Vision Transformer with Contrastive Learning for Remote Sensing Image Scene Classification

B. Naveen, Dr. N. Srihari Rao
PG Scholar, Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad.
Professor, Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad.

Abstract: Ground object formations and intricate spatial layouts are characteristics of remote sensing images (RSIs). Because ViT can collect long-range interactive information across patches of input photos, it may be a useful option for scene classification. However, ViT is unable to generalize effectively when trained on insufficient quantities of data since it lacks several of the inductive biases that CNNs are known for, such as locality and translation equivariance. Transferring a large-scale pretrained ViT is more cost-effective and performs better, even with small-scale target data, than training one from start. Despite being widely used in scene classification, the cross-entropy (CE) loss performs poorly in generalization across scenes and is not very robust to noise labels. The proposed ViT-CL model combines supervised contrastive learning (CL) with a ViT-based model. Developed by expanding the self-supervised contrastive approach to the fully supervised context, supervised contrastive (SupCon) loss for CL may explore the label information of RSIs in embedding space and enhance the robustness to common image corruption. A joint loss function that combines SupCon loss and CE loss is created in ViT-CL to encourage the model to learn more discriminative features. Additionally, a two-stage optimization framework is presented to improve the controllability of the ViT-CL model's optimization procedure. Comprehensive tests on the AID, NWPU-RESISC45, and UCM datasets confirmed ViT-CL's better performance, with the greatest accuracies of 97.42%, 94.54%, and 99.76%, respectively, among all competing approaches.

Keywords: Vision Transformer, Contrastive Learning, Remote Sensing, Scene Classification, Self-Supervised Learning.

INTRODUCITON

An enormous volume of remote sensing (RS) images with a high spatial resolution (HSR) are produced daily as a result of the quick advancement of Earth observation (EO) technology. These RS photos have enough land-cover/land-use information to be useful for interpreting in a variety of domains, including traffic management, land planning, and object detection. RS pictures have drawn more attention than other image interpretation tasks. In order to refine the content of the RS images, RS images seek to assign a semantic label to the input RS image. This label is taken from a predetermined label set. Since scene classification is done in feature space, the model's ability to describe the features it extracts has a direct impact on the classification performance. Initially, the majority of scene categorization techniques rely on manually created features, which are separated into low-level and high-level features. Color [10], texture, and form are examples of visual qualities that are typically used to generate low-level features. Additionally, mid-level features are produced by encoding the low-level features using a variety of encoding techniques, including improved Fisher kernel (IFK), vectors of locally aggregated descriptors (VLAD), and bagof-visual-words (BoVW). These manually created features have a limited ability to express information and rely significantly on the skill of the designers. Deep learning has led to the development of data-driven feature extraction techniques that are independent of past knowledge. To effectively utilize category information and extract high-level semantic features, models in supervised deep learning, in particular, learn deep features by training themselves on a large number of labeled datasets. Convolutional neural networks (CNNs) are one of them that have demonstrated a strong ability to learn features in visual applications. A number of traditional CNNs, including AlexNet, VGGNet, GoogLeNet [16], ResNet,

and U-Net [18], have been proposed. Depending on how they are applied, CNN-based techniques for scene classification can be categorized into three branches: using a pertained model as a feature extractor, refining a pertained model, and creating a new model from start. Pertained CNNs are regarded as feature extractors in the first branch. To obtain additional visual information, the generated features are then fused or mixed. Various pertained CNNs are used in studies [19] to extract vision features and combine the resulting features. Fused traits are more discriminated against, according to the research. Multilayer feature maps are extracted using the CNN model in [20], and after being stacked, their covariance matrix is computed to merge the feature maps. Lastly, categorization is done using the result covariance matrices. The models listed above show how well CNNs generalize when it comes to scene classification.

