A Fusion Approach for Enhanced Remote Sensing Image Classification

Vian Abdulmajeed Ahmed[1][a], Khaled Jouini[1][b], Amel Tuama[2] and Ouajdi Korbaa[1][c]

[1]University of Sousse, MARS Research Lab, LR17ES05, ISITCom, 4011 H. Sousse, Tunisia
[2]Northern Technical University, Computer Engineering Techniques Department, Iraq

Keywords: Remote Sensing, Land Cover Mapping, Features Fusion, Convolutional Neural Networks (CNNs), Scale-Invariant Feature Transform (SIFT), Image Classification.

Abstract: Satellite imagery provides a unique and comprehensive view of the Earth's surface, enabling global-scale land cover mapping and environmental monitoring. Despite substantial advancements, satellite imagery analysis remains a highly challenging task due to intrinsic and extrinsic factors, including data volume and variability, atmospheric conditions, sensor characteristics and complex land cover patterns. Early methods in remote sensing image classification leaned on human-engineered descriptors, typified by the widely used Scale-Invariant Feature Transform (SIFT). SIFT and similar approaches had inherent limitations in directly representing entire scenes, driving the use of encoding techniques like the Bag-of-Visual-Words (BoVW). While these encoding methods offer simplicity and efficiency, they are constrained in their representation capabilities. The rise of deep learning, fuelled by abundant data and computing power, revolutionized satellite image analysis, with Convolutional Neural Networks (CNNs) emerging as highly effective tools. Nevertheless, CNNs' extensive need for annotated data limits their scope of application. In this work we investigate the fusion of two distinctive feature extraction methodologies, namely SIFT and CNN, within the framework of Support Vector Machines (SVM). This fusion approach seeks to harness the unique advantages of each feature extraction method while mitigating their individual limitations. SIFT excels at capturing local features critical for identifying specific image characteristics, whereas CNNs enrich representations with global context, spatial relationships and hierarchical features. The integration of SIFT and CNN features helps thus in enhancing resilience to perturbations and generalization across diverse landscapes. An additional advantage is the adaptability of this approach to scenarios with limited labelled data. Experiments on the EuroSAT dataset demonstrate that the proposed fusion approach outperforms SIFT-based and CNN-based models used separately and that it achieves either better or comparable results when compared to existing notable approaches in remote sensing image classification.

1 INTRODUCTION

Satellite imagery plays a pivotal role in various applications, including land cover mapping, environmental monitoring, disaster assessment, and urban planning. Thanks to advances in Earth observation technology, the volume of remote sensing images is rapidly increasing. Understanding these vast and complex images has become an increasingly important and challenging task (Janga et al., 2023). At the heart of this challenge lies scene and images classification, a complex task that has garnered significant attention from the research community. The central goal of remote sensing scene classification is to accurately assign predefined semantic categories to images. Scene classification requires a high degree of accuracy and adaptability, as the scenes encountered in practice are typically diverse, spanning both rural and urban landscapes(Wang et al., 2022). To meet the requirements of these diverse applications, a classification model must be equipped to handle

---

[a] https://orcid.org/0009-0002-5924-6139
[b] https://orcid.org/0000-0002-3802-9074
[c] https://orcid.org/0000-0003-4462-1805

variations in scale, atmospheric conditions, and noise, while also being capable of recognizing complex spatial patterns(Weiss et al., 2020).

In the early stages of remote sensing scene classification, many methods relied heavily on human-crafted features, with Scale-Invariant Feature Transform (SIFT) being a prominent example. SIFT and similar methods faced the challenge of directly representing an entire image due to their inherent local nature (Weinzaepfel et al., 2011). To address this limitation, local descriptors often employed encoding methods, such as the popular Bag-of-Visual-Words (BoVW). While these encoding methods offered simplicity and efficiency, they simultaneously had limited representation capabilities, especially as they do not allow to represent spatial relationships (Cheng et al., 2019). In response to these limitations, unsupervised learning methods, which autonomously learn features from unlabeled images, emerged as an attractive alternative to human-crafted descriptors. These methods, often employing techniques like k-means clustering, presented a promising avenue for scene classification. Nevertheless, unsupervised methods lack the supervised learning's advantage of having class labels to guide the feature learning process, which often lead to learn features that are not relevant to the classification task (Cheng et al., 2019).

