TexMobile: A Texture-Aware Lightweight Framework for Campus Scene Classification

Jana Godieh Eithar Al-Salameh Leen Almasarweh Dana Abu Al-Ruz
Artificial Intelligence Department, University of Jordan

Instructor: Tamam AlSarhan

Abstract

Fine-grained visual recognition remains challenging in complex real-world environments where scene context and object-level semantics are highly entangled. Existing approaches often rely on single-level supervision and global feature pooling, which limits their ability to capture distributional structures and hierarchical relationships between scene and object categories. In this paper, we propose a hierarchical classification framework that integrates coarse-level semantic supervision with fine-grained encoding-based representation learning. A texture encoding layer is employed to model the distribution of local patterns, enabling the fine-grained head to capture subtle intra-class variations. Extensive experiments conducted on a custom dataset collected from real-world environments at the University of Jordan, comprising both scene categories (laboratory and building environments) and object categories (person, car, and tree), demonstrate the effectiveness of the proposed approach. The proposed model achieves stable convergence behavior, high coarse-level accuracy, and strong fine-grained classification performance.

Figure 1. Representative challenges in campus scene classification. The top row illustrates substantial intra-class variation among laboratory environments, while the bottom row highlights architectural homogeneity, inter-class similarity, and foreground occlusions in outdoor campus scenes.

1. Introduction

The unprecedented global expansion of higher education has forced universities to evolve from collections of buildings into complex, multi-functional ecosystems, necessitating equally sophisticated management systems. This expansion is quantified by an increase in the worldwide gross enrollment rate in higher education from 19% to 40% between 2000 and 2020, with more than 235 million students enrolled worldwide in 2020— more than double the 100 million enrolled at the turn of the century [1]. Operating on this scale introduces unprecedented complexity—larger footprints, greater functional diversity, and higher security demands—that overwhelm traditional manual supervision. Visual scene understanding—the ability to automatically recognize and reason about campus environments—has thus become an essential infrastructure.

However, campus scene recognition introduces unique challenges beyond general computer vision tasks. These environments encompass both structured indoor spaces, such as classrooms and laboratories, and complex outdoor scenes with varying architectural elements, lighting conditions, and seasonal changes. Although humans demonstrate remarkable efficiency in categorizing natural scenes [2, 3], this remains a formidable challenge for automated systems due to semantic ambiguity and large intra-class variations caused by illumination changes, viewing angles, scale differences, and seasonal variations. Furthermore, campus scenes exhibit domain-specific characteristics—including architectural homogeneity across multiple buildings, functional diversity within similar structures, and inherent hierarchical organization distinguishing object-centric views

from scene-oriented landscapes—that are not adequately addressed by general scene recognition approaches. Figure 1 illustrates representative examples from our campus dataset that visually demonstrate these challenges like visual variability and semantic ambiguities that complicate automated scene classification.

While modern CNN-based approaches have significantly advanced scene classification, most operate over flat label spaces and rely on global pooling for final prediction, limiting their ability to explicitly model hierarchical semantics or preserve fine-grained discriminative cues [4, 5]. Hierarchical classification strategies have been proposed to address these limitations [6, 7], but often incur additional computational complexity or depend on cascaded architectures. Moreover, prior work has shown that textural information provides complementary cues for scene understanding, particularly in architecturally similar environments [8].

Campus scenes naturally exhibit a coarse-to-fine semantic organization, with a high-level distinction between object-centric views and scene-oriented landscapes, followed by fine-grained categorization within each group. In addition, subtle variations in materials, surfaces, and vegetation frequently play a decisive role in distinguishing visually similar campus environments.

To address these characteristics, this paper proposes a hierarchical CNN architecture that explicitly models coarse-to-fine semantic structure through dual classification heads operating on shared feature representations. Given an input image, a CNN backbone extracts a shared feature map that is simultaneously fed into (i) a coarse classification head, which performs global average pooling followed by binary classification of object-centric versus scene-centric views, and (ii) a fine classification head, which incorporates a Deep Texture Encoding Network (DeepTEN) to capture discriminative textural patterns for fine-grained scene recognition. The model is trained end-to-end using a weighted multi-task objective that jointly optimizes coarse and fine classification losses.

The main contributions of this work are as follows:

- A hierarchical CNN architecture for campus scene classification that jointly models coarse object–scene discrimination and fine-grained semantic categorization using shared backbone features.
- Integration of a Deep Texture Encoding Network within the fine classification head to enhance sensitivity to texture cues critical for distinguishing visually similar campus environments.
- A joint training strategy with a weighted multi-task loss that balances coarse and fine objectives while maintaining computational efficiency.

