Check for updates

RESEARCH ARTICLE

Fine-grained visual recognition of residential air conditioner units for building energy management

[version 1; peer review: awaiting peer review]

Rafail Daskos, Nikos Dimitropoulos, Ioannis Papias, Vangelis Marinakis

Decision Support Systems Laboratory, School of Electrical & Computer Engineering, National Technical University of Athens, Athens, Greece

v1 **First published:** 06 Oct 2026, **6**:389
https://doi.org/10.12688/openreseurope.24354.1
Latest published: 06 Oct 2026, **6**:389
https://doi.org/10.12688/openreseurope.24354.1

Abstract

Background

Accurate identification of installed appliances is important for energy analytics, demand-side management, automated building inventories, and energy auditing. Residential air-conditioning systems are among the most energy-intensive household appliances, yet detailed information about installed units is often unavailable or difficult to collect at scale. This study investigates the recognition of residential air-conditioning indoor units directly from images as a fine-grained visual classification problem.

Methods

A curated dataset comprising 720 images from 216 air-conditioner manufacturer-model categories was constructed by combining field-acquired photographs captured through a custom mobile application with supplementary web-sourced images. Five representative deep learning architectures were evaluated under a unified experimental protocol. The best-performing baseline was subsequently refined through controlled ablation experiments examining activation functions, classification heads, learning-rate schedulers, and backbone architectures.

Results

Cross-layer feature aggregation provided the strongest baseline performance. Further improvements were obtained using GELU activation, CosFace cosine-margin supervision, OneCycleLR scheduling, and a DenseNet-161 backbone. The final CN-CNN

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

configuration achieved a validation Top-1 accuracy of 74.03%, a Top-3 accuracy of 87.01%, and a macro F1-score of 0.7054, representing an improvement of 11.69 percentage points in Top-1 accuracy over the initial CN-CNN baseline.

Conclusions

The findings demonstrate that fine-grained visual recognition can support scalable and non-intrusive identification of residential air-conditioning units. The proposed framework provides a basis for integrating computer vision into automated building inventories, energy-auditing workflows, and smart energy-management applications.

Keywords
Fine-Grained Visual Classification, Air Conditioner Recognition, CN-CNN, Appliance Identification, Computer Vision, Deep Learning, Building Energy Management

HE This article is included in the Horizon Europe gateway.

Corresponding author: Rafail Daskos (rdaskos@epu.ntua.gr)

Author roles: Daskos R: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Supervision, Visualization, Writing – Original Draft Preparation; **Dimitropoulos N**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Papias I**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Marinakis V**: Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101136216 (project HEDGE-IoT), grant agreement No 101104141 (project BuildON), and grant agreement No 101138373 (project META BUILD).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Daskos R *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License , which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Daskos R, Dimitropoulos N, Papias I and Marinakis V. **Fine-grained visual recognition of residential air conditioner units for building energy management [version 1; peer review: awaiting peer review]** Open Research Europe 2026, **6**:389 https://doi.org/10.12688/openreseurope.24354.1

First published: 06 Oct 2026, **6**:389 https://doi.org/10.12688/openreseurope.24354.1

Nomenclature

AcronymsAC: Air Conditioning
AI: Artificial Intelligence
CN-CNN: Cross-layer Navigation Convolutional Neural Network
CNN: Convolutional Neural Network
CosFace: Large Margin Cosine Loss
CV: Computer Vision
DL: Deep Learning
DSM: Demand-Side Management
FGVC: Fine-Grained Visual Classification
FDD: Fault Detection and Diagnosis
GELU: Gaussian Error Linear Unit
HVAC: Heating, Ventilation and Air Conditioning
LR: Learning Rate
MixDCNN: Mixture of Deep CNNs
OneCycleLR: One Cycle Learning Rate Scheduler
ReduceLROnPlateau: Adaptive Learning Rate Scheduler
ReLU: Rectified Linear Unit
ResNet: Residual Network
NILM: Non-Intrusive Load Monitoring

Variables and parameters

D: Image dataset used for training and evaluation
N: Total number of samples in the dataset
C: Number of AC model classes
x_i: Input image sample i
y_i: Ground-truth class label of sample i
$f_\theta(\cdot)$: Neural network parameterized by weights θ
θ: Trainable parameters of the neural network
z_i: Feature embedding vector of image x_i
W: Weight matrix of the classification layer
w_j: Weight vector corresponding to class j
$\hat{y}_{i,j}$: Predicted logit score for sample i and class j
θ_j: Angle between embedding z_i and class weight w_j in cosine space
m: Cosine margin parameter in CosFace loss
s: Scaling factor applied in cosine-margin classification

Evaluation metrics

Top-1: Percentage of samples where the correct class is the highest-scoring prediction
Top-3: Percentage of samples where the correct class appears among the top three predictions
Precision: Ratio of correctly predicted positive observations to total predicted positives
Recall: Ratio of correctly predicted positive observations to all actual positives
F1-score: Harmonic mean of precision and recall
M_{GPU}: GPU Memory Consumption
Train Time/E: Train Time per Epoch (sec)
Val Time/E: Test/Validation Time per Epoch (sec)

1. Introduction

Improving energy efficiency and enabling demand-side flexibility have become central priorities in modern power systems. The increasing penetration of renewable energy sources introduces variability and uncertainty in electricity generation, creating the need for mechanisms that can balance supply and demand across the grid. In this context, demand-side management (DSM) and demand response strategies have emerged as key tools for improving grid stability and enabling more efficient utilization of energy resources.[1,2,3] Residential buildings play a particularly important role in this transition, as they account for a substantial share of global electricity consumption and therefore offer considerable potential for flexible load management.[4,5,6]

Within residential buildings, household appliances account for a large portion of electricity consumption. Major end-use categories include heating, cooling, lighting, refrigeration, and various plug-in devices.[4,7] Among these loads, space

conditioning systems such as air conditioners (ACs) represent some of the most energy-intensive appliances. Global demand for cooling is expected to increase substantially in the coming decades due to climate change, urbanization, and rising living standards.[8] Because of their thermal inertia and controllable operation, AC systems are widely considered promising candidates for demand response programs and building energy management strategies.[9,10]

Despite their importance in residential energy consumption, detailed information about installed appliances is often unavailable. Traditional approaches for identifying household equipment typically rely on manual surveys, equipment documentation, or intrusive monitoring techniques that require dedicated sensors installed on individual devices. These approaches are often costly, difficult to scale, and dependent on user participation.[11,12,13,14] As a result, considerable research effort has focused on the development of non-intrusive techniques capable of identifying appliances without requiring additional sensing infrastructure.

Several approaches have been explored for appliance identification in residential environments. Non-intrusive load monitoring (NILM) techniques attempt to infer appliance operation from aggregated electrical measurements, while computer vision methods aim to identify devices directly from visual data captured by cameras or mobile devices.[11,15,16,17,18] While NILM-based approaches have demonstrated promising results for detecting appliance usage, they often struggle to distinguish between devices with similar electrical signatures and rarely provide identification at the manufacturer or model level.

Recent advances in computer vision and deep learning have significantly improved the ability of machines to recognize objects directly from images. Convolutional neural networks (CNNs) have enabled major progress in visual recognition tasks and have been successfully applied in numerous real-world applications.[19,20,21] However, recognizing appliances at the manufacturer or model level remains challenging because many devices exhibit highly similar visual characteristics.

This problem belongs to the domain of fine-grained visual classification (FGVC), where the objective is to distinguish between visually similar categories based on subtle differences in appearance or structure.[22,23,24] While FGVC methods have been widely studied in domains such as species recognition and vehicle classification, their application to residential appliance identification remains relatively limited.

In particular, the automatic recognition of indoor air-conditioning units at the manufacturer or model level has received little attention in the literature, despite its potential benefits for energy analytics and smart building applications. Accurate identification of installed equipment can support the creation of appliance inventories, facilitate building energy audits, and enable more precise estimation of appliance-level energy consumption. Such information is particularly valuable for demand response and flexibility estimation, where knowledge of device characteristics can improve load management strategies and support intelligent energy control systems.[2,9,10]

To address this gap, this work investigates the problem of recognizing residential air-conditioning indoor units using computer vision techniques. The task is formulated as a fine-grained visual classification problem, and a deep learning-based framework is developed to distinguish between visually similar AC models. A dedicated dataset of indoor AC units was constructed by combining images collected from real-world installations through a mobile data collection application with additional samples obtained from online sources. This dataset enables systematic benchmarking of deep learning architectures for appliance-level visual recognition.

