Beyond Flat Classification: A Light Incremental Hierarchical CNN Framework for Vegetation and Urban Scene Recognition

Wasan Alhababheh, Layan Nofal, Maha Qaddoumi
Artificial Intelligence Department
University of Jordan
Amman, Jordan
Instructor: Tamam AlSarhan

***Abstract*—Hierarchical image classification remains a challenging problem due to high inter-class similarity and the growing scale of modern datasets. Conventional flat classification models struggle to scale efficiently and often require complete retraining when new classes are introduced. In this paper, we propose an incremental deep learning framework that decomposes the classification task into multiple stages using transfer learning and hierarchical decision routing. The framework employs independently trained convolutional neural network models to progressively refine predictions from coarse-grained to fine-grained categories. Additionally, we introduce a light incremental inference strategy that enables system extensibility without retraining previously learned models. While not designed for continual weight adaptation, this strategy provides a practical and efficient solution for scalable hierarchical classification. Experimental results demonstrate that the proposed approach improves classification reliability, reduces computational overhead during inference, and supports scalable deployment. The modular design makes the framework suitable for real-world applications where efficiency and adaptability are essential.**

***Index Terms*—Hierarchical Image Classification, Incremental Learning, Convolutional Neural Networks, Transfer Learning, Efficient Inference, Urban Scene Recognition.**

I. Introduction

The rapid evolution of Convolutional Neural Networks (CNNs) has established a new paradigm in automated image recognition. However, traditional ”flat” classification architectures (which attempt to distinguish all target classes within a single output layer) frequently encounter performance bottlenecks due to high inter-class similarity [2]. As noted in our research, distinguishing between visually similar urban entities like Buildings and Laboratories requires a more nuanced approach than standard multi-class models provide.

To address these challenges, this research proposes a Modular Hierarchical CNN framework. By decomposing the classification problem into a sequence of specialized binary decision nodes, our approach mimics the ”coarse-to-fine” cognitive process of human visual perception [2]. Instead of forcing a single network to learn the nuanced differences between all classes simultaneously, we partition the problem into logical stages. This structure isolates decision boundaries, significantly reducing feature confusion between visually similar categories.

Our methodology leverages Transfer Learning using the EfficientNet-B0 architecture [1]. The system is designed as a modular pipeline where high-level environmental domains are resolved before moving to fine-grained urban classification. This modularity ensures that the system is not only accurate but also computationally accessible for deployment on commodity hardware. While this paper focuses on the foundational three-tier architecture, the design is specifically intended to support incremental updates, a feature we explore in our future work [6].

II. Related work

Convolutional Neural Networks (CNNs) have become the standard for image classification, with architectures such as EfficientNet [1] achieving state-of-the-art performance through optimized compound scaling. To handle complex datasets with high inter-class similarity, hierarchical methods like HD-CNN [2] exploit semantic relationships to reduce decision complexity through a coarse-to-fine strategy.

Furthermore, ensuring model reliability is critical in urban environments; recent research has introduced methods for network calibration [3] to reduce Expected Calibration Error (ECE), while selective prediction frameworks [4] allow models to optimize the risk-coverage trade-off. While standard architectures often encounter performance bottlenecks, traditional models like ResNet [5] are frequently used as benchmarks for comparison. To address the issue of catastrophic forgetting in incremental learning, a wide range of strategies has been explored. Methods such as Learning Without Forgetting and iCaRL preserve prior knowledge by constraining the training loss when new classes are introduced, while approaches like Expert Gate and Progressive Neural Networks expand the network by adding new task-specific modules. These techniques have shown promise, but they often depend on rehearsal buffers, complex loss regularization, or continual architectural growth, which can make them difficult to deploy in practical settings.

Motivated by this limitation, our work adopts a modular hierarchical structure in which each decision node functions as an independent classifier. When a new category is introduced, only the corresponding branch is retrained while all other nodes remain frozen. This selective update strategy reduces interference between unrelated classes and helps mitigate catastrophic forgetting without requiring full network retraining.

III. Methodology

This work proposes a Hierarchical Convolutional Neural Network (HI-CNN) framework that replaces flat multiclass classification with a structured, multi-stage decision process. The methodology emphasizes architectural design, node configuration, and inference logic that enable progressive refinement from coarse semantic separation to fine-grained urban scene classification.

A. HIERARCHICAL ARCHITECTURE

The HI-CNN is organized as a directed hierarchical tree [9], [11], where each node performs a specialized classification task over a reduced label space. As illustrated in Fig. 1, the proposed HI-CNN follows a three-stage hierarchical decision process with an early-exit mechanism[10], [11].