The advances in deep learning theory, coupled with the increased availability of remote sensing data and parallel computing resources, ushered in a new era for remote sensing image scene classification. Deep learning models have demonstrated their prowess in feature description across various domains, and remote sensing image scene classification was no exception (Aksoy et al., 2023). Convolutional Neural Networks (CNNs) emerged as a powerful tool, pushing the boundaries of classification accuracy in the field. However, the data-hungry nature of CNNs and their extensive need for annotated training data limit their scope of application.

In this study, we investigate the fusion of two distinctive feature extraction methods, namely SIFT and CNN, within the framework of Support Vector Machines (SVM) for remote sensing images classification. This approach aims to harness the benefits of both feature extraction approaches, while overcoming the limitations of each method used separately. SIFT excels in capturing unique local features that are essential for recognizing unique characteristics within an image. The global context and the hierarchical features learned by CNNs contribute to better generalization, ensuring that the

model can accurately classify scenes exhibiting complex patterns that are challenging to capture with local features alone. An additional advantage of this approach is its adaptability to situations with limited labeled data, a common issue in remote sensing.

The EuroSAT dataset (Cheng et al., 2020) is a widely recognized and extensively employed collection of satellite images containing 10 classes of land cover. It consists of 27,000 images collected by the Sentinel-2 satellite, having a spatial resolution of 10 meters. Experiments on EuroSAT dataset demonstrate that our fusion approach outperforms, not only SIFT and CNN used separately, but also existing remote sensing image classification approaches.

The remainder of this paper is organized as follows. Section II briefly reviews related work. Section III presents our features fusion approach. Section IV provides a comparative experimental study on EuroSAT dataset. Finally, section V concludes the paper.

2 RELATED WORK

Land Use and Land Cover (LULC) classification has garnered substantial attention within the scientific community, with numerous studies and reviews dedicated to the comparison of various approaches and emerging trends. For the sake of conciseness and due to lack of space, we mainly focus in the sequel on approaches that employ the EuroSAT dataset used in our experimental study or presenting similarities with our approach. Existing approaches and studies can be broadly classified into two families: Machine Learning (ML)-based algorithms and Deep Learning (DL)-based methods(Yaloveha et al., 2023).

The studies presented in (Hu et al., 2014), (Chen & Tian, 2015), and (Thakur & Panse, 2022) are representative of ML-based approaches. Hu et al. (2014) proposed a method that utilizes randomly sampled image patches for Unsupervised Feature Learning (UFL) in image classification. They applied the BOVW model to this approach and conducted experiments on an aerial scene dataset. The experiments on the dataset present encouraging results with an accuracy of 90.03%. (Chen & Tian, 2015) introduced the Pyramid of Spatial Relations (PSR) model, designed to incorporate both relative and complete spatial information into the BOVW framework for LULC classification. Experiments conducted on a high-resolution remote sensing image revealed that the PSR model achieves an average classification accuracy of 89.1%. In (Thakur &

Panse, 2022), the authors evaluate the performance of four machine learning algorithms: decision tree (DT), k-nearest neighbor (KNN), support vector machine (SVM), and random forest (RF). The results indicate that RF exhibits superior performance compared to DT, KNN, and SVM, while SVM and DT exhibit similar levels of effectiveness.

The studies (P Helber et al., 2018), (Dewangkoro & Arymurthy, 2021) and (Temenos et al., 2023) are representative of DL-based studies. The authors in (Temenos et al., 2023) introduce an interpretable deep learning framework for LULC classification using SHapley Additive exPlanations (SHAPs). (Temenos et al., 2023) uses a compact CNN model for images classification and then feeds the results to a SHAP deep explainer. Experimental results on the EuroSAT dataset demonstrate the CNN's accurate classification with an overall accuracy of 94.72%, whereas the classification accuracy on three-band combinations on each of the dataset's classes highlight its improvement when compared to standard approaches. The SHAP explainable results of the proposed framework shield the network's predictions by showing correlation values that are relevant to the predicted class, thereby improving the classifications occurring in urban and rural areas with different land uses in the same scene.