2. Related Work

2.1. Human Scene Understanding

Human visual perception of scenes has been extensively studied in psychology and neuroscience. Greene and Oliva demonstrated that humans can rapidly extract the gist of a scene—its global semantic properties—from very brief visual exposures, often preceding detailed object-level recognition [2]. This efficiency highlights the hierarchical nature of human scene understanding, where coarse structural information is accessed early and refined by additional visual cues. Consistent with this view, Walther *et al*. showed that even sparse contour-based representations preserve sufficient structural information to support scene category representations in the human visual cortex, underscoring the importance of layout and edge information in early scene processing [3].

These studies highlight the remarkable efficiency of human scene understanding, a capability that automated systems strive to replicate.

2.2. General Scene Recognition

Scene classification—the task of assigning semantic labels to images depicting environments—has evolved from hand-crafted representations to deep learning. Early work emphasized global scene structure using holistic descriptors such as GIST, which captured the spatial envelope of scenes via oriented energy across multiple scales [9]. In parallel, local invariant features such as SIFT enabled robust matching under scale and viewpoint changes [10], and mid-level representations emerged through Bag-of-Visual-Words (BoVW) models that quantize local descriptors into visual vocabularies [11]. Lazebnik *et al*. advanced this paradigm with Spatial Pyramid Matching (SPM), introducing hierarchical spatial partitioning to preserve coarse-to-fine layout information while remaining computationally efficient [12]. Further improvements to aggregation and encoding were achieved using VLAD and Fisher Vector representations, which capture higher-order statistics of local descriptors and became strong pre-deep-learning baselines for recognition tasks [13–15].

Despite their success, hand-crafted pipelines required careful feature engineering and struggled with semantic ambiguity in complex real-world scenes. The emergence of deep learning marked a major shift: convolutional neural networks (CNNs) enabled end-to-end learning of hierarchical representations and significantly improved robustness to variations in illumination, viewpoint, and scale [4]. Subsequent architectural advances, including very deep networks and residual learning, further strengthened feature transferability and recognition performance [5, 16, 17]. While CNNs reduced the reliance on manual feature design, standard scene classifiers often still employ global pooling with

flat label spaces, which can limit their ability to explicitly model hierarchical semantics or retain fine-grained visual cues needed for distinguishing visually similar environments.

2.3. Hierarchical and Coarse-to-Fine Classification

Hierarchical approaches address limitations of flat classifiers by organizing categories into taxonomic structures that reflect semantic relationships. Yan *et al*. [6] proposed HD-CNN, which embeds category hierarchies directly into network architecture through cascaded classifiers operating at progressively finer semantic granularities. This coarse-to-fine strategy reduces the effective search space at each decision stage and mirrors human categorization processes.

Building on this idea, Taoufiq *et al*. introduced HierarchyNet, which jointly learns feature extraction and hierarchical classification through a multiplicative mechanism that conditions fine-level predictions on coarse outputs [7]. Alternative formulations include tree-structured networks employing hierarchical softmax [18] and semantic hierarchy methods that leverage external knowledge bases such as WordNet to define category relationships [19].

2.4. Texture and Material Recognition

Texture plays a crucial role in scene understanding, particularly in distinguishing between similar architectural environments. Early texture analysis methods included filter banks [20] and textons. More recently, deep learning approaches like Deep TEN (Texture Encoding Network) [8] introduced an end-to-end trainable texture encoding layer that captures texture-specific features, proving effective for material recognition in complex scenes. For campus environments, texture analysis helps distinguish between different building materials, indoor surfaces, and natural elements across seasons.

2.5. Domain-Specific Challenges in Campus Scenes

Campus environments present unique challenges not adequately addressed by general scene recognition approaches. As noted by Li *et al*. [21], educational institutions combine structured indoor spaces (classrooms, laboratories, libraries) with complex outdoor areas featuring varying architectural styles, lighting conditions, and seasonal changes. The hierarchical organization of campus scenes-from object-centric views (desks, whiteboards) to scene-oriented landscapes (quadrangles, building facades)-requires specialized approaches.

Recent work has begun addressing these domain-specific challenges. Chen *et al*. [22] proposed a multi-scale feature fusion network for campus scene recognition that handles both indoor and outdoor views. Their approach incorporates attention mechanisms to focus on discriminative architectural elements while suppressing seasonal variations. Similarly, Wang and Zhang [23] introduced a cross-season adaptation framework that learns invariant features across different weather conditions and times of day, crucial for robust campus monitoring systems.