The contributions of this study are summarized as follows:

- **Vision-based framework for fine-grained AC recognition.** We formulate the identification of residential air-conditioning indoor units as a fine-grained visual classification problem and propose a computer vision framework for model-level recognition.

- **Construction of a curated dataset of AC indoor units.** We introduce a dataset consisting of images of residential AC indoor units collected from both real-world installations and web sources, covering multiple manufacturers, models, and environmental conditions.

- **Systematic benchmarking of deep learning architectures.** We conduct a comprehensive evaluation of several deep learning architectures and training configurations, providing insights into the effectiveness of different approaches for fine-grained appliance recognition.

The remainder of this paper is organized as follows. Section 2 reviews related work on appliance identification, fine-grained visual classification, and deep learning approaches relevant to appliance-level recognition. Section 3 presents the proposed vision-based recognition framework. Section 4 describes the experimental setup and reports the benchmarking and ablation study results. Section 5 discusses the implications of the findings and practical considerations for deployment. Finally, Section 6 summarizes the main conclusions and outlines directions for future research.

2. Literature review

The literature on visual recognition of residential AC units can be examined through four complementary perspectives: asset identification practices in buildings, FGVC, CNN-based FGVC architectures, and part-based and cross-layer feature learning strategies. This structure reflects the progression from application-driven asset identification requirements to the methodological advances in deep learning that enable reliable discrimination between visually similar AC models.

2.1 AC unit recognition & asset identification

Accurate identification of installed building assets is essential for energy management, maintenance planning, lifecycle tracking, and automated auditing in the built environment. Heating, ventilation, and air-conditioning (HVAC) systems represent a dominant share of electricity consumption in residential and commercial buildings, making reliable equipment identification particularly important for energy performance analysis and operational optimization.[4]

Traditional asset identification practices rely on manual surveys, equipment documentation, nameplate inspection, and maintenance records. Although such approaches can provide detailed information, they are labor-intensive, error-prone, and difficult to maintain at scale across large building portfolios. The increasing digitization of the construction and facility management sectors has motivated the exploration of automated visual methods for asset documentation and inspection.[25] These studies demonstrate the potential of computer vision (CV) to reduce manual effort and improve consistency in visual assessment tasks; however, they primarily address coarse-grained classification problems, such as structural element recognition or building component labeling.

Within the HVAC domain specifically, existing research has largely focused on operational monitoring, fault detection, and performance diagnostics rather than visual equipment identification. Fault detection and diagnosis (FDD) methods for building systems have been extensively reviewed in the literature.[26,27] Data-driven FDD frameworks use operational sensor measurements and machine-learning techniques to detect and diagnose anomalies in HVAC equipment.[28] Similarly, NILM approaches attempt to infer appliance presence and activity from aggregated electrical consumption signals.[11,12] While effective for detecting appliance categories or abnormal operating behavior, these methods do not provide reliable differentiation between manufacturer and model variants of visually similar air-conditioning units.

In parallel, advances in large-scale object recognition, most notably through datasets such as ImageNet, a large-scale visual recognition benchmark, have enabled deep convolutional neural networks to achieve high performance across thousands of object categories.[19] Nevertheless, generic object recognition systems are optimized for broad semantic distinctions rather than subtle inter-class variations. This limitation becomes critical when distinguishing between visually similar AC indoor units that share common geometric layouts, color schemes, and structural components.

Although computer vision has been successfully applied to indoor object detection and scene understanding,[29] systematic investigation of model-level visual recognition for residential AC units remains limited. Given the high visual similarity across AC models, the task cannot be adequately addressed as a generic object classification problem. Instead, it requires methodologies capable of capturing fine discriminative details between closely related subcategories.

These considerations motivate framing AC unit recognition as a fine-grained visual classification problem, which is discussed in the following subsection.

2.2 Fine-grained visual classification (FGVC)

FGVC refers to the task of distinguishing between subordinate categories within the same superordinate class, where inter-class differences are subtle and intra-class variations can be significant.[22,30] Unlike generic object recognition, which focuses on broad semantic categories (e.g., “car” versus “chair”), FGVC aims to differentiate closely related subcategories, such as bird species, aircraft models, or vehicle variants.

The primary challenge in FGVC lies in the small visual discrepancies between classes combined with variations in pose, illumination, occlusion, and background clutter. Early approaches addressed this challenge by incorporating part-based modeling and explicit localization mechanisms. For example, part-based R-CNN frameworks were proposed to detect

discriminative object parts and improve classification accuracy.[30] Similarly, pose-normalized representations and localized feature extraction strategies were introduced to enhance sensitivity to subtle structural differences.[31]

Subsequent advances in deep learning shifted the focus toward end-to-end CNN architectures capable of learning discriminative representations directly from image data. Bilinear CNN models demonstrated that pairwise feature interactions can significantly improve fine-grained discrimination by capturing higher-order feature correlations.[32] These representations proved particularly effective in domains with subtle inter-class variations.

More recent developments have incorporated attention mechanisms and feature aggregation strategies to automatically emphasize informative regions without requiring explicit part annotations. Weakly supervised localization approaches and attention-based pooling methods have shown strong performance on benchmark FGVC datasets such as CUB-200-2011 and Stanford Cars.[33,34] These methods enable models to focus on discriminative visual cues while remaining robust to background noise.

Despite extensive progress in FGVC for biological species, vehicles, and aircraft, its application to appliance-level equipment recognition remains limited. Residential air-conditioning indoor units present characteristics consistent with fine-grained classification: models share similar global structure, differ in subtle geometric or design features, and may appear under varying lighting and installation conditions. Consequently, framing AC unit recognition within the FGVC paradigm provides a principled foundation for addressing model-level discrimination challenges.

2.3 CNN-based FGVC architectures

Fine-grained visual classification (FGVC) has progressed through a series of architectural developments aimed at improving discrimination between visually similar categories. Since no architecture has been specifically designed for residential AC unit recognition, it is necessary to examine representative FGVC models developed in related domains. As summarized in Table 1, these approaches can be broadly grouped into three categories: bilinear and pooling-based models, attention-based models and deep/hybrid architectures. This categorization reflects the different representational strategies adopted to enhance sensitivity to subtle inter-class variations.

2.3.1 Bilinear and pooling-based models

Bilinear pooling approaches were among the earliest deep learning methods to demonstrate strong performance in FGVC. Bilinear CNN (B-CNN) models compute outer products between feature channels, capturing second-order feature

Table 1. Summary and comparison of FGVC models on the CUB-200-2011 dataset.

Model	Key innovations	Accuracy (%)
Category: Bilinear and Pooling-Based Models		
B-CNN (2015)	Bilinear pooling of pairwise feature interactions[32]	84.1
LCBP-CNN (2016)	Compact bilinear pooling via Tensor Sketch projection[35]	84.0
LRBP (2017)	Low-rank matrix decomposition for efficient bilinear pooling[36]	86.2
HBP-CNN (2018)	Hierarchical/high-order bilinear pooling via tensor decomposition[37]	87.1
MixDCNN (2016)	Mixture of deep CNN experts capturing complementary features[38]	85.0
Category: Attention-Based Models		
DVAN (2017)	Diversified attention across multiple discriminative regions[39]	84.8
RA-CNN (2017)	Recurrent attention for iterative part localization[33]	85.3
MA-CNN (2017)	Multi-attention learning of localized regions jointly[40]	86.5
DCAL (2022)	Dual cross-attention across spatial and channel dependencies[41]	92.0
Category: Deep and Hybrid Architectures		
ResNet-50 (2016)	Residual learning with skip connections[21]	84.5
CN-CNN (2021)	Cross-layer navigation for adaptive feature integration[42]	91.6
CoAtNet (2021)	Hybrid CNN-Transformer combining convolution and self-attention[43]	92.3

interactions that significantly enhance discrimination between visually similar categories.[32] Such representations are particularly effective when class differences are expressed through subtle texture, contour, or structural patterns.