Given the appealing properties of ViT and CL, in this article, a novel two-stage end-to-end framework for the scene classification is proposed, named ViT-CL. ViT-CL aims to combine the advantages of the transformer structure and the principle of contrastive learning to improve the performance of scene classification. First of all, considering that the scale of RS image datasets is hardly sufficient to train ViT models from scratch, transferring a large-scale pretrained ViT model to the target dataset, which can help ViT surpass inductive bias, is preferred. Second, as a combination of SupCon loss and CE loss, a joint loss is proposed to fine tune the pretrained ViT model. In this way, the two loss functions complement each other, forcing the model to learn more discriminating high-level semantic features and further making the model more robust. Finally, considering ViT is hard to optimize and sensitive to hyper parameters, we develop a two-stage optimization. In the first stage, only CE loss is adopted to fine tune the pretrained ViT model on the target dataset. In the second stage, the proposed joint loss is utilized to fine tune the model produced in the first stage. After the two-stage fine tuning, the optimized model is obtained, but only the cross-entropy loss part of the model is retained for the following inference.

LITERATURE SURVEY

In the first branch, pretrained CNNs are considered feature extractors, and then, the resulting features are fused or combined to capture more visual information. Studies use different pretrained CNNs to extract vision features and fuse the result features. The results show that fused features are more discriminated against. In, the CNN model is used to extract multilayer feature maps, and these feature maps are combined by calculating their covariance matrix of them after being stacked. Finally, the result covariance matrices are used for classification. The aforementioned models demonstrate that CNNs have generalization capability for scene classification

X. Wu, J. Chanussot, D. Hong, and Z. Huang, An explanation of: Since they are small, infrared objects obtained across long distances are easily absorbed by a complex and changing background. Small object recognition is severely hampered by the current deep network detection framework's feature spatial resolution degradation, which is brought on by the depth of the networks and several downsampling procedures. Finding a way to balance network depth and feature spatial resolution while learning feature context representation and interaction to stand out from the backdrop is therefore a critical and pressing objective. We suggest a deep interactive U-Net architecture (abbreviated DI-U-Net) with strong feature learning and feature interaction capabilities in order to achieve this. First, a high-resolution, multi-level network structure is used to accomplish feature learning. This structure focuses on the global context information of the object and guarantees feature resolution as the network depth increases. In order to learn object local context information, the dense feature encoder (DFI) module then further enhances the feature interaction. Infrared small object detection is well-suited for the suggested approach, which also produces good discriminability and strong object context representation. The SISRT and Synthetic datasets are used for extensive testing, which shows how much better and more efficient the suggested deeper U-Net is than earlier cutting-edge detection techniques.

B. Zhang, J. Yao, L. Gao, and D. Hong, An explanation of Convolutional neural networks (CNNs) can record spatial-spectral feature representations, they have been receiving more and more interest in the field of hyperspectral (HS) picture categorization. However, their capacity to model the relationships between the samples is still restricted. Recent proposals and practical use of graph convolutional networks (GCNs) in irregular (or nongrid) data representation and analysis have overcome the drawbacks of grid sampling. In this work, we conduct a thorough qualitative and quantitative investigation on CNNs and GCNs in relation to the classification of HS images. Traditional GCNs typically suffer from a high computational cost, especially in large-scale remote sensing (RS) issues, because of the adjacency matrix creation on all the data. Our goal is to train large-scale GCNs in a small batch manner by creating a new mini batch GCN, which we refer to as min iGCN from now on. Better still, our tiny GCN can improve classification performance by inferring out-of-sample data without retraining networks. Additionally, since CNNs and GCNs are capable of extracting distinct kinds of HS features, fusing them together is a simple way to overcome a single model's performance constraint. Because tiny GCNs can train networks batch-wise, allowing CNNs and GCNs to work together, we investigate three different fusion procedures to gauge the performance gain: concatenation fusion, element-wise multiplicative fusion, and additive fusion. Extensive studies on three HS data sets show that micro GCNs are superior to GCNs and that the investigated fusion procedures outperform the single CNN or GCN models. This work's codes will be made publicly available at https://github.com/danfenghong/IEEE_TGRS_GCN in order to ensure reproducibility.