2.6. Current Limitations and Research Gaps

Despite these advances, several gaps remain in campus scene understanding research. First, most existing datasets (MIT Places, SUN397) contain limited campus-specific categories and lack the fine-grained distinctions needed for educational facility management. Second, current methods often fail to capture the functional semantics of spaces-recognizing that two visually similar rooms may serve different purposes (lecture hall vs. auditorium). Third, there is insufficient attention to the security and accessibility aspects of scene understanding, such as identifying blocked pathways or unauthorized access points.

Recent work by Gupta *et al*. [24] addresses some of these limitations through hierarchical graph-based representations that encode spatial relationships between campus elements. However, their approach requires extensive manual annotation and struggles with real-time processing-a critical requirement for campus management systems.

3. Hierarchical Texture-Aware Campus Scene Classification

3.1. Problem Definition

In our dataset, images are classified into five semantic categories: *Person*, *Tree*, *Building*, *car*, and *lab*. While CNNs have several advantages, they also have some limitations. First, a flat CNN treats all target classes as mutually independent and equally distant in the label space. However, the classes in our dataset naturally share semantic and visual relationships. For instance, *Building* and *Labs* often exhibit similar structural patterns such as rectangular shapes, windows, and rigid edges, while *Person* and *Cars* may co-occur in urban scenes. A flat classifier ignores these relationships and attempts to separate all classes simultaneously, increasing confusion between visually similar categories.

Second, training a flat CNN forces a single classifier to handle both coarse semantic discrimination (e.g., natural vs. artificial objects) and fine-grained recognition (e.g., distinguishing *Cars* from *Labs*). This increases optimization difficulty, often leading to unstable convergence and reduced generalization, particularly when some classes share dominant visual cues.

Finally, in a flat classification setting texture features extracted by convolutional layers compete globally, forcing the model to use the same texture cues to separate visually dissimilar classes. This often results in misclassification when dominant texture patterns are shared, for example between *Building* and *Labs*, or when shape information

is insufficient to resolve ambiguity.

3.2. Overall Architecture

To address the limitations of flat classification, we propose a hierarchical texture-aware architecture that explicitly models the two-tier semantic structure inherent in campus environments. Figure 2 illustrates our complete framework.

Our approach consists of three main components: (1) a shared MobileNetV2 backbone for feature extraction, (2) a coarse classification head using global average pooling for high-level semantic discrimination, and (3) a fine classification head incorporating Deep Texture Encoding Network (DeepTEN) for material-level feature aggregation.

Architectural Design Rationale. The proposed dual-head hierarchical architecture is designed to address key limitations of flat CNN classifiers when applied to visually complex campus environments. In particular, flat models are required to simultaneously separate semantically distant classes while resolving subtle intra-class variations, which increases optimization difficulty and often leads to confusion between visually similar categories. To alleviate this, our framework decomposes the learning task into complementary coarse- and fine-grained objectives.The coarse classification branch focuses on learning broad semantic groupings (e.g., object versus scene), which constitutes a comparatively simpler discrimination task. This auxiliary supervision encourages the shared backbone to capture globally discriminative representations that reflect high-level semantic structure. In parallel, the fine classification branch specializes in resolving subtle material and texture differences within these semantic groups, which are critical for distinguishing architecturally homogeneous environments where spatial layout alone is insufficient. The fine branch incorporates a texture encoding pathway that aggregates orderless material-level features, such as building surfaces, pavement patterns, and vegetation textures. This enables robust fine-grained recognition by emphasizing discriminative texture statistics rather than relying solely on spatial configurations. Both branches are jointly optimized using a weighted multi-task loss, allowing the backbone to learn representations that support both global semantic discrimination and fine-grained texture refinement.

Importantly, the hierarchical structure is enforced implicitly through joint supervision rather than explicit feature modulation or gating mechanisms. This formulation acts as a regularizer during training, stabilizing optimization and improving generalization. At inference time, the model produces interpretable two-tier predictions, consisting of a coarse semantic grouping followed by fine-grained classification, achieving strong performance at both levels while maintaining architectural simplicity and computational efficiency.