However, full bilinear representations are high-dimensional and computationally expensive. Compact bilinear pooling (LCBP-CNN) mitigates this issue using Tensor Sketch projections to approximate bilinear features efficiently.[35] Low-rank bilinear pooling (LRBP) further reduces complexity by factorizing bilinear matrices through low-rank decomposition.[36]

Hierarchical Bilinear Pooling (HBP-CNN) extends bilinear modeling by integrating cross-layer feature interactions, enabling the aggregation of multi-scale semantic information.[37] This hierarchical formulation allows both low-level structural cues and high-level semantic representations to contribute jointly to classification decisions. Such multi-scale feature fusion is particularly relevant in scenarios where discriminative information may be distributed across multiple feature depths.

MixDCNN introduces a mixture of experts strategy, combining multiple CNN branches trained to capture complementary feature subsets.[38] Rather than relying on explicit part detection, this architecture models intra-class variation implicitly through expert diversity, offering robustness in unconstrained visual conditions.

2.3.2 Attention-based models

Attention mechanisms were introduced to automatically localize and emphasize discriminative regions within an image. Diversified Visual Attention Network (DVAN) learns multiple complementary attention maps that focus on distinct object parts.[39] Recurrent Attention CNN (RA-CNN) iteratively refines attention regions, enabling progressive localization of informative areas.[33]

Multi-Attention CNN (MA-CNN) further advances this paradigm by jointly learning multiple attention branches within a unified framework, encouraging complementary regional representations.[40] More recently, Dual Cross-Attention Learning (DCAL) integrates spatial and channel attention mechanisms to model dependencies across both spatial regions and feature channels.[41] Such mechanisms improve sensitivity to subtle geometric variations while suppressing irrelevant background information.

Attention-based strategies are particularly advantageous in fine-grained tasks where discriminative cues are localized (e.g., slight curvature differences, minor structural elements) and may be masked by environmental clutter.

2.3.3 Deep and hybrid architectures

Standard deep convolutional backbones such as ResNet-50 provide strong baseline performance through residual learning and effective gradient propagation.[21] Their generalization capability makes them widely adopted starting points for FGVC tasks.

Cross-layer Navigation CNN (CN-CNN) enhances this paradigm by dynamically integrating features across multiple layers, enabling adaptive selection of discriminative representations.[42] By leveraging cross-layer dependencies, CN-CNN captures both local structural detail and high-level semantic context, offering improved robustness in fine-grained settings.

Hybrid convolution-attention architectures have recently emerged to combine the inductive biases of CNNs with the global modeling capacity of self-attention. CoAtNet integrates convolutional stages with Transformer-style attention modules in a unified framework.[43] This hybrid design enables simultaneous modeling of local texture patterns and long-range feature interactions, which is beneficial when discriminative cues depend on both local design elements and global structural layout.

Overall, these architectural developments reflect a progression from explicit second-order feature modeling to adaptive attention mechanisms and hybrid convolution Transformer representations. The models selected in this study represent key milestones across these categories, enabling systematic evaluation of different feature aggregation strategies for fine-grained AC unit recognition.

2.4 Representation principles for fine-grained discrimination

Beyond architectural taxonomies, fine-grained visual classification fundamentally depends on representational mechanisms capable of separating highly similar categories in feature space. Two complementary principles have emerged in the literature as critical for fine-grained discrimination: (i) localized discriminative representation and (ii) multi-level feature aggregation.

First, fine-grained categories often differ only in subtle, spatially localized regions. Unlike coarse object recognition, where global shape or dominant structures may suffice, fine-grained tasks require sensitivity to small geometric variations, texture patterns, contour curvature, or structural details. Effective representations must therefore emphasize informative regions while suppressing irrelevant background information. This can be achieved either through explicit part localization or implicitly through attention-guided feature weighting, enabling the model to focus on discriminative cues without requiring strong supervision.[30,33]

Second, discriminative information in fine-grained tasks is frequently distributed across multiple semantic levels. Lower convolutional layers encode fine texture and edge information, while deeper layers capture more abstract structural representations. Aggregating features across different network depths enables the integration of local structural detail with higher-level contextual semantics. Hierarchical feature interactions and cross-layer aggregation mechanisms enhance the expressiveness of learned representations by modeling correlations between low-level and high-level descriptors.[37,42]

Another important representational aspect concerns higher-order feature interactions. Standard convolutional networks primarily rely on first-order feature activations. However, modeling second-order statistics-such as feature covariance-can substantially increase discriminative capacity in settings where class differences are subtle. Bilinear and related pooling mechanisms exploit this principle by capturing channel-wise interactions, effectively encoding fine structural variations in a more expressive feature space.[32]

These representation principles are particularly relevant for residential air-conditioning indoor units. AC models often share nearly identical global layouts, with differences emerging in minor design elements such as vent segmentation, panel curvature, grille patterns, or branding placement. Such variations may appear at different spatial scales and feature depths, requiring models capable of combining localized discrimination, multi-scale aggregation, and higher-order feature interactions.

Collectively, these insights motivate the architectural choices explored in the subsequent sections, where models incorporating cross-layer integration and advanced feature aggregation are evaluated for fine-grained AC unit recognition.

3. Proposed framework

Building upon the representation principles and architectural strategies discussed in the previous section, this section presents the proposed framework for fine-grained recognition of residential AC indoor units. The objective is to adapt and systematically evaluate representative FGVC architectures within the context of appliance-level model discrimination, and to identify configurations that effectively capture subtle inter-class variations. The following subsections describe the overall pipeline, formal problem definition, baseline models, proposed architectural refinements, and implementation details.

3.1 Overview of the framework

Building upon the fine-grained visual classification principles discussed in Section 2, this section presents the AC Vision-Based Classification Framework developed for residential indoor unit recognition. The framework translates general FGVC methodologies into a structured pipeline tailored to the morphological characteristics and practical constraints of AC model identification.

The framework, as shown in Figure 1, is organized into five sequential stages.

1. **Dataset Construction.** A dedicated dataset of residential AC indoor units is assembled through a combination of real-world field acquisition and curated web-sourced imagery. This dual-source strategy ensures diversity in installation environments, lighting conditions, and manufacturer designs, while preserving label reliability through manual verification.

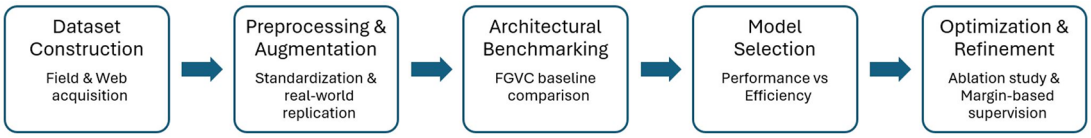


Figure 1. Framework overview.

2. **Data Preprocessing and Augmentation.** All images are standardized to a unified spatial resolution and normalized to ensure architectural compatibility. Grayscale conversion is applied to reduce color-induced bias and emphasize structural features. Data augmentation techniques, including geometric transformations, brightness and contrast adjustments, and noise injection, are applied during training to enhance robustness against real-world variability.

3. **Baseline Architectural Benchmarking.** Representative FGVC architectures from different methodological families are trained under a unified experimental protocol. Identical preprocessing steps, optimization settings, and evaluation metrics are employed to ensure fair comparison. This stage isolates the impact of architectural design on fine-grained AC recognition performance.

4. **Model Selection.** Based on quantitative performance and computational efficiency indicators, the most suitable baseline architecture is selected as the foundation for further refinement. The selection criterion balances discriminative accuracy with practical feasibility.

5. **Controlled Optimization and Refinement.** The selected architecture is subjected to systematic ablation studies, where individual components of the training pipeline are modified in isolation. These controlled experiments examine the influence of activation functions, classification heads with margin-based supervision, learning-rate scheduling strategies, and alternative backbone networks. This stage identifies the configuration that maximizes inter-class separability while maintaining computational tractability.