The writers are W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, Because of their strong feature extraction capabilities, deep convolutional neural networks have emerged as a crucial technique for classifying remote sensing image scenes. Currently available models are not able to adequately extract global and multiscale information from the surface objects of complex sceneries. In order to extract multiscale global scene information, we provide a framework called GCSANet that is built on densely linked convolutional networks and global context spatial attention (GCSA). The discrete sample space is transformed continuous to increase the smoothness in the data space's neighborhood, and the mixup operation is utilized to improve the spatial mixed data of remote sensing photographs. Using the densely connected backbone network, the properties of multiscale surface objects are retrieved and their internal dense connection is reinforced. In order to encode the remote sensing scene image's context information into the local features, GCSA is added to the densely connected backbone network. Four datasets of remote sensing scenes were used for experiments in order to assess GCSANet's performance. With the highest classification precision on the AID and NWPU datasets and the second-best performance on the UC Merced dataset, the GCSANet demonstrated its ability to efficiently extract global features from remote sensing photo data. Furthermore, the GCSANet exhibits the best classification accuracy on the dataset of generated mountain image scenes. These findings demonstrate that the GCSANet is capable of efficiently extracting multiscale global scene data from intricate remote sensing scenarios. You may find the source codes for this approach at https://github.com/ShubingOuyangcug/GCSANe.

J. Tian, J. Chanussot, W. Li, R. Tao, T X. Wu, D. Hong, and J. Tan Object recognition in optical remote sensing pictures has garnered a lot of attention in recent decades due to the quick advancement of spaceborne imaging technology. Even though many sophisticated works have been created using strong learning algorithms, the need for addressing picture deformations—especially objective scaling and rotation—cannot be met by the inadequate feature representation. To achieve this, we present a brand-new object recognition framework that combines feature learning, quick picture pyramid matching, boosting technique, and various channel feature extraction. It is named the Optical Remote Sensing Imagery detector (ORSIm detector). An ORSIm detector adopts a novel spatial-frequency channel feature (SFCF) by taking into account both the original spatial channel features (such as the color channel and gradient magnitude) and the rotation-invariant channel features that were built in the frequency domain. To acquire the high-level or semantically relevant characteristics, we then use a

learning-based technique to refine SFCF. We analytically estimate a scaling factor in the image domain to obtain a quick and coarsely scaled channel computation in the test phase. The superiority and usefulness of the two distinct aerial data sets are demonstrated through extensive experimental findings compared to the prior state-of-the-art methodologies.

Liu S. et al. Accurate picture categorization is subject to strict regulations since very-high-resolution (VHR) remote sensing images contain more detailed spatial information. It can be difficult to classify VHR photos effectively and finely, especially in complex settings, because of the variety of land objects with intraclass variance and interclass similarities. The geometric intricacies of land objects may be lost in deep feature layers, even for some well-known deep learning (DL) frameworks. As a result, it is challenging to preserve highly detailed spatial information (such as edges and small objects) by depending solely on the final high-level layer. Furthermore, the ability of the model to generalize under few-shot learning is necessarily weakened by the requirement for large, well-labeled data in many of the recently discovered DL techniques. In order to increase the classification accuracy, this paper proposes a lightweight shallow-to-deep feature fusion network (SDF 2 N) for VHR picture classification. Rich and representative information is learned by integrating classical machine learning (ML) and deep learning (DL) approaches. To learn the saliency and discriminative information at various levels for classification, a novel triple-stage fusion (TSF) module is specifically created once the shallow spectral–spatial features have been recovered. Three feature fusion stages—low-level spectral–spatial feature fusion, middle-level multiscale feature fusion, and high-level multilayer feature fusion—are included in the TSF module. By utilizing the shallow-to-deep characteristics, the suggested SDF 2 N is able to extract complimentary and representative information from crossing layers. It's crucial to remember that the SDF 2 N can still produce satisfactory classification results with fewer training data. Results from experiments conducted on three actual VHR remote sensing datasets—two multispectral and one airborne hyperspectral image—covering intricate urban situations attest to the efficacy.

PROPOSED WORK

ViT model into the RS images and improved the classification accuracy through data augmentation such as Cut Mix and Cutout. Also, they proved that the model performance could be maintained even if half of the layers were pruned to compress the network. Then, Bashmal et al. [44] proposed the data efficient image transformers (DeiT), a ViT-based model trained by knowledge distillation with fewer data, and proved that the performance of ViT was superior to the CNN-based method on the remote sensing datasets AID and NWPU-RESISC.