Another interesting DL-based study is presented in (Dewangkoro & Arymurthy, 2021). The approach of (Dewangkoro & Arymurthy, 2021) uses different CNN architectures for feature extraction, namely VGG19, ResNet50, and InceptionV3. Then, the extracted feature is recalibrated using Channel Squeeze & Spatial Excitation (sSE) block. The approach also uses SVM and Twin SVM (TWSVM) as classifiers. VGG19 with sSE block and TWSVM achieved the highest experimental results with 94.57% accuracy, 94.40% precision, 94.40% recall, and 94.39% F1-score.

In the study by (P Helber et al., 2018), the authors compares vaious CNN architectures, namelya shallow CNN, a ResNet50-based model model and a GoogleNet-based model. The overall classification accuracy achieved is 89.03%, 98.57%, 98.18% respectively. The authors also evaluated the performance of the Bagof-Visual-Words (BoVW) approach using SIFT features and a trained SVM. The study of (P Helber et al., 2018) shos that all CNN approaches outperform the BoVW method and, overall, deep CNNs perform better than shallow CNNs which achieves an overall accuracy of 89.03% on EuroSAT dataset.

The aforementioned studies demonstrate that there is no one-size-fits-all algorithm that can attain the highest accuracy for all the classes under consideration. Furthermore, as this section highlights, existing approaches tend to concentrate on either classical machine learning methods or deep learning algorithms, with none delving into the advantages that can be derived from integrating classical methods with deep learning algorithms. Our study aims to underscore and quantify the potential benefits of such an integration.

3 FEATURES FUSION VS. HAND-CRAFTED AND CNN-LEARNED FEATURES

Hand-crafted features have played a significant role in computer vision applications, particularly image classification. These features are derived through non-learning processes, directly applying various operations to image pixels. They offer advantages like rotation and scale invariance, achieved by efficiently encoding local gradient information. However, hand-crafted features have three notable limitations (Tsourounis et al., 2022): (*i*) They provide a low-level representation of data and lack the ability to offer an abstract representation crucial for recognition tasks; (*ii*) Local descriptors like SIFT do not yield a fixed-length vector representation of input images, necessitating additional logic for local descriptor encoding, such as Bag-of-Visual-Words (BoVW); and (*iii*) Their capacity is fixed and limited by a predefined mapping from data to feature space, regardless of specific recognition needs.

In the past decade, hand-crafted methods have been largely supplanted by deep Convolutional Neural Networks (CNNs). CNNs employ an end-to-end learning approach, typically in a supervised manner. Each input image is associated with a ground-truth label, and the CNN model's weights are updated iteratively until the model's output aligns with the label. This way, CNNs construct hierarchical feature representations through a learning process that minimizes a defined cost function. CNNs learn feature representation and encoding directly from images, resulting in a model that provides high-level feature representations once trained on a particular dataset and task. However, CNNs demand extensive data and are sensitive to data quality, making them dependent on large annotated datasets while posing challenges related to achieving scale, rotation, or geometric invariance.

In this study, we investigate the synergy between local descriptors (SIFT) and CNN-learned

descriptors. To this end, we compare the fusion of CNN-SIFT features to the cases where SIFT and CNN are used separately. The framework of the proposed models is shown in Figure1. It is worth noting that to gain a more comprehensive understanding of the isolated impact of the fusion approach without any additional considerations or optimizations, we opted for a basic SIFT-based model and a straightforward CNN architecture for feature extraction. While we acknowledge that more complex CNN architectures, such as those used in (Patrick Helber et al., 2019) and (Dewangkoro & Arymurthy, 2021), have the potential to further improve predictive performance, such complex architectures make it difficult to quantify the specific advantages gained from incorporating SIFT-based descriptors. The three models that we study are presented in the sequel.