3.3. Backbone Feature Extraction

The backbone of the proposed framework is implemented using the feature extraction layers of a pretrained MobileNetV2 network [25] , chosen for its favorable trade-off between computational efficiency and representational capacity. Specifically, only the convolutional feature layers are retained, while the original classification head is discarded

Given an input image x, the backbone $\mathcal{B}(\cdot)$ produces a dense feature map representation:

$$\mathbf{F} = \mathcal{B}(x), \quad \mathbf{F} \in \mathbb{R}^{C \times H \times W}, \tag{1}$$

where $C = 1280$ corresponds to the output channel dimension of MobileNetV2 [25], and H and W depend on the input resolution. These feature maps encode multi-level semantic information, ranging from low-level edge and texture primitives to high-level object- and scene-related patterns. The backbone parameters are fine-tuned during training using a lower learning rate than the task-specific heads, enabling stable adaptation to the target dataset while preserving the generalization capability inherited from large-scale pretraining.

3.4. Coarse Classification Head

To enhance the robustness of fine-grained recognition, we introduce a *Coarse Classification Head* as an auxiliary branch that guides the backbone network to first learn high-level semantic structures before focusing on detailed class distinctions. This design follows the principles of hierarchical learning and multi-task supervision, enabling the model to better generalize in challenging imaging scenarios.

Given an input image x, the backbone network $\mathcal{B}(\cdot)$, produces feature maps $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$. These features are globally aggregated using adaptive average pooling:

$$\mathbf{f}_c = \text{GAP}(\mathbf{F}), \tag{2}$$

By operating on globally pooled feature maps, the coarse branch captures the overall distribution of visual patterns across the image, providing a holistic description of both the scene context (laboratory and buildings) and the object categories (persons, cars, and trees). This encourages the network to model high-level semantic cues such as global layout, background structure, and object co-occurrence statistics, which are essential for distinguishing broad environmental conditions before performing fine-grained recognition.

This hierarchical supervision strategy is conceptually inspired by Taoufiq *et al.*'s HierarchyNet [7], which emphasizes encoding distributional statistics to capture global structural patterns, and is further supported by recent studies in hierarchical remote sensing image classification [26]. By incorporating coarse-level supervision, the backbone is

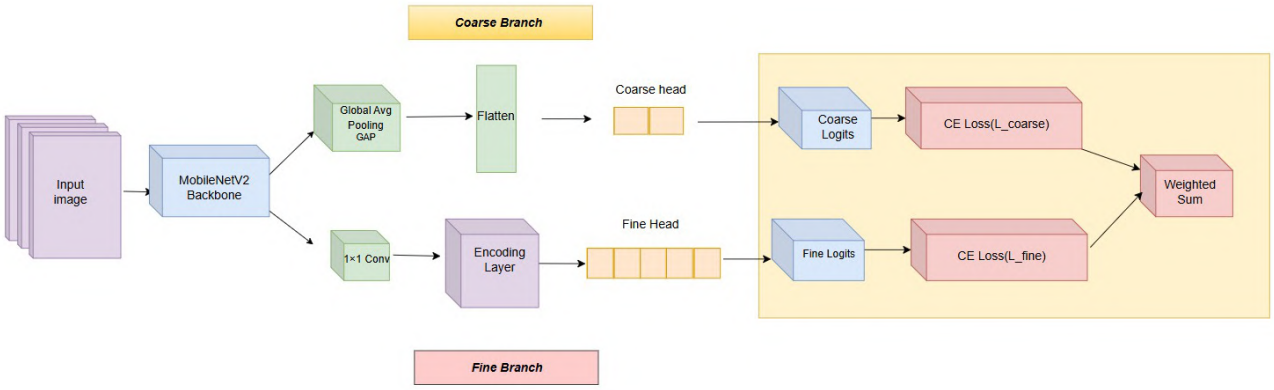

Figure 2. Overview of the proposed hierarchical texture-aware CNN. A shared MobileNetV2 backbone extracts convolutional features that are processed by two parallel branches. The coarse branch performs global semantic discrimination using global average pooling and a coarse classification head, while the fine branch employs a texture encoding layer to capture discriminative material cues for fine-grained recognition. During training, the model is optimized using a weighted joint loss combining fine- and coarse-level cross-entropy objectives.

encouraged to learn discriminative global representations, which act as a strong regularizer and facilitate more stable optimization for the fine-grained classification task.

3.5. Fine Classification with Texture Encoding

To explicitly enforce hierarchical consistency, the coarse logits are used to modulate the fine-level logits through a multiplicative interaction. A shared backbone network extracts a feature vector $\mathbf{z} = f(x; \theta_b)$ from the input image x. The fine classification head maps $\mathbf{z}$ to logits over N_f fine classes:

$$\mathbf{v}^{(f)} = W_f \mathbf{z} + \mathbf{b}_f. \tag{3}$$

The resulting logits are partitioned into N_c disjoint subvectors $\{\mathbf{v}_i^{(f)}\}_{i=1}^{N_c}$, each corresponding to the subclasses of a given coarse class as defined by the hierarchical mapping H.