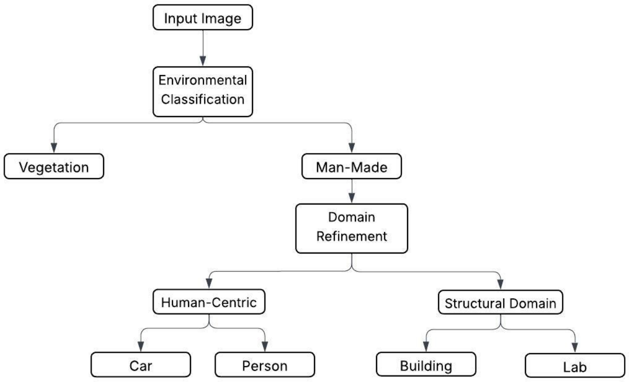


Fig. 1: Overview of the proposed HI-CNN hierarchical classification framework. The system performs a three-stage decision process, progressing from environmental classification to domain refinement and fine-grained scene recognition. Vegetation is treated as a terminal class, enabling early exit and reduced inference cost.

Stage 1: Environmental Classification. The input image is first classified into Vegetation or Man-Made categories. Vegetation is treated as a terminal class, allowing early termination of the inference process and preventing unnecessary downstream computation.

Stage 2: Domain Refinement. Images classified as Man-Made are further separated into Human-Centric and Structural domains, which reduces visual ambiguity and constrains the subsequent classification tasks to semantically coherent subsets.

Stage 3: Fine-Grained Classification. Two domain-specific branches perform final classification. The Human-Centric branch discriminates between Car and Person, while the Structural branch distinguishes between Building and Laboratory.

This hierarchical decomposition simplifies complex decision boundaries by constraining each classifier to a semantically coherent subset of classes.

B. Dataset and Experimental Setup

Dataset Description. We evaluate our framework on a custom urban scene dataset comprising five fine-grained categories: *Tree*, *Building*, *Laboratory*, *Car*, and *Person*. The dataset consists of real-world images captured by undergraduate students from the University of Jordan across various locations on and around the university campus, reflecting realistic urban and semi-urban environments. The dataset is divided into training/validation (80%) and test (20%) sets, with balanced class distribution maintained across all splits [13]. For incremental learning validation, we augment the Human-Centric branch with *Bus* images sampled from the CIFAR-100 dataset.

Splits and Preprocessing. The training/validation set is further split 75%/25% within the 80% allocated for training (resulting in 60% training, 20% validation, and 20% test of the total dataset). All images are resized to 224×224 pixels and normalized using EfficientNet's `preprocess_input` function, which applies the same input scaling used during ImageNet pretraining. Data augmentation is applied during training and includes random rotation ($\pm 10°$), zooming (up to 30%), horizontal flipping, and minor spatial shifts.

Experimental Configuration. All experiments were conducted on Google Colab using a Tesla T4 GPU. Each hierarchical node was trained independently with the Adam optimizer (learning rate 10^{-3}, batch size 32) for up to 15 epochs, with early stopping (patience = 3) based on validation loss.

C. NODE CONFIGURATION AND TRANSFER LEARNING

Each node in the hierarchy employs an EfficientNet-B0 backbone initialized with ImageNet pretrained weights. Transfer learning is used to preserve low-level visual features while fine-tuning higher-level representations for node-specific tasks[14]. The decision function at node N [11] is defined as:

$$f_N(x) = \phi(W_N \cdot x + b_N) \quad (1)$$

where x denotes the extracted feature vector, W_N and b_N are node-specific parameters, and $\phi(\cdot)$ is a non-linear activation function.

D. HIERARCHICAL INFERENCE LOGIC

The final class prediction is obtained by traversing a single path through the hierarchy[11]. The probability of the final classification is modeled as a conditional probability chain:

$$P(C_{final}|I) = \prod_{i=1}^{3} P(S_i|S_{i-1}) \tag{2}$$

where $S_0 = I$ represents the input image and S_i denotes the decision at stage i. An early-exit mechanism allows inference to terminate once a terminal class is reached, improving computational efficiency[10].

This design enables light incremental inference, as new branches can be integrated without retraining previously learned nodes.

E. Training Strategy and Optimization

Each node in the HI-CNN hierarchy is trained independently using a supervised learning paradigm. All decision nodes are formulated as binary classification problems, where each classifier discriminates between two semantically coherent categories. This design ensures that each model focuses exclusively on a narrow and well-defined decision boundary, reducing feature competition and improving convergence stability.