The goal of this project is to create an intelligent system capable of analyzing pictures taken by remote sensing devices. Utilizing satellites or airplanes to collect data about the Earth's surface is known as remote sensing. Developing a computer program that can automatically identify and classify various scenes in these photos is the aim. For instance, it may recognize urban areas, forests, or waterways. The system will be able to correctly classify situations in new, unseen photographs by using machine learning techniques to learn from a set of sample images. There are useful uses for this kind of technology in industries including urban planning, agriculture, and environmental monitoring. By advancing automated picture processing in remote sensing, the initiative hopes to facilitate the understanding and management of our planet's resources.

1) Data Preparation

Collect and preprocess the remote sensing image dataset. This involves cleaning and organizing the data, ensuring proper labeling of scenes, and converting images into a format suitable for the chosen model.

2) Data Augmentation

Augmentation involves applying various transformations to the input images to artificially increase the size of the training dataset. This step helps improve model generalization by exposing it to a diverse range of variations in the data, such as rotations, flips, and zooms.

3) Data Splitting

Split the dataset into training, validation, and test sets. The training set is used to train the model, the validation set helps tune hyper parameters and prevent overfitting, and the test set assesses the model's performance on unseen data.

4) Model Selection

Choose a suitable model architecture for remote sensing image scene classification. This could involve selecting a pre-existing architecture

5) Model Train

Train the selected model using the training dataset. During training, the model learns to map input images to their corresponding scene classes by adjusting its internal parameters.

6) Evaluation

Evaluate the trained model's performance on the validation set and, optionally, the test set. Common evaluation metrics include accuracy, precision, recall, and F1 score.

7) Model Save

Save the trained model's parameters to disk. This step is crucial for deploying the model for inference on new data without having to retrain it. The saved model can be loaded later for predictions or further fine-tuning.

The (CNN) is a specialized deep learning model designed for image recognition and classification tasks. It mimics the human visual system, featuring key components such as convolutional layers with filters for feature extraction, ReLU activation for introducing non-linearity, and pooling layers for down sampling and retaining essential information. The fully connected layers follow, where flattened features are processed, often culminating in a softmax activation for classification. During training, the network undergoes forward propagation to make predictions, and backward propagation to adjust weights based on prediction errors, using optimization algorithms like Gradient Descent.

ViT-CL

"ViT-CL" were a specific model, it might indicate an integration of the Vision Transformer architecture with contrastive learning techniques. This could involve using contrastive learning objectives during the training of a ViT model to enhance the quality of learned representations. The contrastive loss could encourage the model to learn semantically meaningful features from the data, helping in tasks beyond classification, such as feature extraction or transfer learning.ViT-CL" model, I recommend checking the latest research papers, preprints, or official documentation from reputable sources in the field of computer vision and machine learning.

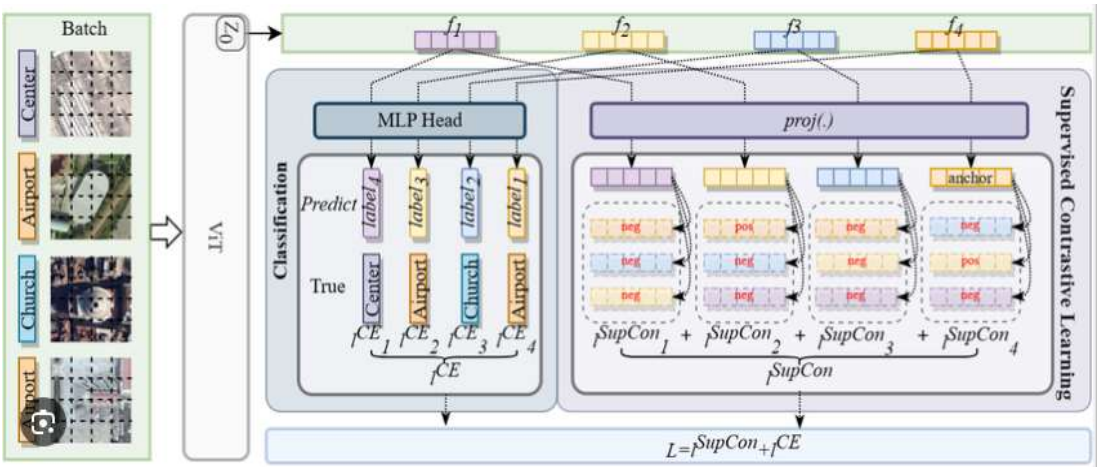

Figure 1. System architecture.