3.1 Model 1: CNN-Based Remote Sensing Image Classification

The architectural components of the explored CNN model are depicted in Figure 2. The model incorporates two convolutional layers to capture essential image features. The initial convolutional layer operates on input images with dimensions of (64, 64, 3) and employs 32 filters with Rectified Linear Unit (ReLU) activation functions, each having a size of 3x3. This layer effectively extracts fundamental characteristics from the input data. The output features from the first layer are further refined by a second convolutional layer, consisting of 64 filters, each with a size of 3x3. The model integrates two max-pooling layers for spatial dimension reduction. The first max-pooling layer reduces the spatial dimensions of the feature maps by a factor of two, enhancing computational efficiency. The second max-pooling layer further compresses the spatial dimensions, facilitating more abstract feature extraction. A flattened layer precedes the fully connected layers, transforming the 2D feature maps into a 1D vector. The network architecture comprises also two dense layers, with the first layer housing 128 neurons activated by ReLU. This configuration allows the model to learn intricate representations from the data. For multi-class classification tasks, the final layer encompasses ten neurons, utilizing the SoftMax activation function to generate class probabilities. During training, we employ the Adam optimizer and a sparse categorical cross-entropy loss function to optimize the model. The primary objective is to minimize the loss and ensure accurate categorization through the training process. "Rather

than training the investigated CNN model from scratch, we employ transfer learning and use a pre-trained model with weights acquired from the ImageNet dataset (Abou Baker et al., 2022) .

3.2 Model 2: SIFT-Based Remote Sensing Image Classification

The different steps of the second studied approach are illustrated in Figure 3. The first step involves the conversion of the original satellite images into grayscale format, simplifying the data while retaining essential visual characteristics. Following this conversion, the SIFT algorithm is applied to identify key points and extract local feature descriptors. These SIFT descriptors represent distinctive image regions and are crucial for capturing unique visual patterns. To further streamline the feature representation and enable efficient classification, we adopt the Bag-of-Visual-Words (BoVW). Here, the extracted SIFT descriptors are quantized into visual words, reducing the feature dimensionality and forming the basis for image representation. The quantization process is facilitated by k-means clustering, which groups similar descriptors into clusters, and the cluster centers become the visual words. Finally, we employ a Support Vector Machine (SVM) model to train on the BoVW-represented satellite images.

3.3 Model 3: Fusion of SIFT and CNN Features

This paper introduces a novel hybrid model that synergizes the strengths of CNN-learned features with SIFT descriptors to enhance remote sensing image classification (Figure 4). The proposed approach harnesses the power of the CNN to extract high-level, semantically rich features, providing a global understanding of the image. It also employs the SIFT detector (Open CV SIFT) to capture fine-grained, local details, benefiting from its robustness to various transformations (*e.g.* scale and rotation).

The SIFT features are flattened and have lengths that are either zero-padded or truncated to 128. To extract deep features, we leverage a pre-trained CNN architecture with weights sourced from ImageNet. The base model is modified by excluding the final fully connected layers (*i.e. Include top=False*), retaining only the convolutional layers. The input images, which are initially of varying dimensions, are pre-processed and resized to meet the (224, 224, 3) input shape requirement of the CNN model.

Once the SIFT and the CNN features are generated, a unified feature vector is produced by

horizontally stacking the CNN features with the truncated/flattened SIFT features. Each image is represented by this feature vector combination, which is a rich representation, encapsulating both global and local information. For the task of classification, we employ a straightforward Support Vector Machine (SVM) with a radial basis function (RBF) kernel. The RBF kernel's flexibility enables the model to capture complex decision boundaries in the feature space. As demonstrated in our experimental study, this synergy between deep learning-based features from the CNN a conventional computer vision characteristic from SIFT yields enhanced classification performance.
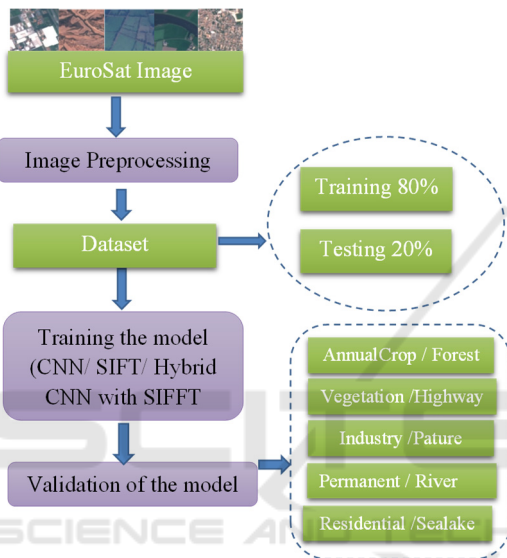

Figure 1: Frame Work for the proposed model.