The fine-level class probabilities are obtained using a softmax function:

$$p_j^{(f)} = \frac{\exp\left(\tilde{v}_j^{(f)}\right)}{\sum_{k=1}^{N_f} \exp\left(\tilde{v}_k^{(f)}\right)}, \quad j = 1, \ldots, N_f. \tag{4}$$

The fine classification probability is obtained by applying a softmax function to the hierarchically modulated fine logits, yielding a valid probability distribution over fine classes that is explicitly conditioned on the coarse prediction.

3.6. Joint Training Objective

The coarse branch is supervised using a cross-entropy loss $\mathcal{L}_{coarse}$ computed over the mapped superclass labels y_c,which are derived from the original fine-grained ground truth categories. This supervision strategy enforces the backbone to learn global scene-level representations rather than focusing solely on subtle local variations. The fine-grained branch is similarly supervised using a cross-entropy loss $\mathcal{L}_{fine}$ calculated on the original fine-grained labels y_f. This helps the network capture subtle discriminative features between closely related categories, enabling accurate fine-level classification. The overall training objective combines the fine-grained and coarse losses as:

$$\mathcal{L} = \mathcal{L}_{fine} + \lambda \mathcal{L}_{coarse}, \tag{5}$$

where λ is a dynamically scheduled weighting factor that gradually decreases during training to shift the model's focus from global semantic discrimination to fine-grained categorization. This formulation allows for dynamic adjustment of the loss weights during training to stabilize optimization and improve fine-level classification performance.

4. Dataset and Experimental Setup

4.1. Campus Scene Dataset

To evaluate our approach, we introduced the Campus Scene Dataset, comprising 2,405 manually collected images from the University of Jordan campus. The dataset features five fine-grained categories (Building, Car, Lab, Person, Tree) organized into two coarse super-classes (Objects and Scenes) to facilitate hierarchical learning. For experimentation, images were partitioned into 70% training, 15% validation, and 15% testing sets as detailed in Table 1.

Table 1. Statistics of the dataset collected from the **University of Jordan**, showing image distribution across fine-grained classes.

Fine Class	Coarse Class	Train	Val	Test	Total
Building	Scene	266	57	57	380
Lab	Scene	313	67	68	448
Car	Object	389	83	84	556
Person	Object	308	66	67	441
Tree	Object	406	87	87	580
Total	**-**	**1682**	**360**	**363**	**2405**

4.2. Data Preprocessing and Augmentation

To prepare the dataset ,we applied a series of preprocessing and augmentation techniques to ensure model robustness and generalization. All input images were normalized using the standard ImageNet mean ($\mu = [0.485, 0.456, 0.406]$) and standard deviation ($\sigma = [0.229, 0.224, 0.225]$) to facilitate stable convergence during training. During the training phase, we employed data augmentation strategies to mitigate overfitting and improve the model's ability to generalize to unseen data. Specifically, we used Random Resized Crop, which crops a random portion of the image and resizes it to 224×224 pixels, introducing scale invariance. Additionally, Random Horizontal Flip was applied with a probability of 0.5 to introduce rotational invariance. For the validation and testing phases, we adopted a deterministic preprocessing pipeline to ensure consistent evaluation. Images were first resized to 256 pixels along the shorter edge, followed by a Center Crop to extract a 224×224 patch, ensuring that the central subject of the scene is preserved for classification.

Figure 3. Data preprocessing visualization. **(a)** A raw sample from the Building class. **(b)** The same image after resizing, random cropping to 224×224, and normalization (displayed in standard colors for clarity).

4.3. Implementation Details

Our experimental setup was executed using the PyTorch framework [27]. The network was trained for 30 epochs with a batch size of 32. For optimization, we employed the AdamW optimizer [28] with a weight decay of 5×10^{-4} to ensure robust regularization.

A key component of our training protocol was the adoption of a differential learning rate strategy. We assigned a conservative learning rate of 1×10^{-5} to the pre-trained MobileNetV2 backbone [25] to preserve its feature extraction capabilities, while a higher rate of 1×10^{-4} was applied to the randomly initialized hierarchical heads to facilitate rapid convergence. These rates were dynamically modulated using a Cosine Annealing scheduler [29]. Additionally, we implemented an adaptive loss balancing mechanism where the coarse supervision weight, λ, was initialized at 0.3 and decayed following a cosine schedule. This allowed the model to prioritize structural learning in the early stages before refining its focus on fine-grained texture details as training progressed.