Performance evaluation is embedded throughout all stages of the framework, guiding architectural comparison and optimization decisions. The outcome of this structured process is an optimized configuration specifically adapted to the fine-grained recognition of residential AC indoor units.

3.2 Problem formulation

The recognition of residential AC indoor units is formulated as a supervised multi-class fine-grained visual classification problem. Let the dataset be defined as:

$$D = \{(x_i, y_i)\}_{\{i=1\}}^{\{N\}} \tag{1}$$

where N denotes the total number of samples, $x_i \in \mathbb{R}^{H \times W}$ represents an input image, and $y_i \in \{1, \ldots, C\}$ corresponds to one of C AC manufacturer-model categories.

A deep neural network parameterized by θ learns a nonlinear mapping:

$$f_\theta : \mathbb{R}^{H \times W} \to \mathbb{R}^d \tag{2}$$

producing a d-dimensional feature embedding:

$$z_i = f_\theta(x_i) \tag{3}$$

A linear classification layer with weight matrix $W = [w_1, \ldots, w_C] \in \mathbb{R}^{d \times C}$ maps embeddings to logits:

$$\hat{y}_{i,j} = w_j^T z_i \tag{4}$$

which are converted to posterior probabilities via the softmax function.[44]

Cross-Entropy Objective. The baseline model is optimized using the Cross-Entropy (CE) loss,[44] defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{exp\left(\widehat{y}_{i,y_i}\right)}{\sum_{j=1}^{C} exp\left(\widehat{y}_{i,j}\right)}\right) \tag{5}$$

As shown in Equation (5), Cross-Entropy maximizes the log-likelihood of the ground-truth class while suppressing competing logits. Although widely adopted for multi-class classification, CE does not explicitly enforce angular separation between class embeddings.

CosFace Margin-Based Objective. To enhance discriminative embedding learning in the fine-grained setting, margin-based supervision is investigated. Following the CosFace formulation,[45] both embeddings and class weights are L2-normalized:

$$\tilde{z}_i = \frac{z_i}{\|z_i\|}, \quad \tilde{w}_i = \frac{w_i}{\|w_i\|} \tag{6}$$

The cosine similarity between normalized vectors is then:

$$\cos\left(\theta_j\right) = \tilde{w}_j^T \tilde{z}_i \tag{7}$$

CosFace introduces an additive cosine margin m and a scaling factor s, leading to the modified objective

$$\mathcal{L}_{CosFace} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{s\left(\cos\left(\theta_{y_i}\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i}\right)-m\right)} + \sum_{i\neq y_i} e^{s\cos\left(\theta_j\right)}}\right) \tag{8}$$

As shown in Equation (8), the cosine margin m enforces stricter angular separation between classes, promoting tighter intra-class clustering and increased inter-class distance in the embedding space. Such angular-margin formulations have demonstrated improved discriminative performance in high-similarity recognition problems.[45]

3.3 Baseline architectures

To evaluate the suitability of different representation mechanisms for fine-grained AC unit recognition, five representative architectures were implemented and benchmarked under identical training conditions. Each model embodies a distinct strategy for discriminative feature learning. All models are trained under a unified protocol (Section 3.5) to ensure comparability.

ResNet-50[21] is a deep residual convolutional architecture that employs identity skip connections to facilitate stable gradient propagation in very deep networks. By learning residual mappings instead of direct transformations, the architecture improves optimization efficiency and feature reuse across layers. In this study, ResNet-50 serves as a standard convolutional baseline for evaluating more specialized fine-grained architectures. The network was initialized with ImageNet-pretrained weights, and its final fully connected layer was replaced with a task-specific classifier of dimension C corresponding to the number of AC model categories.

MixDCNN[38] adopts a mixture-of-experts strategy in which multiple convolutional branches learn complementary representations. The outputs of these branches are aggregated before classification, enabling the network to capture diverse local and global patterns. Local and global patterns. Given the variability in installation environments and lighting conditions present in the AC dataset, MixDCNN was evaluated for its potential robustness to appearance diversity and intra-class variability.

HBP-CNN[37] extends bilinear pooling by modeling interactions between feature maps extracted at different convolutional depths. Given feature tensors $F^{(l1)}$ and $F^{(l2)}$ from two layers, bilinear pooling computes outer-product interactions followed by normalization, enabling second-order feature statistics to enhance discriminative capacity. This mechanism increases sensitivity to subtle geometric variations that may distinguish visually similar AC models.

CN-CNN[42] aggregates feature maps across multiple network depths through a progressive cross-layer fusion mechanism. Instead of relying solely on final-layer features, intermediate representations are adaptively integrated, allowing both low-level structural cues and higher-level semantic patterns to contribute to classification. Since the target task involves static image recognition rather than temporal sequences, the temporal aggregation component of the original CN-CNN formulation was removed. The architecture was adapted to operate in a single-frame configuration while preserving cross-layer fusion mechanisms.

CoAtNet[43] combines convolutional blocks with self-attention modules in a hierarchical structure. Early stages employ convolution to capture local texture and edge information, while later stages incorporate self-attention to model long-range dependencies. This hybrid design was evaluated to assess whether global context modeling improves discrimination among AC models with similar local structures. Recent FGVC research also explores Transformer-dominant or attention-heavy designs to capture long-range dependencies while maintaining sensitivity to fine detail.[46]

Overall, these baselines provide complementary inductive biases for fine-grained discrimination: residual backbones offer strong general-purpose representations,[21] mixture-based learning captures diverse appearance modes,[38] bilinear/cross-layer aggregation increases sensitivity to subtle structural cues,[37,42] and hybrid attention models incorporate global context modeling.[41,43] The comparative performance of these models on the AC dataset is reported in Section 4.3.

3.4 Proposed architecture

The proposed architecture is not defined as a single predefined network configuration, but as a structured model design and refinement framework tailored to fine-grained AC unit recognition. Its objective is to systematically identify the architectural and optimization components that most effectively enhance discrimination among visually similar AC models. The architecture is built upon three core principles as shown in Figure 2:

1. *Baseline Evaluation and Selection.* Multiple representative FGVC architectures are first benchmarked under a unified experimental protocol (Section 4.3). This ensures that model comparison is not influenced by inconsistent preprocessing or optimization settings. The best-performing baseline serves as the foundation for further refinement.

2. *Controlled Architectural Refinement.* Starting from the selected baseline, individual architectural dimensions are systematically modified and evaluated. These include:
   - Activation function choice (e.g., ReLU[47] vs. GELU[48]),
   - Classification head formulation (standard Cross-Entropy[44] vs. cosine-margin supervision[45]),
   - Learning-rate scheduling strategies,
   - Backbone network selection (residual vs. dense connectivity).

Each modification is evaluated in isolation to determine its direct impact on fine-grained classification performance.

3. *Integration into a Unified End-to-End Model.* The final architecture integrates the components that demonstrate the strongest contribution to inter-class separability and embedding stability. Feature extraction, cross-layer aggregation (inspired by CN-CNN[42]), and margin-based classification are trained jointly in an end-to-end manner as defined in Section 3.2.

This structured refinement process ensures that architectural decisions are driven by empirical validation rather than a priori assumptions. The resulting configuration represents the most suitable model for fine-grained AC indoor unit recognition under the constraints of the available dataset and deployment considerations. The quantitative evaluation of each refinement step is presented in Section 4.4.

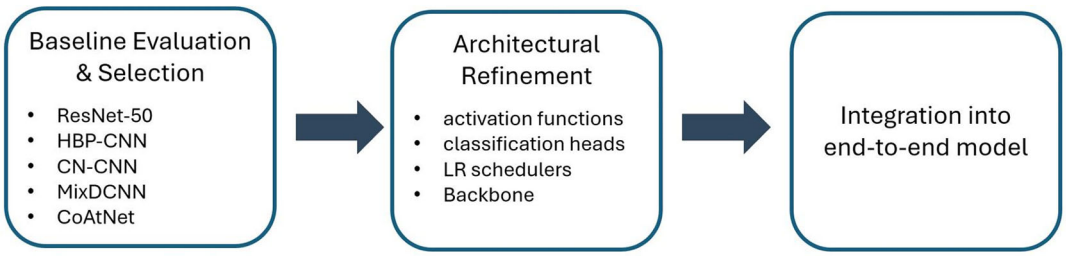


Figure 2. Architecture Principles.