Input (64x64x3 Image Data)
Conv2D (32) 3x3,ReLU
MaxPooling 2x2
Conv2D (64) 3x3, ReLU
MaxPooling 2x2
Flatten
Dense (128) | ReLU
Dense (10) Softmax
Output (10 classes)

Figure 2: Proposed CNN architecture for classification Remote sensing images.
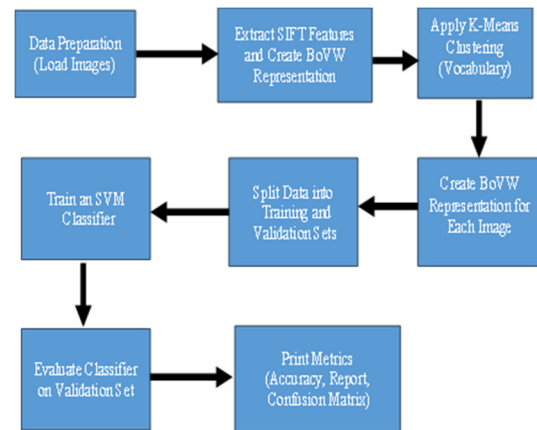

Figure 3: Proposed SIFT procedure for classification Remote Sensing Images.
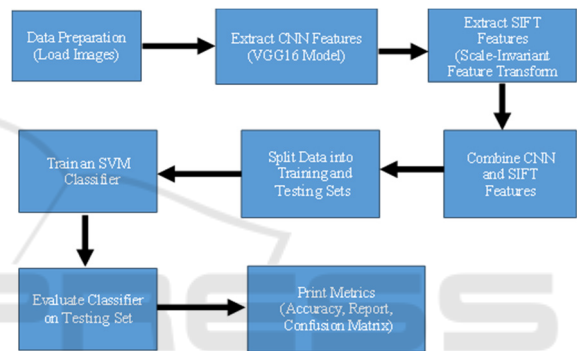
Figure 4: Proposed Hybrid CNN with SIFT models.

4 EXPERIMENTAL STUDY

The EuroSAT dataset (Patrick Helber et al., 2019) used in our experiments, is openly and freely provided by the Copernicus Earth observation program. The dataset is generated with 27,000 labeled and georeferenced image patches, where the size of each image patch is 64_64 m. To ech of the 10 classes of the dataset corresponds 2000 to 3000 images. The LULC classes in this dataset are *permanent crop*, *annual crop*, *pastures*, *river*, *sea & lake*, *forest*, *herbaceous vegetation*, *industrial building*, *highway* and *residential building* (Patrick Helber et al., 2019).

In our experiments, 80% of the dataset was allocated for training, while the remaining instances were reserved for testing. The studied CNN architecture was implemented using Tensorflow (Yu et al., 2019) and Keras (Lee & Song, 2019). OpenCV (Culjak et al., 2012) was used for SIFT features generation. All methods were run using their default settings, and no special tuning was done.

Figure 5: Sample images of Eurosat dataset (Helber et al., 2019).

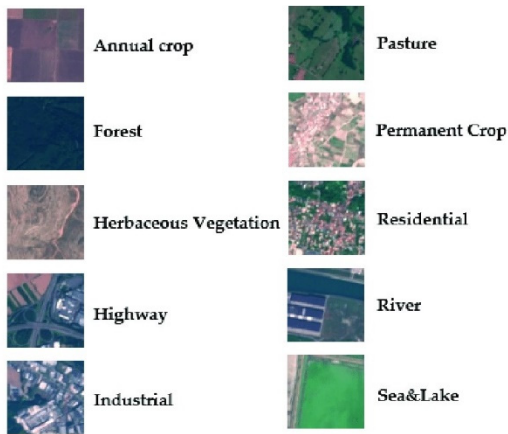The primary goal of our study is to bring to light the advantages that can be drawn from combining SIFT and CNN features. We then compare and contrast three different remote sensing classification models: CNN-based, SIFT-based, and a fusion approach combining SIFT and CNN features. Figures 6, 7, 8, 9, 10 and 11 illustrate, respectively, the confusion matrix and the detailed classification report of the CNN-based model (Model 1), The SIFT-based model (Model 2) and the model based on a fusion of CNN and SIFT features (Model 3). Table 4.1. provides the overall accuracy of the different models. As shown in these figures and in Table 4.1, the features fusion approach by far outperforms the SIFT-based model and the CNN-based model, allowing an enhancement of accuracy of 64.29% and 84.10% respectively.

