prevalent. However, the specific estimates differ due to the fine-tuning variations of the encoder and structural differences in the algorithm. As shown in the tables, the CNN encoder in all four methods has some feature bias, and almost all of them exhibit texture bias. However, the specific estimated values differ, mainly due to the fine-tuning variation of the encoder and the algorithm structure variation.

The model was retrained based on feature bias estimation, and its performance during retraining was recorded, as depicted in Fig. 2. We monitored and updated the accuracy information of the model every m epochs throughout the retraining process. The figure reveals that the model does not produce accuracy improvement when the training data is updated in a few cases, which can primarily be attributed to the randomness of the added data. As we continued to update the training data, the additional data’s volume gradually decreased, and the classification accuracy eventually stabilized at a certain level. Notably, under equal conditions, the final stable values of the smaller Tr curves exceeded the initial values of the larger Tr curves, thereby demonstrating that the proposed method in this paper offers a better classification performance gain than simply adding more data.

TABLE 1. An estimation of the proportion of latent semantic representation factors for forest, harbor, and residential area in the trained CNN encoder on AID dataset. ($Tr = 50\%$).

Model (stage 5)	Factor $ z_{feature} / z $ (%)					
	Shape			Texture		
	Forest	Harbor	Residential area	Forest	Harbor	Residential area
VGG-16	17.55	17.97	17.92	34.62	33.61	34.11
ResNet50	19.35	19.47	18.61	32.64	32.12	33.27
D-CNNs	17.90	18.86	18.37	38.07	36.74	37.92
SCCov	17.01	18.96	19.83	41.45	42.67	43.03

TABLE 2. An estimation of the proportion of latent semantic representation factors for forest, harbor, and residential area in the trained CNN encoder on NWPU-RESISC45 dataset. ($Tr = 20\%$).

Model (stage 5)	Factor $ z_{feature} / z $ (%)					
	Shape			Texture		
	Forest	Harbor	Residential area	Forest	Harbor	Residential area
VGG-16	18.29	18.91	19.00	34.96	33.98	35.31
ResNet50	19.71	19.49	20.89	34.63	34.24	35.33
D-CNNs	19.32	19.77	20.05	38.65	36.90	37.36
SCCov	19.99	20.08	21.22	42.47	41.91	42.49

