Enhancing LULC Classification with Attention-Based Fusion of Handcrafted and Deep Features

Vian Abdulmajeed Ahmed[a], Khaled Jouini[b] and Ouajdi Korbaa[c]
MARS Research Lab, LR17ES05, ISITCom, University of Sousse, Sousse, Tunisia

Keywords: Land Use and Land Cover Classification, Attention-Based Fusion, Early and Late Fusion, Multi-Modal Learning.

Abstract: Satellite imagery provides a unique perspective of the Earth's surface, pivotal for applications like environmental monitoring and urban planning. Despite significant advancements, analyzing satellite imagery remains challenging due to complex and variable land cover patterns. Traditional handcrafted descriptors like Scale-Invariant Feature Transform (SIFT) excel at capturing local features but often fail to capture the global context. Conversely, Convolutional Neural Networks (CNNs) excel at capturing rich contextual information but may miss crucial local features due to limitations in capturing small and subtle spatial arrangements. Most existing Land Use and Land Cover (LULC) classification approaches heavily rely on fine-tuning large pretrained models. While this remains a powerful tool, this paper explores alternative strategies by leveraging the complementary strengths of handcrafted and CNN-learned features. Specifically, we investigate and compare three fusion strategies: (*i*) early fusion, where handcrafted and CNN-learned features are merged at the input level; (*ii*) late fusion, where attention mechanisms dynamically integrate salient features from both CNN and SIFT modalities; and (*iii*) mid-level fusion, where attention is used to generate two feature maps: one prioritizing global context and another, weighted by SIFT features, emphasizing local details. Experiments on the real-world EuroSAT dataset demonstrate that these fusion approaches exhibit varying levels of effectiveness and that a well-chosen fusion strategy not only substantially outperforms the underlying methods used separately but also offers an interesting alternative to solely relying on fine-tuning pre-trained large models.

1 INTRODUCTION

Satellite imagery underpins critical applications like land cover mapping, environmental monitoring, disaster response, and urban planning (Ahmed et al., 2024). At the heart of these applications lies *Land Use and Land Cover* (LULC) classification, which involves assigning predefined semantic classes to remote sensing images. Effective LULC classification requires the capability to discern complex spatial patterns while maintaining robustness against variations in scale, atmospheric conditions, and noise (Xia and Liu, 2019). Early methods in LULC classification relied heavily on handcrafted descriptors like *Scale-Invariant Feature Transform* (SIFT) (Lowe, 2004) and similar approaches, which excel at capturing distinctive local features such as edges and textured regions. These methods, however, often struggle to capture the complex spatial and contextual information present in remote sensing images due to their local nature (Cheng et al., 2019). The advent of deep learning and its models trained on large datasets has revolutionized image classification, achieving accuracy levels far beyond traditional methods. The impressive performance of these models, coupled with their data-intensive nature, has directed much of the current work on LULC classification towards *transfer learning* (Dewangkoro and Arymurthy, 2021; Helber et al., 2019; Wang et al., 2024; Neumann et al., 2020), which involves fine-tuning pre-trained large models on remote sensing datasets.

The *convolutional layers* of deep models operate by applying learned filters (small grids of weights) that slide across the image. As these filters move, their weights are multiplied element-wise with pixel values, and the results are summed to create a *feature map*. The stacking of convolutional layers enables deep models to learn increasingly complex patterns: early layers typically capture basic elements

[a] https://orcid.org/0009-0002-5924-6139
[b] https://orcid.org/0000-0001-5049-4238
[c] https://orcid.org/0000-0003-4462-1805

Ahmed, V. A., Jouini, K. and Korbaa, O.
Enhancing LULC Classification with Attention-Based Fusion of Handcrafted and Deep Features.
DOI: 10.5220/0013098500003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 69-78
ISBN: 978-989-758-737-5; ISSN: 2184-433X

like edges and corners, while subsequent layers interpret these elements to recognize objects. This hierarchical learning process allows Convolutional Neural Networks (CNNs) to excel at capturing global contexts and spatial relationships. Despite these capabilities, CNNs can struggle to preserve very fine-grained details like small textures or subtle variations in color due to the pooling operations that often follow convolutional layers.

In this work, we aim to investigate and quantify the potential benefits of synergizing the complementary strengths of handcrafted SIFT descriptors and features learned from CNNs. Fusing these features is intended to leverage both the local detailed cues provided by SIFT and the global context captured by CNNs. Specifically, we investigate three fusion strategies: a straightforward early fusion approach, and novel late and mid-level fusion approaches integrating *attention mechanisms*. Attention mechanisms enable neural networks to prioritize informative input elements by assigning them weights that reflect their relative importance. The late fusion approach uses attention to dynamically weigh and integrate salient features from both the CNN and SIFT modalities before making final classification decisions.The mid-level fusion approach generates two distinct feature maps: one prioritizing global context and another locally-attended feature map weighted according to SIFT features, emphasizing local details. Our experimental study on the real-world EuroSAT dataset reveals that the different fusion approaches vary in effectiveness. Our study also suggests, that while the prevalent fine-tuning of pre-trained models remains a powerful tool for LULC classification, alternatives such as integrating handcrafted and CNN-learned features warrant exploration.