Table 4.2 compares the accuracy results of our models with some of notable existing approaches. As shown in Table 4.2 our approach achieves an overall accuracy of 92%, and, except for the approach presented in (Dewangkoro & Arymurthy, 2021), it outperforms all other models. The "BoVW (SVM, SIFT, k = 500)" model, based on BoVW with SVM and SIFT, achieves an overall accuracy of 70%. The "UFL" model achieves an overall accuracy of 90%, demonstrating the effectiveness of unsupervised feature learning. The "Pyramid of Spatial Relations"(Chen & Tian, 2015) model reaches 89% accuracy, emphasizing the importance of capturing spatial relationships. The approach presented in (Dewangkoro & Arymurthy, 2021) uses different CNN architectures for feature extraction, namely VGG19, ResNet50, and InceptionV3, and achieves an accuracy of 94%. The approach (Dewangkoro & Arymurthy, 2021) inherits the advantages and limitations of deep neural architectures. Our approach achieves comparable performance to that of

(Dewangkoro & Arymurthy, 2021) while being less resource-intensive and less reliant on the availability of massive labelled data.

As mentioned earlier, in order to better understand the impact of the fusion approach in isolation, we implemented a basic SIFT-based model and a simple CNN architecture for feature extraction. However, it is worth noting the use in our approach of more sophisticated CNN architectures, such as those used in (Helber et al., 2019) and (Dewangkoro & Arymurthy, 2021), have the potential to further enhance predictive performance.

```
[[526    2   16   13    0   15   16    0    4    3]
 [  0  577    2    0    0   21    0    0    1    5]
 [  6    3  479   16    3    8   71   10    5    1]
 [ 16    0   23  358    6   11   42   14   45    0]
 [  0    0   10   41  396    0    7   38    1    0]
 [ 15    5   20   13    0  350    8    0    9    3]
 [ 23    0   54   33    1    9  345    2    6    0]
 [  0    0   14    9    3    0    6  565    0    0]
 [ 37   12   13   84    1   13    6    3  320    1]
 [  4    4    1    1    0    3    0    0    7  586]]
```

Figure 6: Confusion Matrix for CNN model for remote sensing images classification.

	precision	recall	f1-score	support
0	0.84	0.88	0.86	595
1	0.96	0.95	0.95	606
2	0.76	0.80	0.78	602
3	0.63	0.70	0.66	515
4	0.97	0.80	0.88	493
5	0.81	0.83	0.82	423
6	0.69	0.73	0.71	473
7	0.89	0.95	0.92	597
8	0.80	0.65	0.72	490
9	0.98	0.97	0.97	606

Figure 7: Precision, Recall, F1-Score for CNN model for remote sensing images classification.

	precision	recall	f1-score	support
AnnualCrop	0.65	0.71	0.68	520
Forest	0.00	0.00	0.00	66
HerbaceousVegetation	0.40	0.36	0.38	461
Highway	0.53	0.45	0.49	483
Industrial	0.74	0.86	0.79	527
Pasture	0.28	0.65	0.40	244
PermanentCrop	0.51	0.41	0.45	483
Residential	0.77	0.78	0.78	592
River	0.42	0.29	0.35	458
SeaLake	0.00	0.00	0.00	54

Figure 8: Precision, Recall, F1-Score for SIFT model for remote sensing images classification.

```
[[371    0   20   15    0   67   12    0   35    0]
 [  2    0    5    0    0   57    0    0    2    0]
 [ 21    1  167   29   39  100   51   24   29    0]
 [ 46    0   36  216   25   20   56   23   61    0]
 [  2    0    3   15  454    0   21   30    2    0]
 [ 25    0   28    2    0  159    6   13   11    0]
 [ 30    0   59   39   69   35  198   28   25    0]
 [  2    0   34   17   27   24   14  464   10    0]
 [ 66    0   57   73    3   73   34   17  135    0]
 [  7    0    7    1    0   26    0    1   12    0]]
```