5. Experiments

Training Initialization. Given the relatively small scale of the Campus Scene dataset (2,405 images across five categories), all models employ transfer learning from ImageNet-pretrained weights. The MobileNetV2 backbone is initialized with pretrained parameters, providing robust low- and mid-level feature representations (edges, textures, and basic shapes) that generalize well to campus environments.

Preliminary experiments with randomly initialized backbones exhibited unstable optimization behavior, slow convergence, and high variance across runs, often leading to early overfitting and degraded generalization performance. These effects are consistent with prior observations that deep convolutional models require large-scale data to learn robust low-level and mid-level visual representations. [30, 31].

By leveraging pretrained weights, the backbone benefits from transferable feature representations learned on large-scale natural image corpora, enabling stable training and effective adaptation to the target domain. Unless otherwise stated, only the backbone is initialized from ImageNet, while all task-specific heads are trained from scratch.

To further investigate the role of feature refinement in texture-based classification, we evaluate the impact of integrating a channel-wise attention mechanism into the backbone network. Specifically, we employ a Squeeze-and-Excitation (SE) block, which adaptively reweights feature channels based on their global importance as proposed by Hu *et al*. in Squeeze-and-Excitation Networks [32]. The SE block is inserted after the final convolutional block of the backbone and operates solely on intermediate feature representations. The hierarchical classification framework, including the coarse-to-fine structure and loss functions, remains unchanged to ensure a fair comparison.

In addition to channel-wise attention, we evaluate the im-

pact of spatial attention on classification performance. The spatial attention mechanism follows the formulation introduced in CBAM, where a spatial attention map is learned to highlight informative regions in the feature maps [33]. Spatial attention aims to enhance feature representations by emphasizing spatially informative regions while suppressing background responses. We implement a lightweight spatial attention module consisting of a 1×1 convolution followed by a sigmoid activation, producing a single attention map that is applied to the feature maps via element-wise multiplication. The module is inserted after the final convolutional block of the backbone network, and all other components of the hierarchical classification framework remain unchanged.

While SE and spatial attention enhances training accuracy, its impact on validation performance is less consistent. This behavior suggests that emphasizing spatial details may lead to overfitting on localized texture patterns, particularly when training data is limited.

6. Results

6.1. Quantitative Results

This section reports the quantitative evaluation of the proposed hierarchical framework on the test dataset containing five fine-grained classes: *Building, Car, Lab, Person, and Tree*.

Figure 4 illustrates the training and validation loss and accuracy curves over 30 epochs. The results demonstrate stable convergence behavior, where the training and validation losses decrease smoothly without significant divergence, indicating effective regularization and absence of overfitting. The validation accuracy closely follows the training accuracy throughout the training process, achieving approximately 97.5% at the final epochs.

Figure 5 presents the normalized confusion matrix for fine-grained classification on the test set. The diagonal dominance indicates strong discriminative performance across all classes. Notably, the Building and Lab categories achieve perfect recognition rates, while minor confusion is observed between *Tree* and *Building*, where approximately 5% of Tree samples are misclassified as Building. This behavior is expected due to the frequent co-occurrence of trees within building environments, which introduces background context ambiguity.

To evaluate the effectiveness of the coarse supervision, we report the object-versus-scene classification accuracy in Figure 6. The coarse branch rapidly converges within the first few epochs, reaching a validation accuracy above 95% and stabilizing around 97%. This confirms that the model successfully learns to discriminate between scene-level categories (Building, Lab) and object-level categories (Person, Car, Tree), providing reliable high-level semantic guidance to the backbone network.

Overall, the proposed hierarchical framework achieves robust fine-grained classification performance while maintaining high coarse-level accuracy, validating the effectiveness of combining global scene-object supervision with fine-grained encoding-based representation learning.

6.2. Comparison with Baseline Models

To demonstrate the effectiveness of our approach, we compared it against standard baseline architectures including MobileNetV2, DenseNet121, and GoogleNet. Table 2 summarizes the results.

Table 2. Performance comparison with baseline models.

Model	Params (approx.)	Test Accuracy
MobileNetV2 (Vanilla)	∼2.2 M	93.42%
DenseNet121	∼8.0 M	94.18%
GoogleNet	∼6.6 M	96.69%
Ours	**∼2.4 M**	**98.07%**

As shown in Table 2, our texture-aware model achieves the highest accuracy of **98.07%**, outperforming the heavier GoogleNet (96.69%) and DenseNet121 (94.18%) while maintaining a lightweight architecture suitable for resource-constrained environments.