The remainder of this paper is organized as follows. Section 2 provides a concise overview of previous research. Section 3 introduces our features fusion approaches. Section 4 presents a comparative experimental analysis. Finally, Section 5 concludes the paper.

2 RELATED WORK

The EuroSAT dataset (Helber et al., 2018) is a widely recognized and extensively used dataset for LULC classification. It includes 27,000 geotagged image patches, each covering an area of 64x64 meters with a spatial resolution of 10 meters. The dataset comprises ten distinct classes, with each class including 2,000 to 3,000 images. As illustrated in Figure 1, these classes represent a diverse range of land use and land cover types. For the sake of conciseness and due to lack of space, we mainly focus in the sequel on approaches presenting similarities with our work or that use EuroSAT. Existing remote sensing image classification approaches and studies can be broadly classified into two families: Machine Learning (ML)-based and Deep Learning (DL)-based methods.

The study by Chen & Tian (Chen and Tian, 2015), and Thakur & Panse (Thakur and Panse, 2022) are representative of ML-based approaches. (Chen and Tian, 2015) introduced the Pyramid of Spatial Relations (PSR) model, designed to incorporate both relative and complete spatial information into the BoVW (*i.e.* Bag of Visual Words) framework. Experiments conducted on a high-resolution remote sensing image revealed that the PSR model achieves an average classification accuracy of 89.1%. In (Thakur and Panse, 2022), the performance of four machine learning algorithms was evaluated on the EuroSAT dataset: Decision Tree (DT), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Random Forest (RF). The study revealed distinct performance levels among the algorithms: RF achieved the highest overall accuracy of 56.70%, significantly outperforming DT and KNN.

The studies (Temenos et al., 2023), (Dewangkoro and Arymurthy, 2021), (Helber et al., 2019), (Wang et al., 2024) and (Neumann et al., 2020) are representative of DL-based approaches. In (Temenos et al., 2023), the authors introduce an interpretable DL framework for LULC classification using SHapley Additive exPlanations (SHAPs). They employ a compact CNN model for image classification, followed by feeding the results to a SHAP deep explainer, achieving an overall accuracy of 94.72% on EuroSAT. The approach in (Dewangkoro and Arymurthy, 2021) utilizes different CNN architectures for feature extraction, including VGG19, ResNet50, and InceptionV3. These extracted features are then recalibrated using the Channel Squeeze & Spatial Excitation (sSE) block, with Twin SVM (TWSVM) serving as classifier, achieving an accuracy of 94.39% on EuroSAT. In (Helber et al., 2019), various CNN architectures were compared, including a shallow CNN, a ResNet50-based model, and a GoogleNet-based model. The achieved classification accuracies on EuroSAT were 89.03%, 98.57%, and 98.18%, respectively. (Neumann et al., 2020) explored in-domain fine-tuning using five diverse remote sensing datasets and the ResNet50V2 architecture. (Neumann et al., 2020) demonstrated that models fine-tuned on in-domain datasets significantly outperform those pretrained on general purpose datasets like ImageNet. The pretrained ResNet50v2 fine-tuned on in-domain

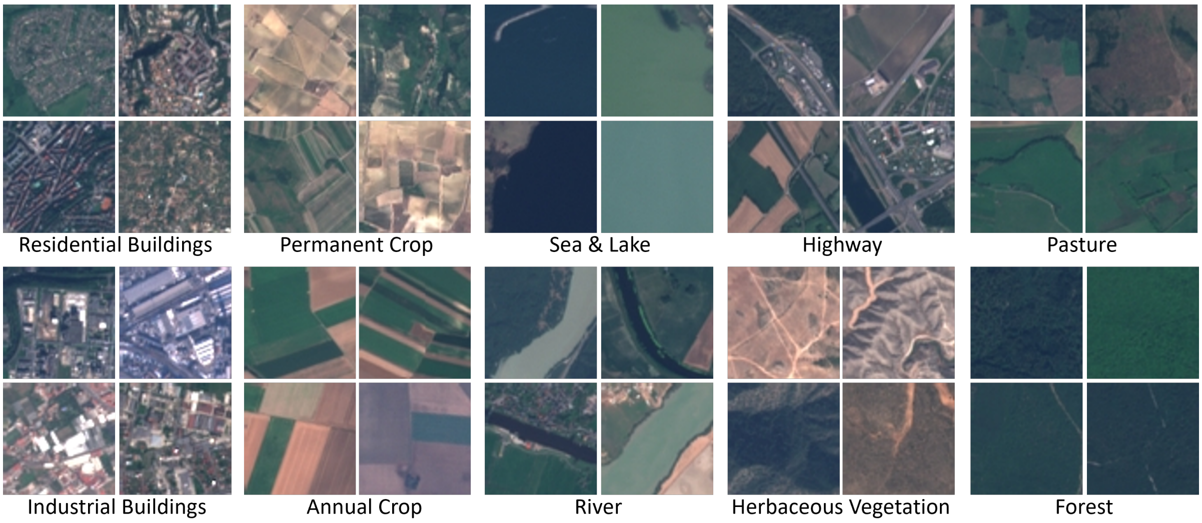

Figure 1: Sample Images Extracted from the EuroSAT Dataset (Helber et al., 2019).

datasets achieved an overall accuracy of 99.2% on EuroSAT.