Binary cross-entropy loss is used for all nodes, as each stage represents a mutually exclusive binary decision. The Adam optimizer is employed due to its adaptive learning rate properties, which provide stable convergence across heterogeneous classification tasks. During training, early stopping based on validation loss is applied to prevent overfitting and reduce unnecessary computation.

This independent, binary optimization strategy further reinforces the modularity of the framework, as individual nodes can be retrained or replaced—such as when introducing a new category in a specific branch—without affecting the remaining hierarchy.

IV. Results and Performance Analysis

This section evaluates the performance of the HI-CNN framework using accuracy, loss convergence, and class-wise precision metrics.

A. Classification Accuracy

To evaluate the incremental extensibility of the proposed framework, an additional *Bus* category was introduced exclusively within the Stage-3 Human-Centric branch at inference time. Specifically, the original Car-versus-Person classifier was replaced with a Car-versus-Bus binary classifier, while all upstream nodes remained frozen. This class was not included in the initial hierarchical design and serves to validate the light incremental capability of the HI-CNN without retraining upstream nodes. The overall framework achieved a top-1 classification accuracy of 95%. The class-wise distribution of predictions, including the incrementally added *Bus* category, is illustrated in the confusion matrix shown in Fig. 2.

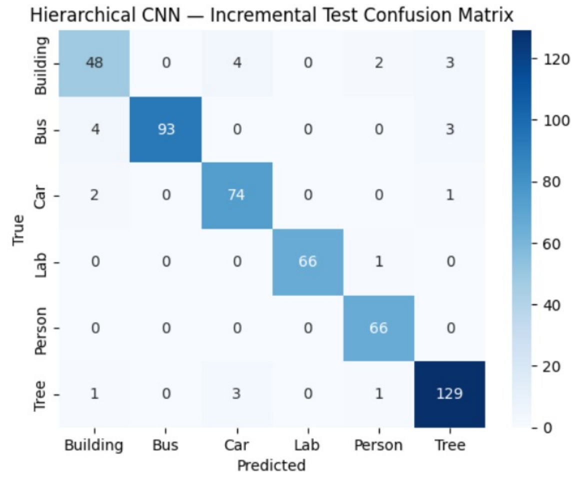


Fig. 2: Incremental Test Confusion Matrix. The Stage-1 environmental gating effectively isolates the ’Tree’ class (129 correct) from urban structures.

The confusion matrix reveals that the hierarchical structure significantly mitigates cross-domain errors. For instance, ”Tree” classification shows minimal leakage into ”Building,” and the Stage-3 nodes successfully distinguish between ”Bus” and ”Car” with 93 and 74 correct predictions, respectively.

To further assess the effectiveness of the proposed hierarchical design, we compare the HI-CNN framework with a flat CNN baseline in terms of classification accuracy and relative inference cost. As summarized in Table I, the proposed model achieves superior accuracy while reducing computational overhead during inference.

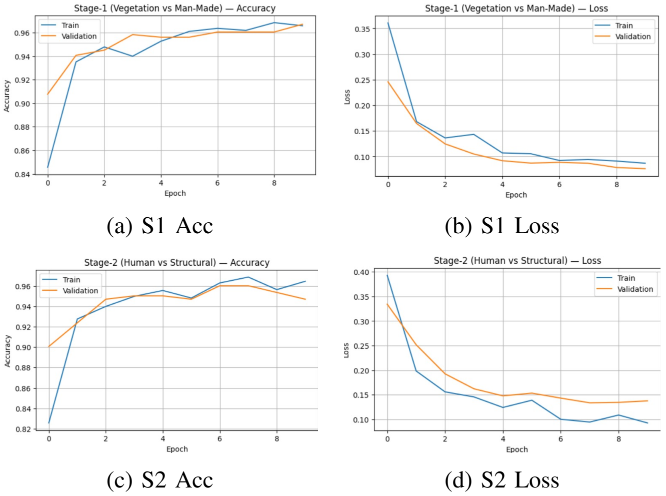
TABLE I: Comparison Between Flat and Hierarchical Classification Models

Model Type	Accuracy (%)	Relative Inference Cost
Flat CNN	92.1	High
HI-CNN (Proposed)	95.0	Reduced

This performance gain confirms that hierarchical decomposition not only improves classification accuracy but also enables more efficient inference by activating only a subset of specialized models per input. Compared to the flat CNN baseline, the proposed HI-CNN framework achieves superior performance while reducing redundant computation, demonstrating the practical benefits of structured decision routing.