Figure 9: Confusion Matrix for SIFT model for remote sensing images classification.

	precision	recall	f1-score	support
AnnualCrop	0.91	0.97	0.94	534
Forest	0.88	0.93	0.90	69
HerbaceousVegetation	0.89	0.90	0.89	473
Highway	0.90	0.87	0.89	479
Industrial	0.93	0.96	0.94	512
Pasture	0.93	0.90	0.91	251
PermanentCrop	0.90	0.87	0.88	486
Residential	0.97	0.97	0.97	591
River	0.91	0.88	0.89	441
SeaLake	0.96	0.88	0.92	52

Figure 10: Precision, Recall, F1-Score for Hybrid CNN with SIFT model for remote sensing images classification.

```
[[516    1    4    1    0    2    5    0    4    1]
 [  0   64    1    0    0    3    0    0    1    0]
 [  2    2  425    7    5    0   20    7    5    0]
 [ 12    2    2  418    5    4    9    2   24    1]
 [  0    0    1    6  489    0    7    9    0    0]
 [  5    3   11    0    0  225    4    0    3    0]
 [ 16    0   28    6   11    1  422    1    1    0]
 [  0    1    3    0   12    0    1  574    0    0]
 [ 13    0    3   27    1    7    2    0  388    0]
 [  3    0    0    0    0    1    0    0    2   46]]
```

Figure 11: Confusion Matrix for Hybrid CNN with SIFT model for remote sensing images classification.

Table 4.1: The results of accuracy for the proposed models.

Model	Accuracy
CNN	0.83
SIFT	0.56
Fusion of CNNs and SIFT features	**0.92**

Table 4.2: The accuracy Result of the proposed model compare with related work.

Models	Accuracy
CNN two layer (Helber et al., 2018)	0.87
BoVW (SVM, SIFT, k = 500) (Helber et al., 2018)	0.70
UFL (Hu et al., 2014)	0.90
Pyramid of spatial relatons(Chen & Tian, 2015)	0.89
Combination of different CNNs deep architectures (Dewangkoro & Arymurthy, 2021)	0.94
Fusion of CNNs and SIFT features (proposed model)	**0.92**

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

Remote sensing image classification is a crucial task for various critical applications, including land cover mapping, environmental monitoring, and disaster response. In this work we investigated the fusion of hand-crafted features (SIFT) and CNN-learned features to enhance remote sensing image classification. SIFT excels in capturing local features essential for discerning specific image attributes. Meanwhile, CNNs' ability to learn global context and hierarchical features, enhances generalization and allows accurate classification of scenes with complex patterns. The experimental study conducted over the EuroSAT dataset shows that our fusion approach allows a substantial classification enhancement with regards to CNN and SIFT used separately: up to 10.84% accuracy enhancement when compared to CNN and up to 64.29% enhancement when compared to SIFT. Although our fusion approach was implemented using straightforward SIFT-based Model and CNN architecture (to better isolate the benefits of features fusion), our experimental study shows that it achieves better or comparable results with notable existing remote sensing image classification approaches.

5.2 Future Work

The promising obtained results pave the way for the exploration of other applications and further forms of collaboration between classical hand-crafted features and modern deep features. We are currently exploring two research directions. The first involves remote sensing images enhancement during registration, which aims at improving the quality of remote sensing images to make them more amenable to subsequent analysis. The second focuses on the detection of changes in images captured within the same geographic areas but at different time points. Such change detection is crucial several critical in domains such as environmental monitoring. To this end we are currently investigating the integration of SIFT and Siamese networks for efficient change detection in remote sensing image.