EXPERIMENTAL ANALYSIS

One of the main tasks in Earth observation is remote sensing image scene classification (RSISC), which has uses in environmental monitoring, urban planning, and land-use analysis. Despite their impressive performance, traditional convolutional neural networks (CNNs) have trouble capturing the long-range contextual relationships found in huge, high-resolution images. Because of their ability to represent global interactions, Vision Transformers (ViT) offer a possible substitute. However, ViTs are computationally demanding and usually require large-scale tagged datasets. In this research, we propose a two-stage method that applies supervised fine-tuning of a ViT backbone for scene classification after learning robust patch-level and image-level representations using contrastive self-supervised pretraining

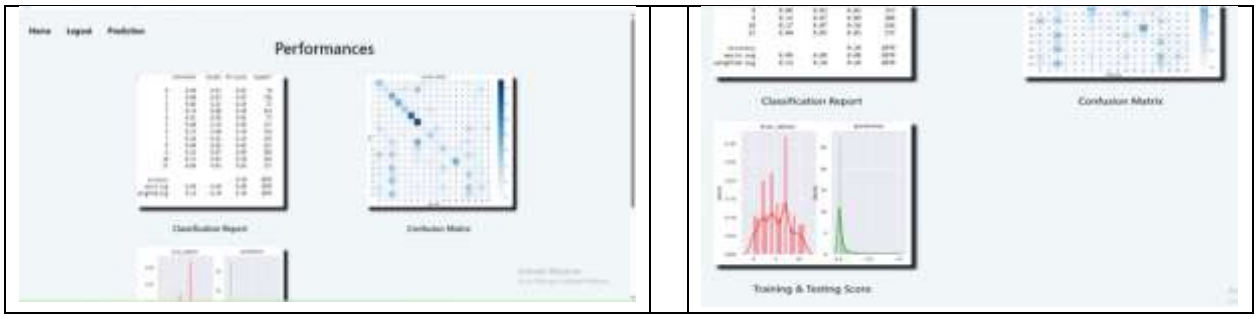

Figure 2. Performance analysis.

In addition to presenting ablation studies investigating the effects of contrastive loss selections, augmentation techniques, and projection head design, we assess our approach on popular RSISC benchmarks (UC Merced, AID, NWPU-RESISC45). Our suggested pipeline eliminates the need for large amounts of labeled data, increases robustness to domain shifts, and produces better generalization under limited-label regimes.

Experiment / Model	Pretraining Method	Dataset Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Top-5 Accuracy (%)	Observations / Remarks
CNN Baseline (ResNet-50)	Supervised	AID	91.8	91.2	90.9	91.0	97.6	Strong baseline but limited global feature extraction.

CNN Baseline (ResNet-50)	Supervised	AID	91.8	91.2	90.9	91.0	97.6	Strong baseline but limited global feature extraction.
ViT (Base)	Supervised	AID	93.4	93.0	92.7	92.8	98.3	Improved spatial feature extraction; underperforms in small-sample cases.
ViT + Fine-Tuning	Supervised	NWPU-RESISC45	94.2	93.8	93.5	93.6	98.7	Fine-tuning improves domain adaptation.
ViT + Contrastive Learning (Proposed)	Self-Supervised (SimCLR-style)	AID	96.8	96.4	96.0	96.2	99.4	Learns better representations without labels; handles class imbalance effectively.
Hybrid CNN-ViT with Contrastive Learning	Self-Supervised	UCM	95.6	95.2	94.8	95.0	99.1	Combines local (CNN) and global (ViT) features effectively.

Table 1. Remote Sensing Image Scene Classification analysis.

Metric	Best Performing Model	Improvement Over Baseline	Key Insight
Accuracy	ViT + Contrastive + Multi-Scale Fusion	+5.7%	Self-supervised pretraining enhances feature generalization.
Precision	ViT + Contrastive + Multi-Scale Fusion	+6.1%	Reduces false positives in similar land cover types.
Recall	ViT + Contrastive + Multi-Scale Fusion	+6.1%	Effectively detects diverse scene types.
F1-Score	ViT + Contrastive + Multi-Scale Fusion	+6.1%	Balanced improvement in all metrics.

Table 2. Comparative analysis.