3.5 Training strategy and implementation details

To ensure fair comparison and reproducibility, all models were trained under a unified optimization protocol so that architectural differences could be evaluated without confounding hyperparameter variations. Model parameters θ were optimized using the Adam optimizer,[49] which updates parameters according to:

$$\theta_{t+1} = \theta_t - \eta \frac{\widehat{m}_t}{\sqrt{\widehat{u}_t} + \epsilon} \tag{9}$$

where $\widehat{m}_t$ and $\widehat{u}_t$ denote the bias-corrected first and second moment estimates of the gradients and η represents the learning rate. The initial learning rate was set to 1×10^{-3}.

Learning-rate scheduling was applied to improve convergence behavior. For baseline experiments, the OneCycle learning-rate policy[50] was adopted. This scheduler cyclically varies the learning rate, enabling rapid exploration of the loss landscape followed by gradual annealing. During architectural refinement experiments, the ReduceLROnPlateau scheduler was also evaluated, which adaptively reduces the learning rate when validation performance stagnates. All models were implemented using PyTorch.[51]

Training was performed using mini-batches of size $B = 16$, balancing gradient stability and computational efficiency in stochastic gradient-based optimization.[44] The maximum number of training epochs was set to 200. To avoid overfitting, early stopping was applied with patience $p = 10$, terminating training when the validation loss failed to improve for p consecutive epochs:

$$Stop\ if\ \mathcal{L}_{val}^{(t)} > \min_{k<t} \mathcal{L}_{val}^{(k)}\ \ for\ p\ epochs \tag{10}$$

The images were resized to (224×224) pixels for the baseline experiments, consistent with standard ImageNet-pretrained backbone architectures.[21] During later refinement stages, higher-resolution inputs of (448×448) pixels were used to better preserve fine spatial details that are particularly important in fine-grained visual classification tasks. To improve model generalization and simulate real-world variability in image acquisition conditions, online data augmentation was applied to the resized images during training. Augmentation operations included random horizontal flipping, random rotations, brightness and contrast perturbations, and mild Gaussian noise injection. Such augmentation strategies are widely used to improve robustness in deep learning models trained on limited datasets.[52]Training was further accelerated by using automatic mixed-precision computation, which reduces GPU memory usage and improves computational efficiency while maintaining numerical stability.[53] Finally, model performance was evaluated using multiple predictive metrics. Top-k accuracy was computed as:

$$Top-k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(y_i \in \widehat{y}_i^{(k)}\right) \tag{11}$$

where $\widehat{y}_i^{(k)}$ represents the set of top-k predicted classes for sample i, and $\mathbb{I}(\cdot)$ is the indicator function. In the experiments, Top-1 and Top-3 accuracy metrics were used. In addition, precision, recall, and F1-score were computed in the multiclass setting.[44] For each class c, precision and recall are defined as:

$$Precision = \frac{TP_c}{TP_c + FP_c}, \quad Recall = \frac{TP_c}{TP_c + FN_c} \tag{12}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c. The F1-score is given by:

$$F1_c = 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{13}$$

To account for potential class imbalance, macro-averaged values of these metrics were reported. In addition to predictive performance, computational efficiency indicators-including training time per epoch, validation time per epoch, and GPU memory consumption-were recorded to assess the practical feasibility of deploying the proposed models.



(a) Field-acquired images captured under real installation conditions



(b) Web-sourced images obtained from manufacturer and retail platforms

Figure 3. De-identified representative samples from the AC indoor unit dataset.

4. Experiments and results

To validate the effectiveness of the proposed fine-grained AC recognition framework, comprehensive experiments were conducted under a unified evaluation protocol. The analysis aims to quantify classification accuracy, investigate the contribution of architectural components, and examine computational efficiency. Particular emphasis is placed on understanding how systematic design choices influence discriminative embedding quality.

4.1 Dataset

A dedicated dataset was constructed to support fine-grained recognition of residential AC indoor units. After cleaning and verification, the final dataset comprises 720 labeled images corresponding to 216 unique AC manufacturer-model categories, examples of which can be seen in Figure 3. Each class represents a specific indoor unit model, reflecting real-world appliance-level identification scenarios.

Data acquisition followed a two-stage process. First, images were collected in real residential and commercial environments through a custom-developed mobile application. During capture, each image was annotated with the corresponding manufacturer and model metadata to ensure labeling reliability. These field-acquired samples reflect realistic installation conditions, including variations in viewpoint, illumination, background clutter, and partial occlusions.

To increase coverage and include models less frequently encountered in field inspections, the dataset was supplemented with additional images obtained from publicly available online sources, including manufacturer catalogs and retail platforms.[54] All web-sourced images were manually inspected to verify authenticity, remove duplicates, and ensure correspondence with the declared model identifiers. This dual-source strategy balances controlled product imagery with unconstrained real-world photographs, increasing both diversity and ecological validity.

The resulting dataset includes models from major manufacturers represented in the Greek and broader European markets. Visual inspection confirms high inter-class similarity across models, with many units sharing nearly identical casing geometry, vent structures, and color schemes. At the same time, moderate intra-class variability arises from environmental factors such as lighting conditions, wall texture, camera angle, and partial visibility. These characteristics collectively reinforce the formulation of AC recognition as a fine-grained visual classification task.

For model development and evaluation, the dataset was partitioned into training and validation subsets using class-balanced sampling. Depending on the initial sample size of each class, either proportional splitting or a leave-one-out strategy was applied to ensure that every model category remained represented in both subsets whenever feasible. The validation set was held out during optimization and used exclusively for performance assessment, ensuring consistent and unbiased comparison across architectures.

4.2 Experimental setup

To ensure a fair and controlled comparison across architectures, all evaluated models were trained under a unified experimental protocol. The goal was to isolate architectural representational capacity from differences in preprocessing, optimization strategy, or hyperparameter tuning.

All models were initialized with ImageNet-pretrained weights and trained using identical input preprocessing, augmentation strategies, optimizer configuration, and batch size, as described in Section 3.5. During the baseline comparison stage, image resolution was fixed at (224 × 224) pixels and a standard Cross-Entropy loss was employed to provide a common evaluation reference.

Each architecture was trained using the same class-balanced training-validation split described in Section 4.1. Validation performance was computed after each epoch on the hold-out subset, and early stopping was applied based on the validation loss to prevent overfitting. The same evaluation metrics, namely Top-1 accuracy, Top-3 accuracy, precision, recall, and F1-score were recorded for all models to ensure direct comparability.

Beyond predictive performance, computational efficiency was assessed for each architecture using mean training time per epoch, validation time per epoch, and peak GPU memory usage. These metrics provide insight into the trade-offs between accuracy and deployability, which are particularly relevant for energy-informatics applications involving edge or embedded systems.

Following the baseline comparison, the best-performing architecture was selected as the reference model for further refinement and ablation analysis. Subsequent experiments were conducted exclusively on this model to evaluate the impact of backbone selection, activation functions, classification heads, and learning-rate scheduling strategies under controlled conditions.

4.3 Comparison of baseline architectures

The validation performance of the evaluated architectures under the unified experimental protocol is summarized in Table 2 and illustrated in Figure 4. All models were trained using identical preprocessing, optimization settings, and data splits to ensure fair and controlled comparison.

Among the tested architectures, CN-CNN achieved the highest validation accuracy (62.34%) and F1-score (0.5683), outperforming residual, bilinear, ensemble, and hybrid convolution-attention baselines. The consistent gains observed across both precision and recall indicate improved class balance and more stable decision boundaries under fine-grained conditions.

HBP-CNN and ResNet-50 demonstrated moderate performance, suggesting that residual learning and bilinear pooling partially enhance discrimination but remain limited when inter-class differences are subtle and intra-class variability is driven by environmental factors. In contrast, MixDCNN and CoAtNet exhibited significantly lower accuracy and F1-scores. Their performance degradation is likely attributable to over-parameterization relative to dataset scale and reduced generalization under limited training samples.