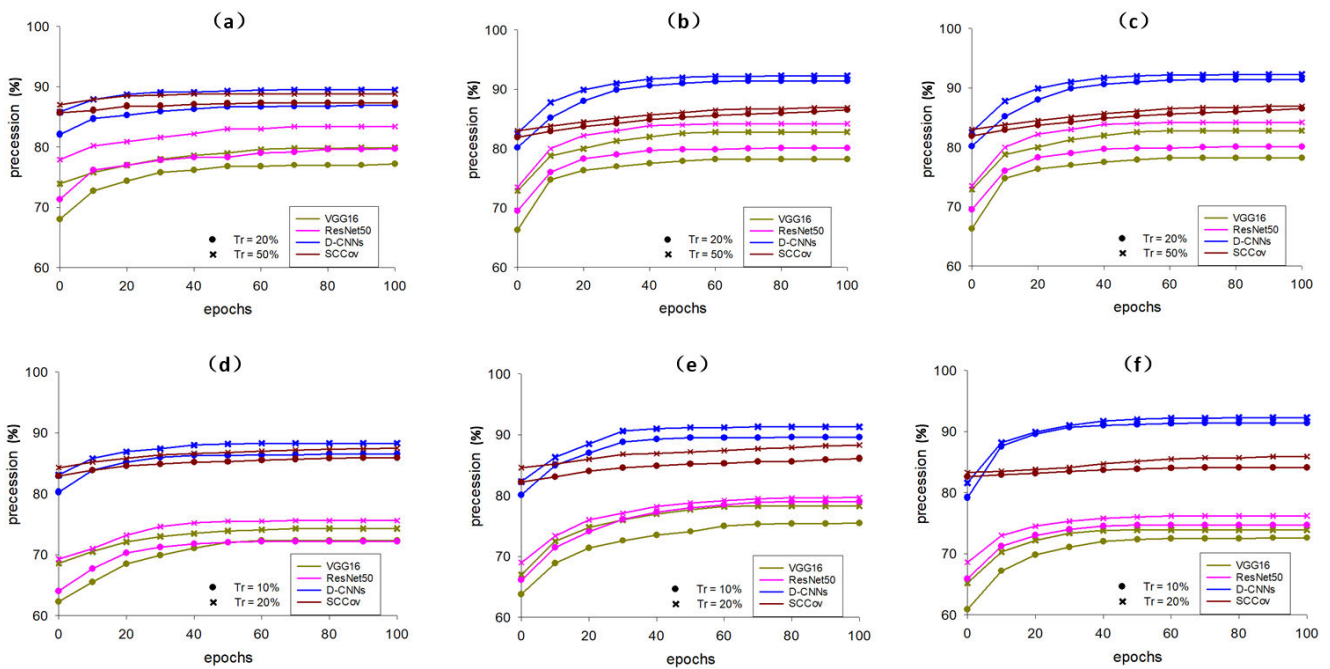


FIGURE 3. For every m epochs ($m = 10$), the performance changes of various baseline models during the retraining process were recorded (forest, harbor, and residential class only). Specifically, data for (a), (b), and (c) were from AID dataset, while data for (d), (e), and (f) were from NWPU-RESISC45 dataset.

Although the change in retraining accuracy is relatively flat compared to the early training phase, the models' classification performance improved to some extent after several data increments. Moreover, SCCov's CNN encoder structure's cross-layer design helped to preserve the local texture features in the early convolutional layers, but the total

feature dimension increased after feature stitching, leading to a flatter change in accuracy during retraining.

2) FULL-CLASS CLASSIFICATION EXPERIMENT

In the full-class classification experiments, it is difficult to accurately estimate the proportion of semantic representation

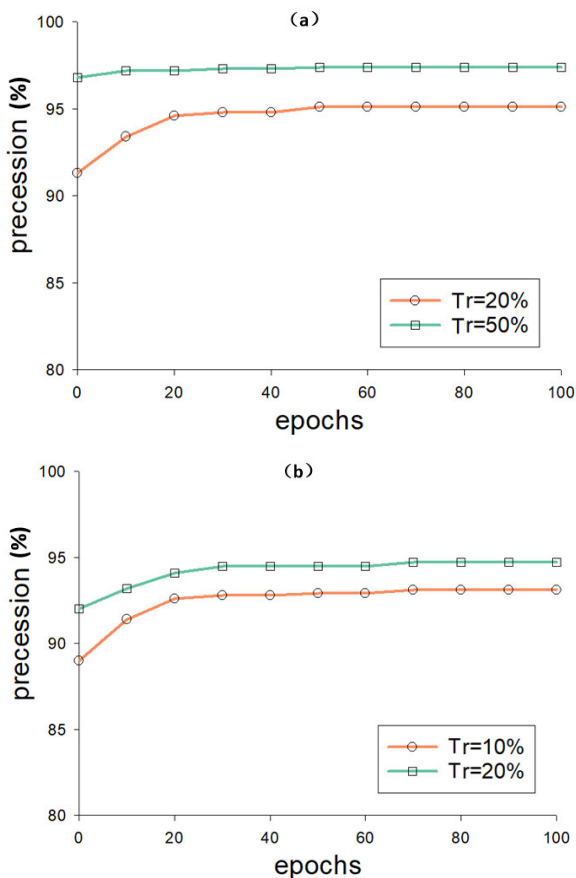


FIGURE 4. The accuracy changes of the proposed model during retraining on AID dataset (a) and NWPU-RESISC45 dataset (b).

factors for both shape and texture on each class when the model proposed in this paper is applied. Thus, we adopted a similar strategy as in the previous section, where we combine all the generated stylized data in an unordered manner and only estimate the feature bias of all the data as a whole. Fig. 3 displays the performance variation of the proposed model retrained on the AID and NWPU-RESISC45 datasets. It can be observed from the figure that Tr is positively correlated with model accuracy, and the curve change is flatter for larger Tr . This is primarily because more training data can offer richer features, which somewhat mitigates the performance loss caused by the feature bias problem of the classifier. The larger amount of data and smaller resolution in NWPU-RESISC45 provide relatively weaker quality of features. Therefore, the retraining on AID with the same Tr value of 20% is more effective than on NWPU-RESISC45, yielding a classifier with a slightly better gain effect.