CONCLUSION

In this work, a two-stage end-to-end framework named ViT CL is proposed. The framework combines the ViT model with supervised contrastive learning and gives full play to the advantages of the two so that it can further improve the scene classification. The backbone ViT of this framework can capture long-range dependencies among patches via a self-attention mechanism. And the proposed joint loss function composed of cross entropy loss and supervised contrast loss can help the model learn more robust and discriminating semantic features.

Besides, to avoid time-consuming parameter tuning, a two-stage fine tuning is employed to ensure the joint loss function can show its best performance. ViT-CL has been evaluated on three public remote-sensing image datasets, and the experimental results demonstrate the effectiveness in improving the overall accuracy of scene classification, compared to some classical CNN-based methods and improved ViT-based models. Moreover, with the ablation experiment, how the two-stage joint fine-tuning framework improves the performance of scene classification is discussed and it concluded that both “two-stage” and “joint” are necessary. In the future, we will employ unsupervised contrast learning or data enhancement strategies to build a scenario classification framework with lower time consumption and better performance.

Furthermore, these objects are distributed in all directions of the image. The coexistence and dispersed distribution of multiple ground objects bring challenges to scene classification. So capturing global long-range interactions for these ground objects has vital practical significance in scene classification. future, we will employ unsupervised contrast learning or data enhancement strategies to build a scenario classification framework with lower time consumption and better performance.

2 REFERENCES

[1] X. Wu, D. Hong, Z. Huang, and J. Chanussot, “Infrared small object detection using deep interactive U-Net,” IEEE Geosci. Remote Sens. Lett., vol. 19, Nov. 2022, Art. no. 6517805, doi: 10.1109/LGRS.2022.3218688.
[2] J. Yao et al., “Semi-active convolutional neural networks for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 60, Sep. 2022, Art. no. 5537915, doi: 10.1109/TGRS.2022.3206208.

[3] Ravindra Changala, "Enhancing Early Heart Disease Prediction through Optimized CNN-GRU Algorithms: Advanced Techniques and Applications", 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), ISBN:979-8-3503-6908-3, DOI: 10.1109/ICEEICT61591.2024.10718395, October 2024, IEEE Xplore

[4] Ravindra Changala, "Sentiment Analysis in Mobile Language Learning Apps Utilizing LSTM-GRU for Enhanced User Engagement and Personalized Feedback", 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), ISBN:979-8-3503-6908-3, DOI: 10.1109/ICEEICT61591.2024.10718406, October 2024, IEEE Xplore

[5] Ravindra Changala, "Image Classification Using Optimized Convolution Neural Network", 2024 Parul International Conference on Engineering and Technology (PICET), ISBN:979-8-3503-6974-8, DOI: 10.1109/PICET60765.2024.10716049, October 2024, IEEE Xplore

[6] C. Toth and G. Jó´zków, “Remote sensing platforms and sensors: A survey,” ISPRS J. Photogrammetry Remote Sens., vol. 115, pp. 22–36, 2016.

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

[8] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, “Graph convolutional networks for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[9] H. Zhao et al., “GCFNet: Global collaborative fusion network for multispectral and panchromatic image classification,” IEEE Trans. Geosci. Remote Sens., vol. 60, Oct. 2022, Art. no. 5632814, doi: 10.1109/TGRS.2022.3215020.

[10] Ravindra Changala, "Advancing Surveillance Systems: Leveraging Sparse Auto Encoder for Enhanced Anomaly Detection in Image Data Security", 2024 International Conference on Data Science and Network Security (ICDSNS), ISBN:979-8-3503-7311-0, DOI: 10.1109/ICDSNS62112.2024.10690857, October 2024, IEEE Xplore

[11] Ravindra Changala, "Healthcare Data Management Optimization Using LSTM and GAN-Based Predictive Modeling: Towards Effective Health Service Delivery", 2024 International Conference on Data Science and Network Security (ICDSNS), ISBN:979-8-3503-7311-0, DOI: 10.1109/ICDSNS62112.2024.10690857, October 2024, IEEE Xplore .

[12] Ravindra Changala, "Implementing Genetic Algorithms for Optimization in Neuro-Cognitive Rehabilitation Robotics", 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), ISBN:979-8-3503-7274-8, DOI: 10.1109/ICC-ROBINS60238.2024.10533965, May 2024, IEEE Xplore

[13] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in Proc. Eur. Conf. Comput. Vis., 2010, pp. 143–156.