The superior performance of CN-CNN supports the hypothesis that cross-layer feature aggregation is particularly beneficial in fine-grained appliance recognition. By integrating low-level structural cues with higher-level semantic representations, CN-CNN appears better equipped to capture subtle morphological differences among visually similar AC indoor units.

Beyond predictive performance, computational efficiency was evaluated to assess practical deployment feasibility. As shown in Table 3, clear trade-offs emerge between representational complexity and resource demand. CN-CNN required the longest training time and highest GPU memory usage, reflecting the computational overhead introduced by cross-layer feature fusion. HBP-CNN exhibited moderate computational demand, while ResNet-50 remained comparatively efficient among the higher-performing models.

Table 2. Validation performance of baseline architectures on the AC indoor unit dataset.

Model	Accuracy (%)	Precision	Recall	F1-score
ResNet-50	53.12	0.4783	0.5312	0.4878
HBP-CNN	55.71	0.4950	0.5571	0.5120
CN-CNN	62.34	0.5455	0.6234	0.5683
MixDCNN	30.57	0.2872	0.3057	0.2230
CoAtNet	31.17	0.2690	0.3117	0.2732

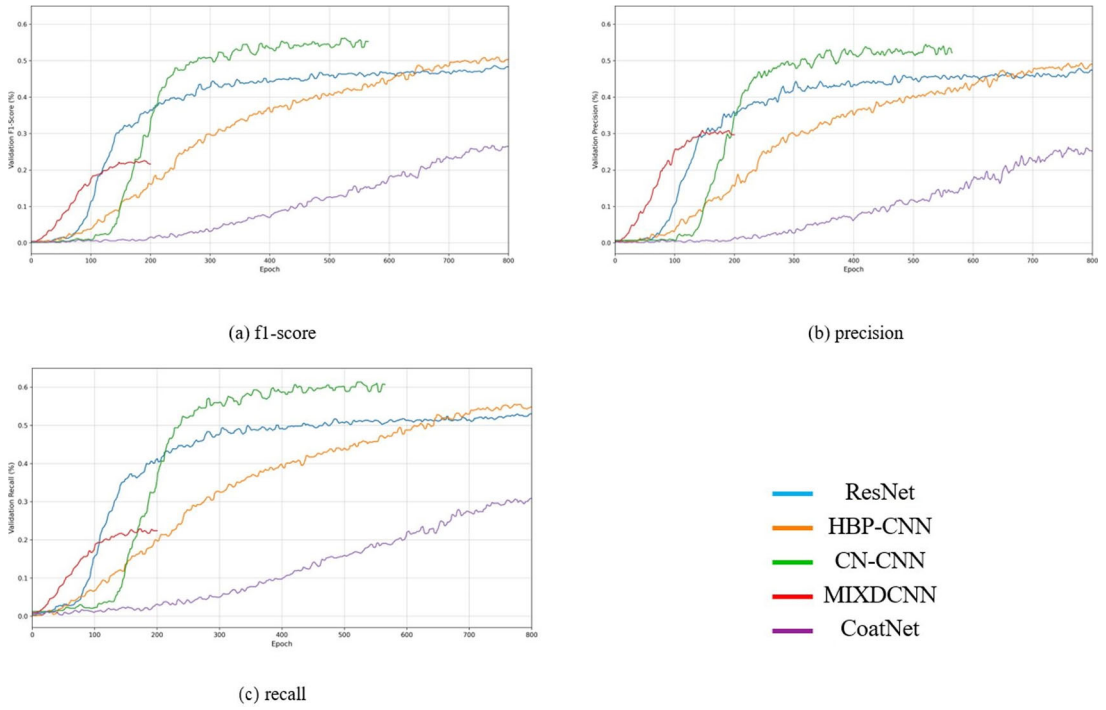


Figure 4. Classification metrics of initial testing of architectures: (a) f1-score, (b) precision, (c) recall.

Table 3. Computational efficiency comparison of baseline architectures.

Model	Train Time/E (s)	Val Time/E (s)	M_{GPU} (MB)
ResNet-50	67.84	15.53	7665
HBP-CNN	80.41	15.62	9324
CN-CNN	150.71	17.05	10970
MixDCNN	139.52	9.72	5539
CoAtNet	56.60	15.43	4802

MixDCNN and CoAtNet demonstrated lower memory requirements but substantially weaker discriminative performance, indicating that computational efficiency alone does not guarantee robustness in fine-grained AC recognition.

Taken together, the results highlight a performance-efficiency trade-off characteristic of deep fine-grained models. Although CN-CNN incurs a higher computational cost, its superior discriminative capability justifies its selection as the reference architecture for further refinement. Given the pronounced visual similarity across AC models and the limited dataset scale, prioritizing hierarchical feature integration is critical to improving generalization performance.

4.4 Ablation study

To quantify the contribution of individual architectural and optimization components to the overall performance of the proposed framework, a series of controlled ablation experiments was conducted on the selected CN-CNN architecture. Following the baseline comparison presented in Section 4.3, CN-CNN was adopted as the reference model due to its superior discriminative performance under fine-grained conditions.

Each experiment isolates a single design variable while keeping all remaining components fixed. This controlled setup ensures that observed performance variations can be directly attributed to the evaluated modification rather than to confounding hyperparameter differences. The ablation analysis examines four principal dimensions: activation functions, classification head formulation, backbone architecture, and learning-rate scheduling strategy.

All experiments employ the same dataset split, preprocessing pipeline, optimizer configuration, and early-stopping criteria described in Sections 3.5 and 4.1. Performance is evaluated using the metrics defined previously, enabling direct comparison across ablation stages. This structured analysis aims to identify the configuration that best balances representational capacity, optimization stability, and computational feasibility for fine-grained AC indoor unit recognition.

4.4.1 Activation functions

The first ablation experiment evaluates the influence of activation function choice on fine-grained AC indoor unit recognition. Activation functions regulate nonlinear feature transformation and gradient flow during training, potentially affecting convergence stability and discriminative representation learning.

Within the selected CN-CNN architecture, the standard Rectified Linear Unit (ReLU) was compared against the Gaussian Error Linear Unit (GELU), while keeping all other architectural components, optimization settings, dataset splits, and preprocessing parameters fixed. This controlled setup ensures that any observed performance variation is attributable solely to the activation function.

As shown in Table 4 and Figure 5, replacing ReLU with GELU yields consistent improvements across all evaluation metrics. Validation accuracy increases by 4.02 percentage points, accompanied by corresponding gains in precision, recall, and F1-score.

The observed improvement suggests that smoother nonlinear activation dynamics contribute positively to discriminative feature formation in fine-grained settings. Unlike ReLU, which enforces hard thresholding of negative responses, GELU introduces a probabilistic weighting of activations, potentially preserving low-magnitude feature signals. Such signals may encode subtle structural variations critical for differentiating visually similar AC models.

Overall, the results indicate that activation function selection exerts a measurable influence on representation quality in fine-grained appliance recognition. Based on these findings, GELU is adopted in subsequent refinement experiments.

4.4.2 Classification head design

This ablation experiment evaluates the influence of classification head formulation on fine-grained AC recognition performance. Three strategies were examined within the refined CN-CNN architecture (with GELU activation): (i) standard linear classification with Cross-Entropy loss, (ii) ArcMargin-based angular supervision, and (iii) CosFace-based cosine-margin supervision. All remaining architectural and optimization parameters were held constant.

To establish a performance baseline, the standard linear head was evaluated against multiple ArcMargin configurations, with the results presented in Figure 6. Across all evaluated metrics, including Top-1 accuracy, Top-3 accuracy, precision, recall, and F1-score, the standard linear head consistently outperformed ArcMargin variants.

The introduction of an additive angular margin did not improve generalization and, in several configurations, resulted in performance degradation. Even when a staged training strategy was applied (Linear → ArcMargin), the model exhibited reduced stability. This behavior suggests that strong angular constraints may not generalize effectively under conditions of limited per-class sample size and relatively small intra-class variance.