V. CONCLUSION

Designing and modifying deep learning models based on brain-inspired principles is a key driver of their gradual evolution. In this work, we propose a data enhancement and retraining approach that incorporates feature bias estimation, drawing inspiration from how the human brain perceives

different features during classification. Our approach has demonstrated a notable positive influence on the performance of CNN-based remote sensing scene classifiers. Nevertheless, its constraints are apparent, particularly in scenarios where the data volume is substantial or feature maps are spliced, it produces quite limited improvement. Although the overall enhancement is not a major breakthrough, it provides a new way of thinking and direction for the problem of RS image scene classification. Additionally, the optimized design principles inspired by our framework can be replicated and applied to similar tasks in other industries, such as industrial device classification and driving scene classification. In the following phase, we may deliberate on utilizing the backdrop of two opposing cases with extremely limited and abundantly ample samples to further investigate and refine the method put forward in this paper.

REFERENCES

- [1] A. Tayyebi, B. C. Pijanowski, and A. H. Tayyebi, "An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran, Iran," *Landscape Urban Planning*, vol. 100, nos. 1–2, pp. 35–44, Mar. 2011.
- [2] T. Zhang and X. Huang, "Monitoring of urban impervious surfaces using time series of high-resolution remote sensing images in rapidly urbanized areas: A case study of Shenzhen," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2692–2708, Aug. 2018.
- [3] 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.
- [4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jan. 2004.
- [5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Soc. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.
- [6] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
- [7] 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. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.
- [8] 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.
- [9] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.
- [10] 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, Sep. 2019.
- [11] 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. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1150–1162, 2022.
- [12] W. Wang, Y. Chen, and P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533918.
- [13] J. Shen, T. Yu, H. Yang, R. Wang, and Q. Wang, "An attention cascade global-local network for remote sensing scene classification," *Remote Sens.*, vol. 14, no. 9, p. 2042, Apr. 2022.
- [14] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Comput. Biol.*, vol. 10, no. 12, Dec. 2014, Art. no. e1003963.
- [15] Y. Xu and M. Vaziri-Pashkam, "Limits to visual representational correspondence between convolutional neural networks and the human brain," *Nature Commun.*, vol. 12, no. 1, pp. 1–16, Apr. 2021.