While transfer learning typically involves adapting a pre-trained model to improve performance on a related dataset, knowledge transfer involves training a single model on multiple tasks simultaneously and leveraging shared representations and knowledge across these tasks. (Gesmundo and Dean, 2022) used knowledge transfer and employed a multitask learning framework in which the model learns from diverse remote sensing datasets concurrently. The evolutionary "mutant multitask network" (µ2Net), introduced by (Gesmundo and Dean, 2022), enhances model efficiency and quality through effective knowledge transfer mechanisms while addressing common challenges such as catastrophic forgetting and negative transfer. Empirical results demonstrate that µ2Net can achieve competitive performance across various image classification tasks. Specifically, on EuroSAT, µ2Net achieved a high classification accuracy of 99.2%. Knowledge transfer is also used by (Wang et al., 2024), where Vision Transformers (ViT) (Steiner et al., 2021) with Rotatable Variance Scaled Attention (RVSA) are used as part of a Multi-Task Pretraining (MTP) framework. When evaluated on EuroSAT, the MTP-enhanced model achieved a high accuracy of 99.2%.

As outlined in this section, most existing LULC classification approaches typically focus on either classical ML or DL methods. Studies (Tianyu et al., 2018) and (Ahmed et al., 2024) have demonstrated the benefits of combining handcrafted and CNN-learned features on the general-purpose CIFAR dataset and EuroSAT dataset, respectively. However, these studies only explored a straightforward early fusion approach. Our work proposes more advanced attention-based fusion methods that can potentially learn to focus on more discriminative features for improved LULC classification accuracy.

3 SYNERGIZING HANDCRAFTED AND CNN LEARENED FEATURES

As mentioned earlier, SIFT is adept at capturing intricate local details and textures but falls short in interpreting broader scene contexts. In contrast, CNNs excel at understanding contextual information and spatial relationships, yet they may overlook fine-grained details. Integrating these features potentially allows the model mitigating the limitations of each method when used alone.

The remainder of this section explores three distinct fusion strategies: (straightforward) early fusion, and novel late and mid-level fusion with attention mechanisms. Broadly, early fusion directly combines features extracted from different modalities before feeding them into a classifier. Conversely, late fusion extracts features independently using separate models for each modality and then fuses these features before classification. Mid-level fusion partially extracts features from each modality before allowing information exchange between them, enabling them to influence each other's feature learning process.

3.1 Baseline Models and Early Fusion Approach

This section provides an overview of the baseline models and the early fusion approach used in our experiments, establishing a foundation for understanding the more advanced late and mid-level fusion approaches discussed later in this paper.

SIFT identifies keypoints in an image that remain stable under scale, rotation, and illumination changes. Initially, SIFT creates a scale-space representation of the image by convolving it with Gaussian filters at multiple scales. Keypoints are then localized as local extrema (peaks or valleys) in the Difference-of-Gaussian (DoG) images computed across these scales (Lowe, 2004). Keypoints are typically found at corners, edges, or distinct texture patterns (Lowe, 2004) and in areas with significant variations in intensity across different directions. In the context of satellite images, keypoints often correspond to transitions between different land covers. To enhance accuracy, each keypoint's precise position and scale are refined through interpolation to achieve subpixel accuracy. Once keypoints are identified, SIFT computes a descriptor for each of them. This descriptor encapsulates information about the gradients or directional changes in intensity surrounding that keypoint within a localized patch of the image (Lowe, 2004). The standard SIFT descriptor is generated by creating a histogram of gradient orientations within this patch, divided into a 4×4 grid, with each of the 16 cells contributing eight orientation bins, resulting in a 128-element descriptor vector. SIFT identifies potentially hundreds or thousands of keypoints per image. To simplify data representation and ensure compatibility with most machine learning algorithms, all individual keypoint descriptors are typically concatenated into a single row vector (*flattening*).

Figures 2 and 3 illustrate the baseline models employed in our work. The first baseline model is a neural network that takes SIFT descriptors as input. It follows a common architecture with two dense layers for feature processing, using ReLU (Rectified Linear Unit) activation functions to introduce non-linearity. Batch normalization is incorporated to stabilize training, and dropout with L2 regularization are applied to prevent overfitting. The second baseline model is a convolutional neural network (CNN) that takes RGB images as input. It features a standard architecture comprising convolutional layers for feature extraction, pooling layers for downsampling, a flattening layer for feature vector transformation, and fully-connected layers for classification. ReLU activation functions are used throughout the network to introduce non-linearity, and dropout with L2 regularization are applied to prevent overfitting.

Figure 4 illustrates the early fusion model used in our work. Early fusion is a prevalent approach for combining features extracted from different modalities (Ahmed et al., 2024) and consists in combining these features before feeding them into the higher-level layers of a neural network. Despite its simplicity, early fusion can achieve good accuracy because it enables the model to learn a unified representation that leverages information from both RGB images and SIFT descriptors during the training process (Ahmed et al., 2024). As shown in Figure 4, the early fusion model we employed comprises two distinct branches, each processing a different modality. After feature extraction in each branch, the model concatenates them. This fused feature vector is then passed through standard neural network layers that integrate regularization techniques (dropout, L2). The final layer employs a softmax activation function for multi-class classification.