B. Training Convergence

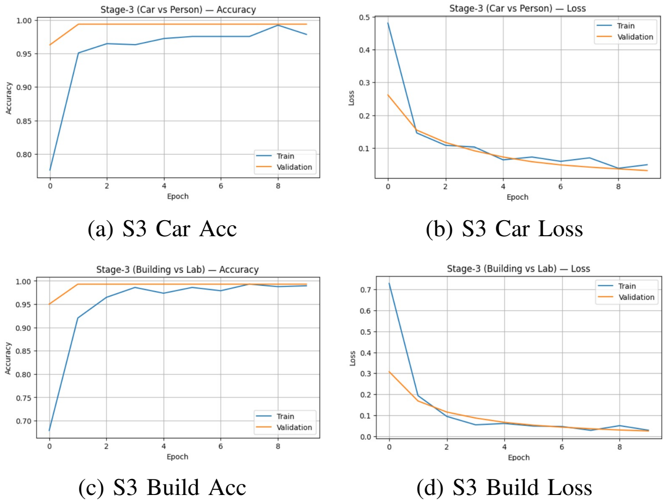
The training health of the HI-CNN hierarchy is evaluated through the accuracy and loss curves of each decision node. As shown in Fig. 3, the high-level nodes (Stage-1 and Stage-2) demonstrate stable learning, with validation accuracy consistently tracking the training curve. This convergence suggests



(a) S1 Acc (b) S1 Loss

(c) S2 Acc (d) S2 Loss

Fig. 3: Performance for High-Level Nodes (Stage-1 and Stage-2).



(a) S3 Car Acc (b) S3 Car Loss

(c) S3 Build Acc (d) S3 Build Loss

Fig. 4: Performance for Stage-3 Leaf Nodes (Car/Person and Building/Lab).

that the EfficientNet-B0 backbone successfully extracts the foundational structural and environmental features required for top-level branching.

The performance of the fine-grained leaf nodes is further detailed in Fig. 4. Notably, the Stage-3 ”Building vs. Lab” node exhibits exceptionally rapid convergence, reaching near-perfect validation accuracy within the first two epochs. This sharp decline in loss (Fig. 4d) compared to the more gradual stabilization seen in Stage-1 indicates that the hierarchical feature gating successfully isolates highly discriminative architectural traits, simplifying the classification task for downstream nodes. Across all stages, the minimal gap between training and validation metrics proves the system’s resilience against overfitting.

C. Impact of Hierarchical Decomposition

To evaluate the contribution of hierarchical decomposition, we analyze model behavior at each stage of the pipeline. Stage-1 environmental classification achieves high recall for Vegetation, confirming the effectiveness of early semantic separation. Domain refinement further reduces ambiguity by isolating human-centric and structural features before fine-grained classification. This staged decision-making process simplifies feature learning and improves prediction consistency compared to flat multiclass models.

V. Discussion on Architectural Efficiency and Deployability

While classification accuracy is a primary metric, the practical utility of our HI-CNN framework is defined by its architectural efficiency and its readiness for real world deployment. By decomposing the five class problem into a hierarchical pipeline using an EfficientNet-B0 backbone, we achieve several critical advantages over standard ”flat” multi-class models.

A. Computational Resource Optimization

A key challenge in multi-stage deep learning pipelines is the cumulative inference latency introduced by sequential model execution. This challenge is effectively mitigated through the use of the EfficientNet-B0 backbone, which leverages compound scaling to balance network depth, width, and input resolution while maintaining a low parameter count and high representational efficiency[16]. As a result, the total inference time of the three-stage hierarchical pipeline remains within real-time constraints, despite the presence of multiple decision stages. Furthermore, the early-exit mechanism employed at higher levels of the hierarchy ensures that not all stages are evaluated for every input, further reducing unnecessary computation[10]. This lightweight architectural design enables practical deployment on commodity hardware and edge platforms[15], such as mobile monitoring units or unmanned aerial systems, without reliance on high-cost GPU infrastructure, thereby lowering deployment barriers and operational costs.

B. Incremental Modularity and System Scalability

The hierarchical organization of the proposed HI-CNN framework enables incremental modularity[9], allowing localized updates and extensions without requiring full system retraining. Unlike flat multiclass models, where modifying a single class often necessitates retraining the entire network, the proposed design isolates classification responsibilities within individual nodes. For instance, refining the decision boundary between Car and Person requires updating only the Stage-3 Human-Centric branch, while upstream classifiers responsible for environmental and domain-level separation remain unchanged[12]. This inference-time modularity significantly simplifies system maintenance, reduces retraining time and computational cost, and supports scalable expansion as new classes or domains are incorporated. Such flexibility is particularly advantageous in dynamic urban monitoring scenarios, where classification requirements may evolve over time.