REFERENCES

Abou Baker, N., Zengeler, N., & Handmann, U. (2022). A Transfer Learning Evaluation of Deep Neural Networks for Image Classification. In *Machine Learning and Knowledge Extraction* (Vol. 4, Issue 1,pp. 22–41). https://doi.org/10.3390/make4010002

Aksoy, M. Ç., Sirmacek, B., & Ünsalan, C. (2023). Land classification in satellite images by injecting traditional features to CNN models. *Remote Sensing Letters*, *14*(2), 157–167. https://doi.org/10.1080/2150704X.2023.2167057

Chen, S., & Tian, Y. (2015). Pyramid of Spatial Relatons for Scene-Level Land Use Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(4), 1947–1957. https://doi.org/10.1109/TGRS.2014.2351395

Cheng, G, Xie, X., Han, J., Guo, L., & Xia, G.-S. (2020). Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3735–3756. https://doi.org/10.1109/JSTARS.2020.3005403

Cheng, Gong, Xie, X., Han, J., Guo, L., & Xia, G. S. (2019). Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*(X), 37353756. https://doi.org/10.1109/JSTARS.2020.3005403

Culjak, I., Abram, D., Pribanic, T., Dzapo, H., & Cifrek, M. (2012). A brief introduction to OpenCV. *2012 Proceedings of the 35th International Convention MIPRO*, 1725–1730.

Dewangkoro, H. I., & Arymurthy, A. M. (2021). Land use and land cover classification using CNN, SVM, and Channel Squeeze & Spatial Excitation block. *IOP Conference Series: Earth and Environmental Science*, *704*(1). https://doi.org/10.1088/1755-1315/704/1/012048

Helber, P, Bischke, B., Dengel, A., & Borth, D. (2018). Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 204–207. https://doi.org/10.1109/IGARSS.2018.8519248

Helber, Patrick, Bischke, B., Dengel, A., & Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(7), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

Hu, F., Xia, G.-S., Wang, Z., Zhang, L., & Sun, H. (2014). Unsupervised feature coding on local patch manifold for satellite image scene classification. *2014 IEEE Geoscience and Remote Sensing Symposium*, 1273–1276. https://doi.org/10.1109/IGARSS.2014.6946665

Janga, B., Asamani, G. P., Sun, Z., & Cristea, N. (2023). A Review of Practical AI for Remote Sensing in Earth Sciences. In *Remote Sensing* (Vol. 15, Issue 16). https://doi.org/10.3390/rs15164112

Lee, H., & Song, J. (2019). Introduction to convolutional neural network using Keras; an understanding from a statistician. *Communications for Statistical Applications and Methods*, *26*(6), 591–610.

Temenos, A., Temenos, N., Kaselimi, M., Doulamis, A., & Doulamis, N. (2023). Interpretable Deep Learning Framework for Land Use and Land Cover Classification in Remote Sensing Using SHAP. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3251652

Thakur, R., & Panse, P. (2022). Classification Performance of Land Use from Multispectral Remote Sensing Images using Decision Tree, K-Nearest Neighbor, Random Forest and Support Vector Machine Using EuroSAT. *nternational Journal of Intelligent Systems and Applications in Engineering IJISAE*, *2022*(1s),67–77. https://github.com/phelber/EuroSAT

Tsourounis, D., Kastaniotis, D., Theoharatos, C., Kazantzidis, A., & Economou, G. (2022). SIFT-CNN: When Convolutional Neural Networks Meet Dense SIFT Descriptors for Image and Sequence Classification. *Journal of Imaging*, *8*(10). https://doi.org/10.3390/jimaging8100256

Wang, X., Xu, H., Yuan, L., Dai, W., & Wen, X. (2022). A Remote-Sensing Scene-Image Classification Method Based on Deep Multiple-Instance Learning with a Residual Dense Attention ConvNet. In *Remote Sensing* (Vol. 14, Issue 20). https://doi.org/10.3390/rs14205095

Weinzaepfel, P., Jégou, H., & Pérez, P. (2011). Reconstructing an image from its local descriptors. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 337–344. https://doi.org/10.1109/CVPR.2011.5995616

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, *236*, 111402. https://doi.org/https://doi.org/10.1016/j.rse.2019.111402

Yaloveha, V., Podorozhniak, A., Kuchuk, H., & Garashchuk, N. (2023). Performance Comparison of Cnns on High-Resolution Multispectral Dataset Applied To Land Cover Classification Problem. *Radioelectronic and Computer Systems*, *2023*(2(106)), 107–118. https://doi.org/10.32620/REKS.2023.2.09

Yu, L., Li, B., & Jiao, B. (2019). Research and Implementation of CNN Based on TensorFlow. *IOP Conference Series: Materials Science and Engineering*, *490*, 42022. https://doi.org/10.1088/1757-899X/490/4/042022.