3.2 Late Fusion: Attention-Enhanced Dual Learning (ADL)

Attention mechanisms enable neural network models to selectively focus on informative aspects within the input data. They achieve this by learning a set of weights for different parts of the input, indicating their relative importance. Broadly, attention mechanisms operate by calculating scores (*e.g.*, element-wise multiplication, dot products) reflecting the potential relevance of each element in the input. These scores are learned dynamically based on the context. A softmax function is then applied to the scores, normalizing them into a probability distribution. The resulting weights sum to 1 and represent the relative importance of each element as a probability. Finally, the original input elements are multiplied by their corresponding weights and then summed. This creates a weighted representation of the input, emphasizing informative aspects based on the learned attention weights.

As depicted in Figure 5, in our proposed late fusion model, features are extracted from each modality independently using separate models, and then the outputs of the branches are fused at the very end of the network before classification. To improve feature learning, our late fusion model leverages adequate attention mechanisms in both branches. A *channel-wise attention* is integrated into the CNN of the RGB branch through *Squeeze-and-Excitation* (SE) (Hu et al., 2018) blocks. The SE block dynamically adjusts the importance of each channel within

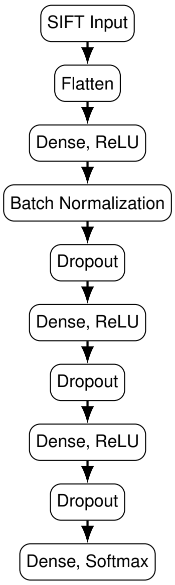

Figure 2: Baseline SIFT-NN Model

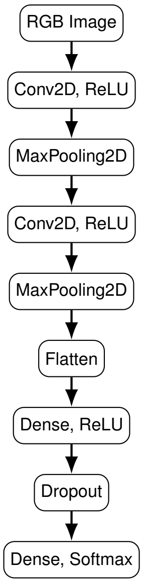

Figure 3: Baseline Shallow CNN Model

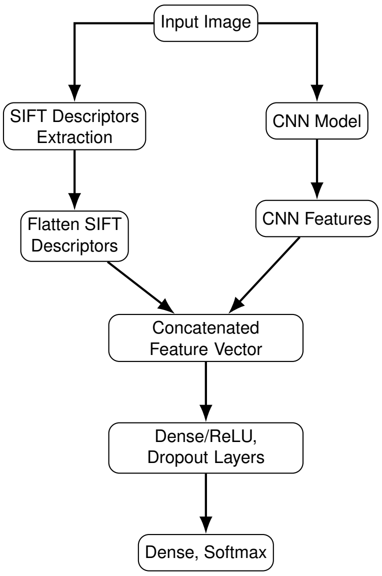
Input Image
SIFT Descriptors Extraction
CNN Model
Flatten SIFT Descriptors
CNN Features
Concatenated Feature Vector
Dense/ReLU, Dropout Layers
Dense, Softmax

Figure 4: Early Fusion of SIFT and CNN Features.

the feature maps. The process involves the following steps (Hu et al., 2018):

1. *Squeeze*: Global average pooling is applied to each feature map, reducing each channel to a single value: $z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_{i,j,c}$, where $x_{i,j,c}$ represents the value at position (i, j) in channel c, and H and W are the height and width of the feature map.
2. *Excitation*: A gating mechanism with a bottleneck structure (two fully connected layers) is applied to capture channel-wise dependencies: $s = \sigma(W_2\delta(W_1 z))$, where W_1 and W_2 are the weight matrices, δ denotes the ReLU activation function, and σ denotes the sigmoid activation function.
3. *Recalibration*: The original feature map is scaled by the learned channel weights: $\tilde{x}_{i,j,c} = s_c \cdot x_{i,j,c}$

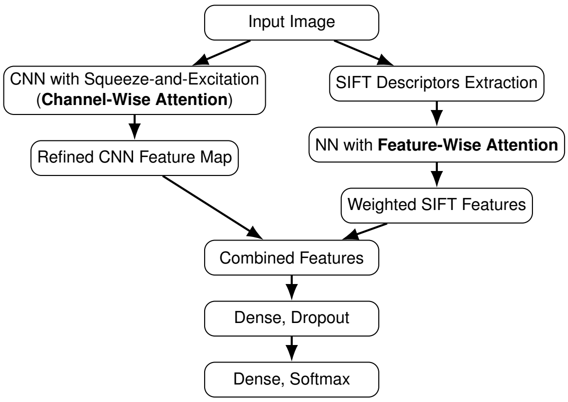
Input Image
CNN with Squeeze-and-Excitation (Channel-Wise Attention)
SIFT Descriptors Extraction
Refined CNN Feature Map
NN with Feature-Wise Attention
Weighted SIFT Features
Combined Features
Dense, Dropout
Dense, Softmax

Figure 5: Late Fusion Approach : Attention-Enhanced Dual Learning (ADL).