6.3. Ablation Study

We conduct a systematic ablation study to quantify the contribution of each component in the proposed hierarchical texture-aware architecture. All experiments employ a MobileNetV2 backbone initialized with ImageNet pretrained weights and are evaluated on identical data splits. Unless otherwise stated, models are trained using the same optimization strategy and data preprocessing pipeline, with training duration adapted to the number of trainable parameters.

We progressively introduce architectural components starting from a flat CNN baseline and evaluate their impact on fine-grained and coarse-level recognition performance.

Experimental Configurations: We evaluate four progressively enhanced variants:

- **Baseline (Flat CNN):** Standard MobileNetV2 with a single classification head for the five fine-grained categories.
- **+ Texture Encoding**: Baseline augmented with the Deep Texture Encoding Network (DeepTEN) module [8] before the classification layer.
- **+ Coarse Head**: Incorporates the hierarchical coarse classification branch for Object/Scene discrimination alongside the fine-grained head.
- **Full Model**: Complete hierarchical texture-aware architecture with all components active.

Table 3 summarizes the ablation results.

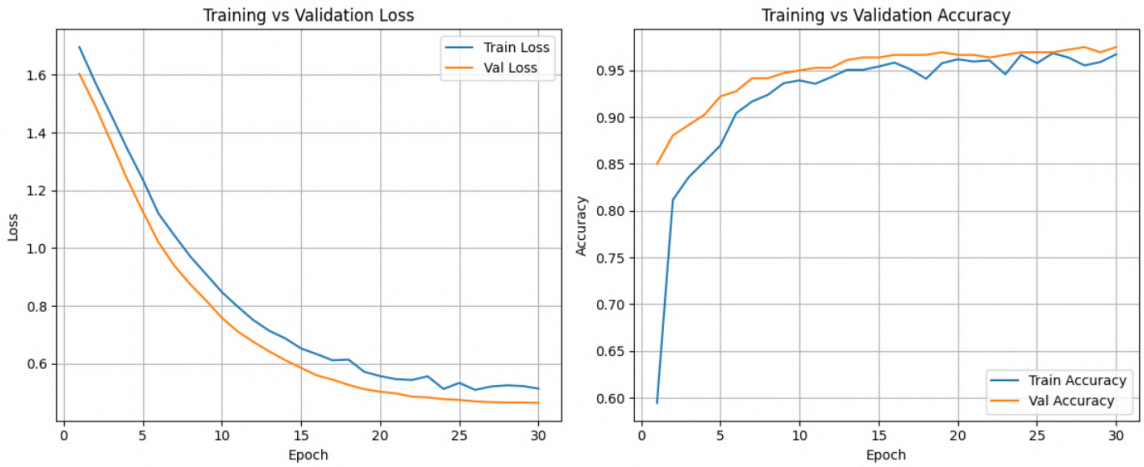

Figure 4. Training and validation loss and accuracy curves over 30 epochs.

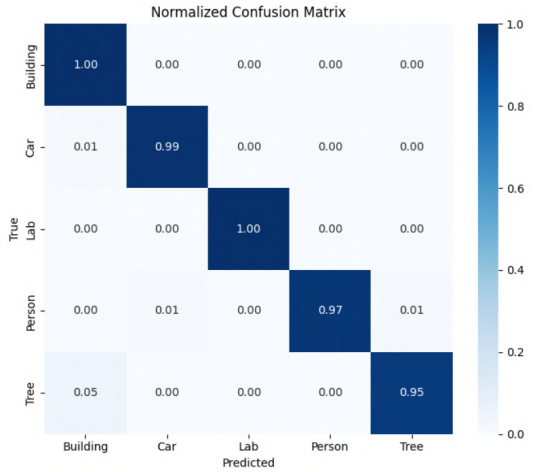

Figure 5. Normalized confusion matrix for fine-grained classification.

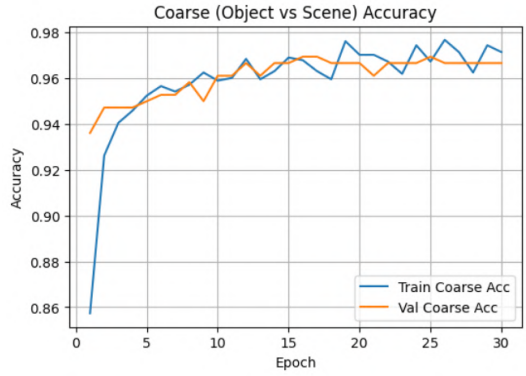

Figure 6. Coarse-level (Object vs Scene) training and validation accuracy.