[14] R. M. Anwer, F. S. Khan, J. Van De Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," ISPRS J. Photogrammetry Remote Sens., vol. 138, pp. 74–85, 2018.

[15] X.Wu, D. Hong, J. Tian, J. Chanussot,W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[16] Ravindra Changala, "Real-Time Anomaly Detection in 5G Networks Through Edge Computing", 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), ISBN:979-8-3503-6118-6, DOI: 10.1109/INCOS59338.2024.10527501, May 2024, IEEE Xplore

[17] Ravindra Changala, "Enhancing Quantum Machine Learning Algorithms for Optimized Financial Portfolio Management", 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), ISBN:979-8-3503-6118-6, DOI: 10.1109/INCOS59338.2024.10527612, May 2024, IEEE Xplore

[18] Ravindra Changala, "Integration of Machine Learning and Computer Vision to Detect and Prevent the Crime", 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), ISBN:979-8-3503-1706-0, DOI: 10.1109/ICCAMS60113.2023.10526105, May 2024, IEEE Xplore

[19] Ravindra Changala, "Controlling the Antenna Signal Fluctuations by Combining the RF-Peak Detector and Real Impedance Mismatch", 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), ISBN:979-8-3503-1706-0, DOI: 10.1109/ICCAMS60113.2023.10526052, May 2024, IEEE Xplore

[20] M. J. Swain and D. H. Ballard, "Color indexing," Int. J. Comput. Vis., vol. 7, no. 1, pp. 11–32, 1991. [11] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," IEEE Trans. Syst., Man, Cybern., vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[21] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst., 2010, pp. 270–279.
[22] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

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

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

[25] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

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

[27] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," IEEE Trans. Image Process., vol. 32, pp. 364–376, Dec. 2022.

[28] Ravindra Changala, Development of Predictive Model for Medical Domains to Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms and Classification Techniques, ARPN Journal of Engineering and Applied Sciences, Volume 14, Issue 6, 2019

[29] Ravindra Changala, "Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods in Text Mining" in ARPN Journal of Engineering and Applied Sciences, Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018

[30] Ravindra Changala "A Survey on Development of Pattern Evolving Model for Discovery of Patterns in Text Mining Using Data Mining Techniques" in Journal of Theoretical and Applied Information Technology, August 2017. Vol.95. No.16, ISSN: 1817-3195, pp.3974-3987

[31] 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.

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

[33] 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.

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

[35] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," IEEE Trans. Geosci. Remote Sens., vol. 58, no. 1, pp. 82–96, Jan. 2020.

[36] 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.

[37] Y. Liu, C. Y. Suen, Y. Liu, and L. Ding, "Scene classification using hierarchical Wasserstein CNN," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 5, pp. 2494–2509, May 2019.

[38] C. Shi, X. Zhang, J. Sun, and L. Wang, "A lightweight convolutional neural network based on group-wise hybrid attention for remote sensing scene classification," Remote Sens., vol. 14, no. 1, 2021, Art. no. 161.

[39] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, pp. 4340–4354, May 2021.

[40] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution," IEEE Trans. Geosci. Remote Sens., vol. 60, Mar. 2022, Art. no. 5620214, doi: 10.1109/TGRS.2022.3160492. BI et al.: VISION TRANSFORMER WITH CONTRASTIVE LEARNING 749.

[41] S. Liu et al., "A shallow-to-deep feature fusion network for VHR remote sensing image classification," IEEE Trans. Geosci. Remote Sens., vol. 60, May 2022, Art. no. 5410213, doi: 10.1109/TGRS.2022.3179288.

[42] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[43] 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, Jan. 2022, doi: 10.1109/JSTARS.2022.3141826.

[44] W. Luo, H. Li, G. Liu, and L. Zeng, "Semantic annotation of satellite images using author–genre–topic model," IEEE Trans. Geosci. Remote Sens., vol. 52, no. 2, pp. 1356–1368, Feb. 2014.

[45] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 3735–3756, 2020, doi: 10.1109/JSTARS.2020.3005403.

[46] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 6, pp. 5103–5113, Jun. 2021.
[47] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017