In the SIFT branch, SIFT descriptors are processed using a neural network integrating a *feature-wise attention layer* that assigns weights to descriptors to emphasize the most informative ones. By assigning higher weights to relevant SIFT descriptors, the attention mechanism emphasizes features that are particularly informative for specific LULC classes. This complements the focus on global spatial relationships learned by the CNN in the RGB branch. Formally, each SIFT descriptor x_i is transformed into query Q_i, key K_i, and value V_i vectors using learned linear transformations:

$$Q_i = W_Q x_i + b_Q, \quad K_i = W_K x_i + b_K, \quad V_i = W_V x_i + b_V$$

where W_Q, W_K, W_V are weight matrices and b_Q, b_K, b_V are bias vectors. These transformations enable the network to effectively compute attention scores, which measure the relevance of each feature within the descriptor relative to others. The attention score for each feature within the SIFT descriptor is computed as the dot product of the query vector with all key vectors: $score_{ij} = Q_i \cdot K_j$. This results in a matrix of attention scores indicating the relevance of each feature with respect to all others. The subsequent softmax normalization of these scores produces attention weights that denote the significance of each descriptor element: $\alpha_{ij} = \frac{\exp(score_{ij})}{\sum_k \exp(score_{ik})}$. Ultimately, a weighted sum of the value vectors, weighted by these attention weights, results in a refined representation of the SIFT descriptors. The final output of the attention mechanism is the weighted sum of the value vectors, where the weights are the normalized atten-

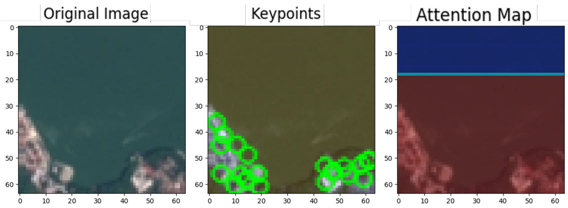

Figure 6: Example Illustrating SIFT-Guided Attention.

tion scores: $attended\_features_i = \sum_j \alpha_{ij} V_j$.

After the independent processing in each branch, the enriched feature outputs from the RGB branch (with channel-wise attention) and the SIFT branch (with feature-wise attention) are concatenated. The fused features combine the spatial relationships captured by the CNN with the local variations captured by the SIFT descriptors, enriched by their respective attention mechanisms. The concatenated feature vector serves as input to dense layers with regularization techniques (dropout, L2).

3.3 Mid-Level Fusion: Fusion of Local Attended CNN Features and Global CNN Features (LFGF) with Gating Mechanism

Early fusion as well as late fusion combine independent feature representations from CNN and SIFT for classification, considering both global and local features equally important and informative. In this section, we propose a novel mid-level fusion approach with the same aim of leveraging local SIFT cues to help identify potentially informative regions, but with a different rationale.

Instead of concatenating separate SIFT and CNN features, the proposed mid-level approach fuses *CNN global features*, which capture the global context, with *localized, attention-weighted CNN features* that specialize in capturing finer-grained local details. To achieve this, we use a custom *SIFT-guided dynamic* and *adaptive attention* mechanism.

As illustrated in the example of Figure 6 (extracted from our experiments), within this attention mechanism, SIFT descriptors and keypoints act as guides, highlighting potentially informative regions within the image that are likely to hold discriminative power for distinguishing between different land cover types. The attention mechanism subsequently focuses on these highlighted areas, selectively amplifying the detailed features captured by the CNN within those specific patches. Most importantly, this approach integrates, rather than discards, the rich feature set extracted by the CNN from the entire image and fuses it with the attended CNN features.

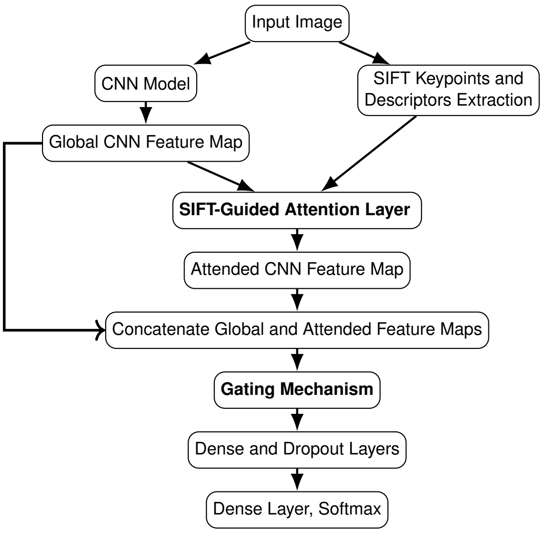

Figure 7: Mid-Level Fusion Approach: Fusion of Local Attended CNN Features and Global CNN Features (LFGF) with Gating Mechanism

As depicted in Figure 7, the feature map extracted by the RGB branch along with the keypoints and descriptors extracted by the SIFT branch form the input of the attention layer. The attention layer acts as a bridge between the global CNN features and localized SIFT information. It has three key components: *projection layers*, *scaled dot-product attention* (Vaswani et al., 2017), and *weighting and aggregation*.