Table 3. Ablation study evaluating the contribution of individual components in the proposed hierarchical texture-aware architecture. All models use a MobileNetV2 backbone pretrained on ImageNet.

Variant	**Val (%)**	**Test Fine (%)**	**Test Coarse (%)**
Baseline (flat)	94.9%	93.4%	–
+ Texture	97.8%	97.5%	–
+ Coarse head	96.7%	98.1%	97.8%
Full Model	**97.5%**	**98.1%**	**98.4%**

Analysis: The baseline flat CNN achieves strong performance (93.4% test accuracy), establishing a competitive starting point given the compact dataset size. Introducing texture encoding yields a substantial improvement of +4.1 percentage points on fine-grained test accuracy, confirming the importance of material-aware representations for distinguishing visually similar campus scenes.

Adding the hierarchical coarse classification head further improves fine-grained recognition, increasing test accuracy to 98.1% while simultaneously enabling accurate coarse-level prediction (97.8%). This auxiliary supervision encourages the backbone to learn globally discriminative representations, reducing semantic ambiguity between object-centric and scene-centric views.

The full model achieves the most balanced performance, attaining 98.1% fine-grained and 98.4% coarse-level test accuracy. While certain ablated variants achieve comparable validation performance, the complete architecture consistently delivers superior generalization across both semantic levels, demonstrating the complementary benefits of texture encoding and hierarchical supervision.

7. Conclusion

In this work, we presented a hierarchical texture-aware framework for campus scene classification, designed to address the challenge of high inter-class similarity. By integrating a MobileNetV2 backbone with a dedicated texture encoding layer, our model effectively captures both global structural features and fine-grained local details. We validated our approach on the Campus Scene Dataset manually collected from the University of Jordan. Experimental results demonstrate that our method achieves a competitive accuracy of 98.07%, confirming that combining hierarchical supervision with explicit texture modeling significantly enhances recognition performance in complex outdoor environments.

References

[1] UNESCO. Higher education global data report (working document). Technical report, United Nations Educational, Scientific and Cultural Organization, May 2022. 1

[2] Michelle R. Greene and Aude Oliva. The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472, 2009. 1, 2

[3] Dirk B. Walther, Barry Chai, Eamon Caddigan, Diane M. Beck, and Li Fei-Fei. Simple line drawings suffice for functional mri decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, 108(23):9661–9666, 2011. 1, 2

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[6] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015. 2, 3

[7] Salma Taoufiq, Balázs Nagy, and Csaba Benedek. Hierarchynet: Hierarchical cnn-based urban building classification. *Remote Sensing*, 12(22):3794, 2020. 2, 3, 4

[8] L. Zhang, Y. Yang, M. Wang, R. Hong, L. Nie, and X. Li. Deep TEN: Texture encoding network. In *CVPR*, pages 2896–2905, 2017. 2, 3, 7

[9] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 2

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

[11] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003. 2

[12] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006. 2

[13] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311, 2010. 2

[14] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[15] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 2

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[17] Christian Szegedy, Wei Liu, Yangqing Jia, et al. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 2

[18] Bolei Zhou, Vignesh Jagadeesh, and Robinson Piramuthu. Conditional probability models for deep learning of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2012. 3

[19] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 3

[20] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001. 3

[21] X. Li, Y. Zhang, H. Wang, and J. Chen. Campus scene understanding: Challenges and opportunities. *IEEE Access*, 8:123456–123467, 2020. 3

[22] J. Chen, M. Wang, and L. Liu. Multi-scale feature fusion for campus scene recognition. In *ICIP*, pages 112–116, 2021. 3

[23] Q. Wang and L. Zhang. Cross-season adaptation for robust campus scene understanding. *PR*, 125:108456, 2022. 3

[24] R. Gupta, S. Patel, and A. Kumar. Graph-based hierarchical representation for campus scene understanding. In *WACV*, pages 234–241, 2023. 3

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 4, 6

[26] C. Zhang, W. Li, and Q. Du. Hierarchical deep learning framework for remote sensing image classification. *Remote Sensing*, 12(22):3794, 2020. 4

[27] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[30] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2661–2671, 2019. 6

[31] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4918–4927, 2019. 6

[32] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2019. 6

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 7