- [16] Y. Mohsenzadeh, C. Mullin, B. Lahner, and A. Oliva, "Emergence of visual center-periphery spatial organization in deep convolutional neural networks," *Sci. Rep.*, vol. 10, no. 1, pp. 1–8, Mar. 2020.
- [17] A. A. Zeman, J. B. Ritchie, S. Bracci, and H. Op de Beeck, "Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Feb. 2020.
- [18] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–13.
- [19] M. Poo, "Towards brain-inspired artificial intelligence," *Nat. Sci. Rev.*, vol. 5, no. 6, p. 785, Oct. 2018.
- [20] K. K. Parhi and N. K. Unnikrishnan, "Brain-inspired computing: Models and architectures," *IEEE Open J. Circuits Syst.*, vol. 1, pp. 185–204, 2020.
- [21] Z.-X. Zhang, B. Luo, J. Tang, S. Yu, and A. Hussain, "Editorial for special issue on brain-inspired machine learning," *Mach. Intell. Res.*, vol. 19, no. 5, pp. 347–349, Sep. 2022.
- [22] G. Jacob, R. T. Pramod, H. Katti, and S. P. Arun, "Qualitative similarities and differences in visual object representations between brains and deep networks," *Nature Commun.*, vol. 12, no. 1, pp. 1–14, Mar. 2021.
- [23] I. Kuzovkin, R. Vicente, M. Petton, J.-P. Lachaux, M. Baciú, P. Kahane, S. Rheims, J. R. Vidal, and J. Aru, "Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex," *Commun. Biol.*, vol. 1, no. 1, pp. 1–12, Aug. 2018.
- [24] R. M. Battleday, J. C. Peterson, and T. L. Griffiths, "Capturing human categorization of natural images by combining deep networks and cognitive models," *Nature Commun.*, vol. 11, no. 1, pp. 1–14, Oct. 2020.
- [25] Y. Ge, Y. Xiao, Z. Xu, X. Wang, and L. Itti, "Contributions of shape, texture, and color in visual recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 369–386.
- [26] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*.
- [27] B. Shi, D. Zhang, Q. Dai, Z. Zhu, and Y. Mu, "Informative dropout for robust representation learning: A shape-bias perspective," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 8828–8839.
- [28] Y. Li, Q. Yu, M. Tan, J. Mei, P. Tang, W. Shen, A. Yuille, and C. Xie, "Shape-texture debiased neural network training," 2020, *arXiv:2010.05981*.
- [29] K. Hermann, T. Chen, and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19000–19015.
- [30] J. Li, G. Zhu, C. Hua, M. Feng, P. Li, X. Lu, J. Song, P. Shen, X. Xu, L. Mei, L. Zhang, S. Afaq Ali Shah, and M. Bennamoun, "A systematic collection of medical image datasets for deep learning," 2021, *arXiv:2106.12864*.
- [31] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jul. 2017.
- [32] W. Zhang, L. Jiao, F. Liu, J. Liu, and Z. Cui, "LHNet: Laplacian convolutional block for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626513.
- [33] X. Chen and J. Zhu, "Land scene classification for remote sensing images with an improved capsule network," *J. Appl. Remote Sens.*, vol. 16, no. 2, May 2022, Art. no. 026510.
- [34] X. Chen, G. Zhu, and M. Liu, "Bag-of-visual-words scene classifier for remote sensing image based on region covariance," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [35] H. Fei, Z. Fan, C. Wang, N. Zhang, T. Wang, R. Chen, and T. Bai, "Cotton classification method at the county scale based on multi-features and random forest feature selection algorithm and classifier," *Remote Sens.*, vol. 14, no. 4, p. 829, Feb. 2022.
- [36] E. A. DeYoe, G. J. Carman, and P. Bandettini, "Mapping striate and extrastriate visual areas in human cerebral cortex," *Proc. Nat. Acad. Sci. USA*, vol. 93, no. 6, pp. 2382–2386, 1996.
- [37] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2414–2423.
- [38] M. Amirul Islam, M. Kowal, P. Esser, S. Jia, B. Ommer, K. G. Derpanis, and N. Bruce, "Shape or texture: Understanding discriminative features in CNNs," 2021, *arXiv:2101.11604*.
- [39] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.
- [40] D. V. Foster and P. Grassberger, "Lower bounds on mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 83, no. 1, Jan. 2011, Art. no. 010101.
- [41] 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, Apr. 2017.
- [42] 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.
- [43] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
- [44] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2020.



ZHONG DONG received the B.E. degree from the Harbin Institute of Technology, Harbin, China, in 2010, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2020. Currently, he is a Postdoctoral Researcher with the Department of Automation, Tsinghua University. His research interests include remote sensing data processing, machine learning, and computer vision.



BAOJUN LIN received the Ph.D. degree from Jilin University, in 1991.

He was a Doctoral Supervisor and a Chief Designer of Beidou's Third-Generation Satellite. He is the Vice Chief Director of the Innovation Academy for Microsatellites, Chinese Academy of Sciences. He has presided over the completion of the development, testing, and launch of two Beidou second-generation experimental satellites and ten Beidou-3 satellites. His main research

interests include satellite navigation technology, aerospace technology, spacecraft system design, and space application system technology.



FANG XIE received the M.S. degree from the Harbin Institute of Technology, in 2017. He is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences.

He is an Assistant Researcher with the Innovation Academy for Microsatellites, Chinese Academy of Sciences. He was a Satellite System Engineer of Beidou's Third-Generation Satellite. His research interests include spacecraft system design, satellite navigation technology, and computer vision.

...