Projection Layer: The first step involves projecting the CNN features, SIFT descriptors, and keypoints into a common latent space. This is achieved through the use of separate fully connected (dense) layers for each feature type. These projection layers transform the input features into vectors of the same dimensionality, enabling subsequent similarity calculations. Formally, let $\mathbf{F}_{\text{CNN}}$ denote the CNN features, $\mathbf{F}_{\text{SIFT}}$ the SIFT descriptors, and $\mathbf{K}_{\text{SIFT}}$ the SIFT keypoints. The projection layers can be represented as:

$$\mathbf{P}_{\text{CNN}} = W_{\text{CNN}}\mathbf{F}_{\text{CNN}} + b_{\text{CNN}}$$
$$\mathbf{P}_{\text{SIFT}} = W_{\text{SIFT}}\mathbf{F}_{\text{SIFT}} + b_{\text{SIFT}}$$
$$\mathbf{P}_{\text{KP}} = W_{\text{KP}}\mathbf{K}_{\text{SIFT}} + b_{\text{KP}}$$

where W and b are the weights and biases of the respective projection layers, and $\mathbf{P}_{\text{CNN}}, \mathbf{P}_{\text{SIFT}}$, and $\mathbf{P}_{\text{KP}}$ are the projected features.

Scaled Dot-Product Attention (Vaswani et al., 2017): The projected features are next fed into a scaled dot-product attention mechanism. This attention mechanism computes a score that measures the similarity between the SIFT-derived features (descriptors and keypoints) and the CNN features. These scores determine the importance of each region in the CNN feature map relative to the SIFT keypoints. The similarity scores are computed using a scaled dot-product operation: $Score(i,j) = \frac{\mathbf{P}_{\text{SIFT},i} \cdot \mathbf{P}_{CNN,j}^{T}}{\sqrt{d}}$, where d is the dimensionality of the projected features, and $\cdot$ denotes the dot product.

The attention weights are then obtained by applying a softmax function to the similarity scores: $\alpha_{ij} = softmax(Score(i,j)) = \frac{\exp(Score(i,j))}{\sum_k \exp(Score(i,k))}$. These attention weights indicate the degree of relevance of each CNN feature region to the SIFT keypoints.

Weighting and Aggregation: Finally, the attention weights are used to modulate the CNN features. The original CNN feature map is element-wise multiplied by the attention weights, effectively highlighting regions deemed important by the SIFT keypoints/descriptors. The refined features, referred to as ”*Attended CNN Features*” are computed as follows:

$$\mathbf{F}_{\text{Attended CNN Features}} = \alpha \odot \mathbf{F}_{\text{CNN}}$$

where α represents the attention weights and $\odot$ denotes the element-wise multiplication.
After obtaining the ”Attended CNN Features” these are concatenated with the original CNN features to form a comprehensive feature vector. As depicted in Figure 7 our model integrates a *gating mechanism* and *L1 regularization* to reduce redundancy in the concatenated feature map. The gating mechanism selectively combines the original and attended CNN features by learning to scale the importance of each feature through a sigmoid-activated gate, thus enhancing feature discrimination. L1 regularizations are applied to the dense layers projecting the SIFT descriptors and CNN features, as well as the gating layer, to promote sparsity in the learned weights. This encourages the model to utilize a smaller, more informative subset of features, improving generalization and reducing the risk of overfitting.

4 EXPERIMENTAL STUDY

4.1 Experimental Setup

This section evaluates the performance of the proposed fusion strategies on the EuroSAT real-world dataset. The experiments include baseline models, the early fusion model, and the proposed late and mid-level fusion models. Fusion approaches are implemented using both, the shallow CNN described earlier, and a pre-trained, fine-tuned MobileNetV2 model (Qamar and Bawany, 2023). While not the most accurate pre-trained model, MobileNetV2 offers a good trade-off between accuracy and speed. To avoid functional redundancy, we opted for a spatial attention mechanism instead of channel-wise attention in our late fusion approach using MobileNet. This choice allows the network to focus not only on the significance of features across channels (a task already managed by the depthwise separable convolutions) but also on their spatial importance.

Table 1: Accuracy achieved by the studied models.

Model	Accuracy
Baseline	
SIFT-NN model	0.619
Shallow CNN	0.845
Fine-tuned MobileNetV2	0.966
Early Fusion	
Shallow CNN	0.887
Fine-tuned MobileNetV2	0.976
Proposed Late Fusion Approach (ADL)	
Shallow CNN	0.911
Fine-tuned MobileNetV2 - SIFT	0.984
Proposed Mid-Level Fusion Approach (LFGF)	
Shallow CNN	0.924
Fine-tuned MobileNetV2	0.985

All models were implemented using Keras and TensorFlow (Abadi et al., 2015). SIFT keypoints and descriptors were extracted using OpenCV (Culjak et al., 2012). To enhance the robustness of our models, we applied common image augmentation techniques, including random flips, random jitters, random rotations, random crop, noise injections for SIFT descriptors, etc.. The EuroSAT images were stratified by land cover class and split into a 70/15/15 training, validation, and test set. Each model was trained for 100 epochs with early stopping and learning rate reduction on plateau strategies.

4.2 Results & Discussion

The results presented in Table 1 confirm the potential benefits of synergizing handcrafted features with learned CNN features for LULC classification. Notably, all the fusion approaches outperform the SIFT-NN and CNN baseline models. The results also show that not all fusion approaches are equally effective. The late fusion approach achieves an improvement of 47.17% over the baseline SIFT-NN and of 7.68% over the baseline CNN, demonstrating the advantage of applying attention mechanisms to dynamically weigh and integrate salient features from both the CNN and SIFT branches before final classification decisions are made. The mid-level fusion approach, which fuses the original CNN-learned feature map with the SIFT-based attended CNN feature map, achieves a 49.27% improvement over SIFT-NN and a

Table 2: Accuracy Achieved by Main Existing Approaches.