Given the limitations observed with ArcMargin, cosine-margin supervision (CosFace) was subsequently evaluated. Unlike ArcMargin, which enforces an additive angular margin, CosFace introduces a cosine margin directly in the normalized embedding space, promoting inter-class separation while maintaining more stable optimization dynamics. Table 5 summarizes the quantitative comparison between the linear and CosFace heads. CosFace improves all key

Table 4. Comparison of validation metrics between ReLU and GELU activations.

Activation	Val Accuracy (%)	Precision	Recall	F1-score
ReLU	62.34	0.5455	0.6234	0.5683
GELU	**66.36**	**0.5911**	**0.6636**	**0.6113**

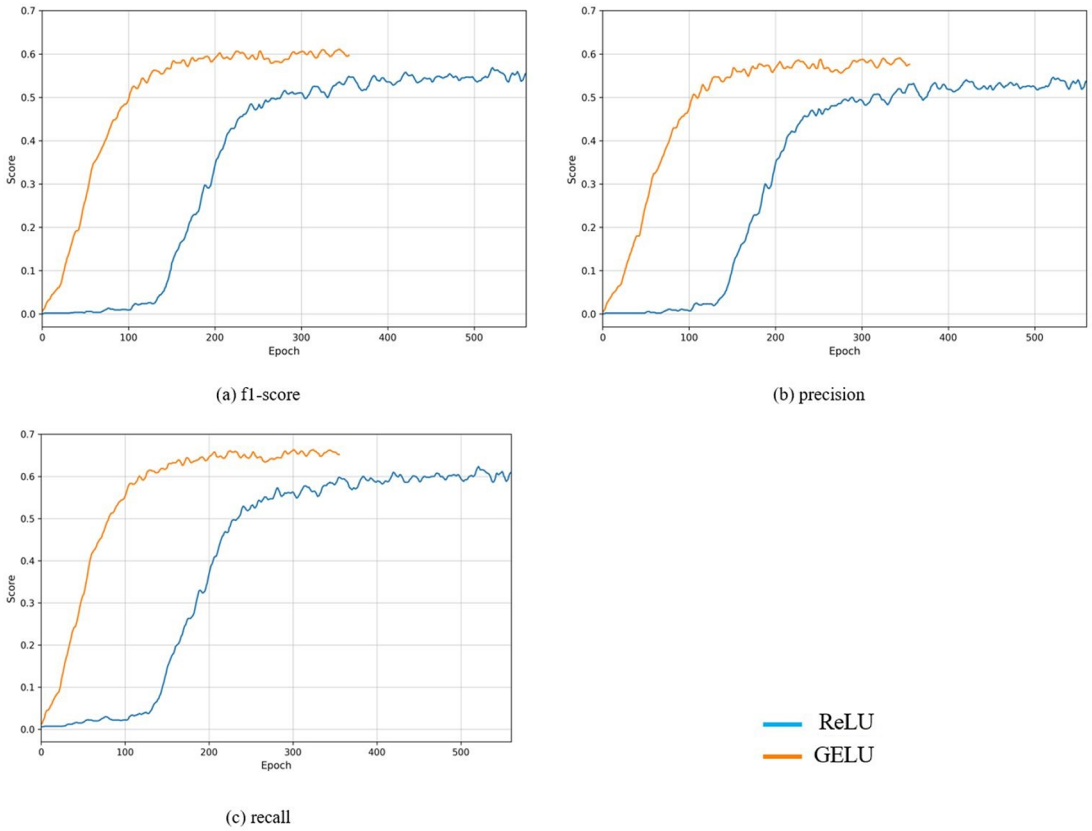


Figure 5. Comparison of validation metrics *between* ReLU and GELU activation functions: (a) f1-score, (b) precision, (c) recall.

metrics: Top-1 accuracy increases to 71.69%, Top-3 accuracy to 84.03%, precision to 0.6444, recall to 0.7169, and F1-score to 0.6621.

Figure 7 further illustrates the consistent performance gains achieved with CosFace across all evaluation metrics.

Compared to ArcMargin, CosFace demonstrates improved stability and faster convergence, achieving higher validation performance within fewer training epochs. The results indicate that cosine-margin supervision is better suited to the dataset scale and fine-grained characteristics of AC indoor unit recognition.

Accordingly, CosFace is adopted as the classification head for all subsequent experiments.

4.4.3 Learning-rate schedulers

This ablation experiment examines the impact of learning-rate scheduling strategy on model convergence and generalization performance. Two widely used scheduling mechanisms were compared within the refined CN-CNN architecture (DenseNet backbone, GELU activation, CosFace head): OneCycleLR and ReduceLROnPlateau. All other training conditions were kept constant.

As shown in Table 6 and Figure 8, OneCycleLR achieves higher Top-1 accuracy and F1-score compared to ReduceLROnPlateau, while maintaining competitive Top-3 performance. Although ReduceLROnPlateau yields a marginally higher Top-3 score, its lower Top-1 and F1 values indicate reduced consistency in correct primary predictions.

Furthermore, OneCycleLR demonstrates faster metric stabilization across epochs, with earlier convergence and reduced oscillation in validation performance. This behavior suggests improved optimization stability under dynamic learning-rate modulation, particularly when combined with margin-based supervision and smooth nonlinear activation.

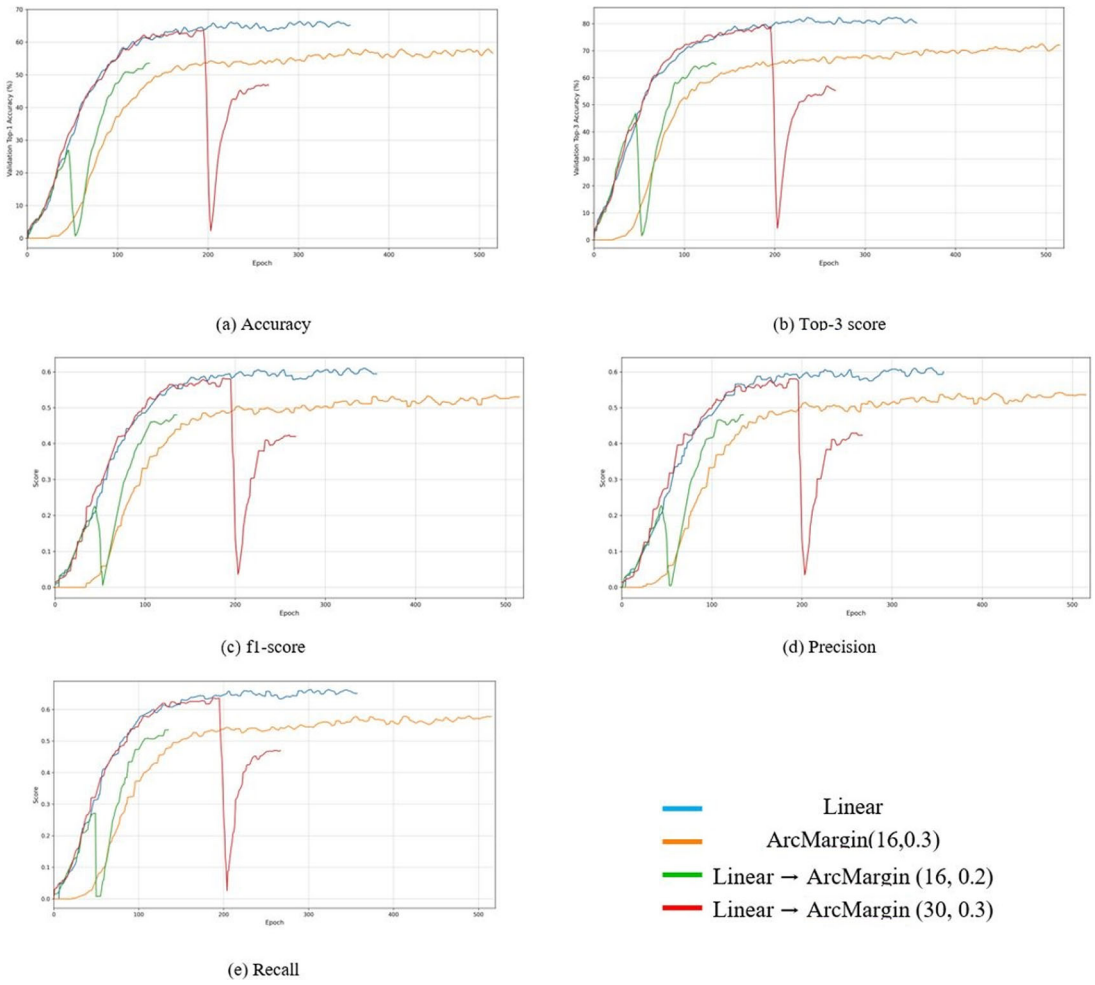


Figure 6. Linear vs ArcMargin head: (a) Accuracy, (b) Top-3 score, (c) f1-score, (d) Precision, (e) Recall.