C. Feature Gating and Error Resilience

The hierarchical structure of the HI-CNN framework acts as a logical feature-gating mechanism that enhances robustness and error resilience. By separating Vegetation from Man-Made entities at the initial classification stage, the system effectively filters out natural textures and patterns before engaging in more complex structural or human-centric analysis. This coarse-to-fine decision process reduces cross-domain confusion, a common failure mode in flat classifiers where visually dissimilar categories compete within a single decision space[8]. Additionally, the explicit separation of semantic domains improves the transparency of the inference process, enabling human operators to trace classification outcomes to specific decision nodes. This interpretability facilitates error diagnosis, system auditing, and targeted refinement, which are critical requirements for deployment in real-world and safety-sensitive applications.

D. Limitations and Future Directions

While the current framework provides a robust foundation for hierarchical classification, it is characterized by certain limitations that define our future research agenda. The present ”light incremental” approach allows for the addition of new nodes and branches without retraining the entire hierarchy; however, it does not yet support full weight adaptation or continual learning across existing nodes, unlike more comprehensive incremental learning frameworks such as iCaRL [12]. This means that while the structure is scalable, the individual feature extractors remain static once deployed.

To address this limitation, future work will focus on transitioning from the current light incremental model toward a fully dynamic learning system. This extension will explore weight-consolidation techniques to mitigate catastrophic forgetting[6], enabling the model to adapt and refine its internal representations as new data becomes available, while preserving its baseline performance of 95% accuracy on previously learned classes.

VI. Conclusion

This study successfully implemented a hierarchical CNN for urban scene recognition, achieving a classification accuracy of 95%. The results demonstrate that Stage-1 environmental gating significantly reduces cross-domain confusion, as evidenced by the high recall for the Tree class. Furthermore, the modular design allows for future expansion into more complex urban categories without requiring full system retraining.

In addition, the modular nature of the system provides a scalable solution for incremental updates, allowing for the addition of new urban categories without exhaustive retraining. Future work will focus on expanding the hierarchy to include more complex urban entities and exploring the deployment of this framework on edge-computing devices for real-time monitoring applications.

References

[1] M. Tan and Q. Le, ”EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *ICML*, 2019.

[2] Z. Yan et al., ”HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition,” *ICCV*, 2015.

[3] C. Guo et al., ”On Calibration of Modern Neural Networks,” *ICML*, 2017.

[4] Y. Geifman and R. El-Yaniv, ”Selective Classification for Deep Neural Networks,” *NIPS*, 2017.

[5] K. He et al., ”Deep Residual Learning for Image Recognition,” *CVPR*, 2016.

[6] G. I. Parisi et al., ”Continual lifelong learning with neural networks: A review,” *Neural Networks*, 2019.

[7] K. Kowsari, ”Hierarchical Medical Image Classification, A Deep Learning Approach,” *BiB*, 2020.

[8] B. Zhao, J. Feng, X. Wu, and S. Yan, ”A Survey on Deep Learning-based Fine-grained Object Classification and Semantic Segmentation,” *International Journal of Automation and Computing*, 2017.

[9] Y. Wei, X. Zhang, W. Li, X. Wang, and B. Zhou, “Tree-CNN: A Hierarchical Deep Convolutional Neural Network for Incremental Learning,” *Neural Networks*, 2020.

[10] Surat Teerapittayanon, Bradley McDanel, and H. T. Kung, ”BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks,” in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2016.

[11] Yoshua Bengio, Nicolas Léonard, and Aaron Courville, ”Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation,” *arXiv preprint arXiv:1308.3432*, 2013.

[12] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. Lampert, “iCaRL: Incremental Classifier and Representation Learning,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

[13] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, ”Places: A 10 Million Image Database for Scene Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2018.

[14] R. Wightman, ”EfficientNet PyTorch: A Comprehensive Transfer Learning Study,” *arXiv preprint arXiv:2104.00298*, 2021.

[15] M. Xu, D. Zhu, Y. Liu, and J. Liu, ”Edge Intelligence: Architectures, Challenges, and Applications,” *IEEE Internet of Things Journal*, 2020.

[16] M. Tan and Q. Le, ”EfficientNetV2: Smaller Models and Faster Training,” *International Conference on Machine Learning (ICML)*, 2021.