Model	Accuracy
SVM (Thakur and Panse, 2022)	0.509
Random Forest (Thakur and Panse, 2022)	0.567
SIFT-SVM (Helber et al., 2018)	0.701
SIFT-CNN (Ahmed et al., 2024)	0.916
Pretrained VGG19 with TWSVM (Dewangkoro and Arymurthy, 2021)	0.946
SHapley Additive exPlanations (SHAPs) (Temenos et al., 2023)	0.947
Pretrained GoogleNet (Helber et al., 2019)	0.960
Pretrained ResNet50 (Helber et al., 2019)	0.964
ResNet50 pretrained on in-domain datasets (Neumann et al., 2020)	0.992
µ2Net (Gesmundo and Dean, 2022)	0.992
Multi-Task Pretraining with Vision Transformers (ViT) (Wang et al., 2024)	0.992
Our Mid-Level Fusion Approach - Shallow CNN	0.924
Our Mid-Level Fusion Approach - MobileNetV2	0.985

9.22% improvement over the baseline CNN. The late and mid-level fusion approaches outperform the more common early fusion approach, which merges information at the initial stages of processing and might discard some feature details before the network can learn their importance.

The improvements achieved by late and mid-level fusion observed with the pre-trained MobileNetV2 are smaller than with the shallow CNN. This is likely because MobileNetV2's pre-trained features already achieve a high baseline performance, leaving less room for enhancement by additional features. However, gains observed across both models demonstrate the generalizability of the proposed fusion approaches.

As shown in Table 2, existing work heavily relies on fine-tuning pre-trained large models such as ResNet50 (Helber et al., 2019; He et al., 2016), Googlenet (*i.e.* InceptionV1,) (Gesmundo and Dean, 2022), and Vision Transformers (ViT) (Wang et al., 2024). These models deliver very high performance, with accuracies ranging from 96.0% to 99.2%. Our proposed mid-level fusion approach with a shallow CNN achieved an accuracy of 92.4%, which surpasses many traditional approaches and is competitive with some pre-trained deep learning approaches. Furthermore, our mid-level fusion with MobileNetV2 reached an accuracy of 98.5%, which is competitive with high-performance models, and only slightly below the top-performing models at 99.2% (Neumann et al., 2020; Wang et al., 2024). Table 2 highlights that, beyond the prevalent fine-tuning of pre-trained large models, alternative approaches such as the integration of handcrafted features deserve exploration.

As mentioned earlier, our mid-level approach fuses two feature maps: the original CNN-learned feature map and a SIFT-based attended CNN feature map. The original feature map captures the broader context, while the attended feature map prioritizes local details. Figure 8 shows the distribution of attention weights in the attended feature map within the mid-level approach. The peaks around zero suggest that the model still relies on the global context provided by the original CNN feature map. The presence of non-zero peaks in attention weights across classes indicates that the mid-level fusion approach effectively utilizes local features captured by SIFT descriptors. This is crucial for enhancing the model's ability to capture fine-grained details that CNNs may not prioritize.

The distribution patterns also reflect the nature of each LULC class, with more complex classes like Industrial showing a broader spread of attention weights. This indicates a more nuanced use of local features. Homogeneous classes like Forest and SeaLake show a narrow distribution, suggesting a consistent pattern of local feature importance, aligning with their more uniform textures.

The analysis of attention weight distributions highlights the potential of the mid-level fusion approach for integrating local details captured by SIFT descriptors with the global context learned by CNNs. This approach demonstrably enhances the model's ability to capture fine-grained information crucial for accurate LULC classification. However, further investigation is necessary to determine the generalizability of these findings across various datasets and LULC tasks. Additionally, exploring alternative attention mechanisms or feature extraction techniques might be beneficial for capturing even more nuanced local features or handling situations where SIFT descriptors might not be optimal.

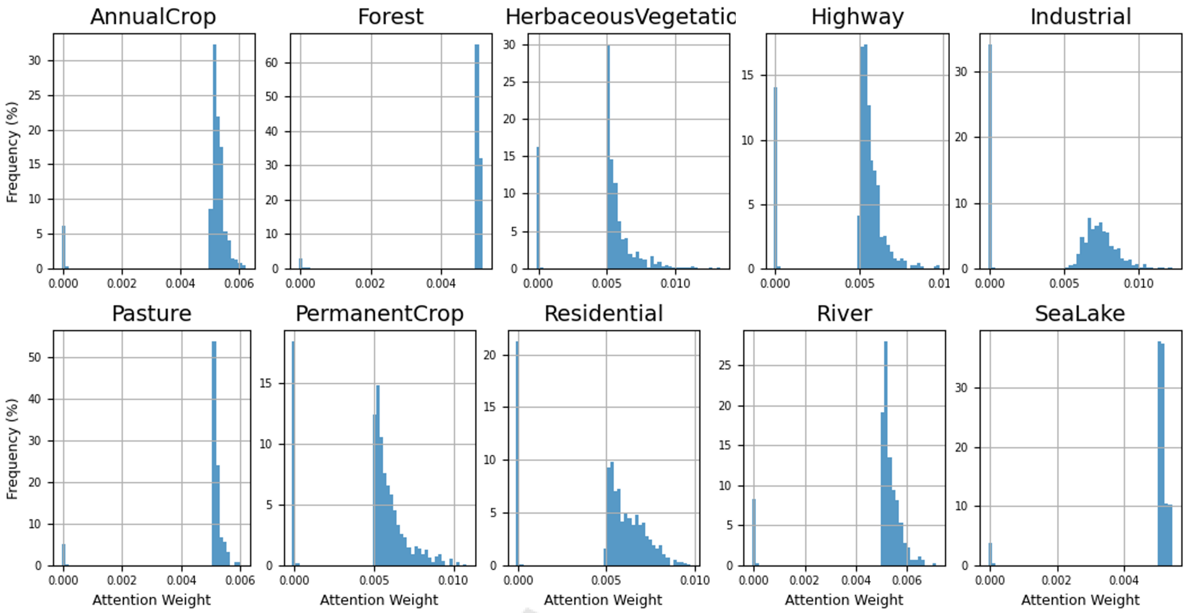

Figure 8: Distribution of Attention Weights (LFGF-CNN).

5 CONCLUSION AND FUTURE WORK

This paper investigated the synergistic integration of handcrafted SIFT descriptors with CNN-learned features for improved LULC classification accuracy. We compared three fusion strategies: early, late, and mid-level fusion. Late fusion dynamically weighs salient features from both modalities before classification. Mid-level fusion further refines this by using a custom SIFT-guided attention mechanism, selectively amplifying detailed features while preserving the rich CNN features. Experiments on real-world data showed that late and mid-level fusion outperform the conventional early fusion approach, demonstrating their efficacy in capturing both fine-grained local details and broader scene context.

The encouraging results of fusion approaches pave the way for several research directions. Moving forward, we plan to delve deeper into the realm of attention-based methods and dynamic fusion approaches. We also envision exploring the application of the proposed method to land-use change detection by analyzing time series data. Another interesting direction involves investigating the generalizability and adaptability of our proposed fusion approaches by applying them to various data-intensive tasks beyond LULC classification.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citron, C., Corrado, G., Davis, A., Dean, J., Devin, M., and et. al. (2015). Tensorflow: Large-scale machine learning on heterogeneous systems.

Ahmed, V. A., Jouini, K., Tuama, A., and Korbaa, O. (2024). A fusion approach for enhanced remote sensing image classification. In Radeva, P., Furnari, A., Bouatouch, K., and de Sousa, A. A., editors, *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2024, Volume 2: VISAPP, Rome, Italy, February 27-29, 2024*, pages 554–561. SCITEPRESS.

Chen, S. and Tian, Y. (2015). Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.*, 53(4):1947–1957.

Cheng, G., Xie, X., Han, J., Guo, L., and Xia, G. S. (2019). Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13(X):3735–3756.

Culjak, I., Abram, D., Pribanic, T., Dzapo, H., and Cifrek, M. (2012). A brief introduction to opencv. In *35th International Convention MIPRO*, page 1725–1730.

Dewangkoro, H. I. and Arymurthy, A. M. (2021). Land use and land cover classification using cnn, svm, and channel squeeze & spatial excitation block. *IOP Conf. Ser. Earth Environ. Sci.*, 704(1).

Gesmundo, A. and Dean, J. (2022). An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of* *the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2018). Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS*, page 204–207.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 12(7):2217–2226.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

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

Neumann, M., Pinto, A. S., Zhai, X., and Houlsby, N. (2020). In-domain representation learning for remote sensing. *CoRR*, abs/1911.06721.

Qamar, T. and Bawany, N. Z. (2023). Understanding the black-box: towards interpretable and reliable deep learning models. *PeerJ Comput. Sci.*, 9.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2021). How to train your vit? data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270.

Temenos, A., Temenos, N., Kaselimi, M., Doulamis, A., and Doulamis, N. (2023). Interpretable deep learning framework for land use and land cover classification in remote sensing using shap. *IEEE Geosci. Remote Sens. Lett.*, 20:1–5.

Thakur, R. and 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 da. *Orig. Res. Pap. Int. J. Intell. Syst. Appl. Eng. IJISAE*, 2022(1s):67–77. [Online]. Available: https://github.com/phelber/EuroSAT.

Tianyu, Z., Zhenjiang, M., and Jianhu, Z. (2018). Combining cnn with hand-crafted features for image classification. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 554–557.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wang, D., Zhang, J., Xu, M., Liu, L., Wang, D., Gao, E., Han, C., Guo, H., Du, B., Tao, D., and Zhang, L. (2024). Mtp: Advancing remote sensing foundation model via multi-task pretraining.

Xia, H. and Liu, C. (2019). Remote sensing image deblurring algorithm based on wgan. In Liu, X., Mrissa, M., Zhang, L., Benslimane, D., Ghose, A., Wang, Z., Bucchiarone, A., Zhang, W., Zou, Y., and Yu, Q., editors, *Service-Oriented Computing – ICSOC 2018 Workshops*, pages 113–125, Cham. Springer International Publishing.