Table 5. Comparison of Linear and CosFace classification heads on CN-CNN.

Head Type	Val Top-1 (%)	Val Top-3 (%)	Precision	Recall	F1-score
Linear	66.36	82.47	0.5911	0.6636	0.6113
CosFace	**71.69**	**84.03**	**0.6444**	**0.7169**	**0.6621**

Overall, OneCycleLR provides a favorable balance between convergence speed and generalization performance for the refined CN-CNN architecture. Consequently, it is adopted as the default scheduling strategy in the final configuration.

4.4.4 Backbone architecture

The final architectural ablation examines the impact of backbone selection on fine-grained AC indoor unit recognition. The CN-CNN framework was evaluated using two convolutional backbones: ResNet-50 and DenseNet-161. All other components, including GELU activation, CosFace classification head, and OneCycleLR scheduling, were held constant.

Table 7 summarizes the validation performance of both backbones, while Figure 9 provides a visual comparison across evaluation metrics.

DenseNet-161 consistently outperformed ResNet-50 across all evaluation metrics, improving Top-1 accuracy by approximately 2.34% and increasing the F1-score by 0.043. These gains suggest that dense connectivity enhances

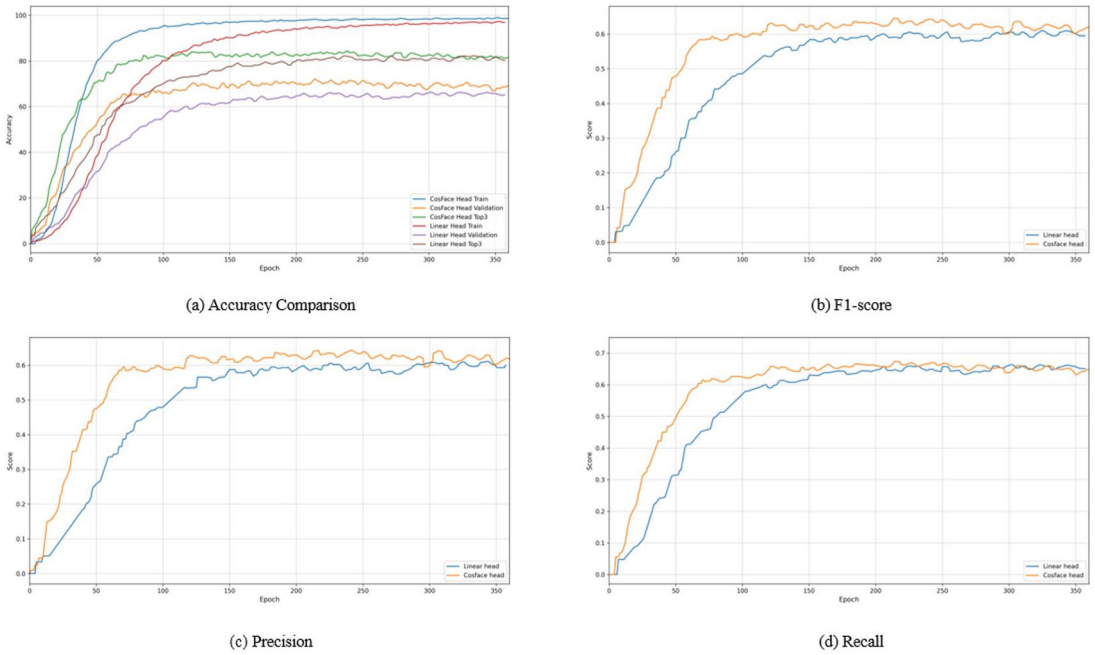


Figure 7. **Linear vs CosFace head: (a) Accuracies, (b) f1-score, (c) Precision, (d) Recall.**

Table 6. **Comparison of learning-rate schedulers within the refined CN-CNN configuration.**

Scheduler	Val Top-1 (%)	Val Top-3 (%)	Precision	Recall	F1-score
OneCycleLR	**71.69**	84.03	**0.6444**	**0.7169**	**0.6621**
ReduceLROnPlateau	67.79	**84.81**	0.6097	0.6779	0.6297

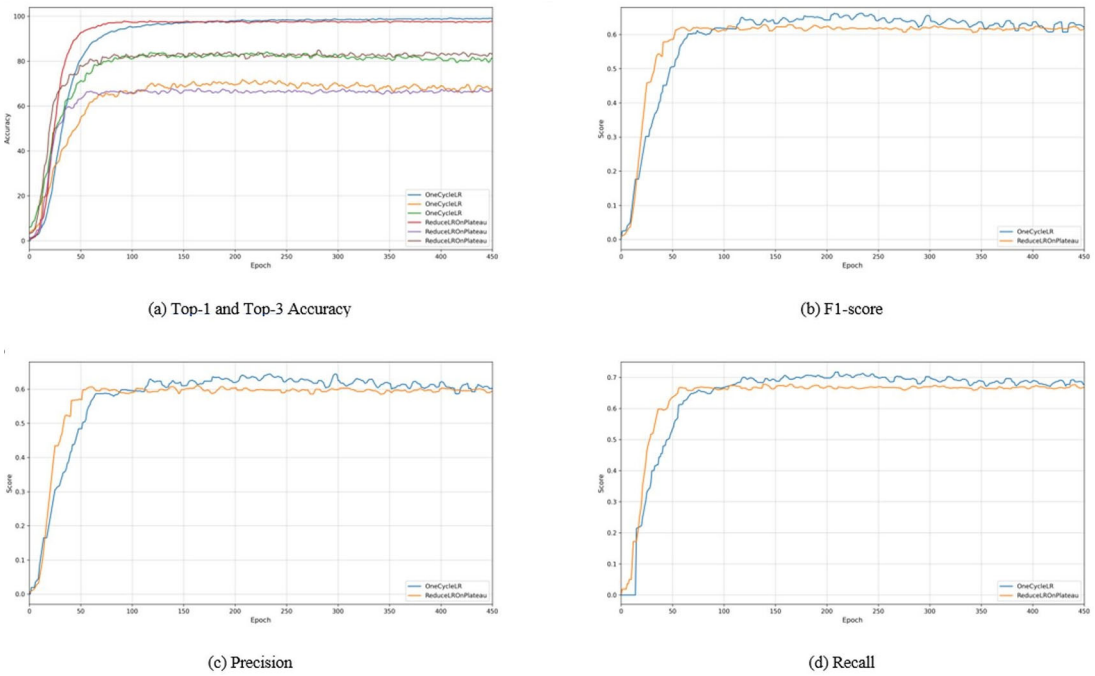


Figure 8. **OneCycleLR vs ReduceLROnPlateau schedulers: (a) Accuracies, (b) f1-score, (c) Precision, (d) Recall.**

feature reuse and multi-scale information propagation, which is particularly beneficial in fine-grained recognition tasks characterized by subtle inter-class differences.

To further assess practical feasibility, computational cost was evaluated for both backbones. The results are summarized in Table 8.

DenseNet-161 required moderately longer training time compared to ResNet-50, reflecting its increased representational complexity. GPU memory usage remained comparable between the two configurations under mixed-precision training. Despite its higher computational demand, DenseNet-161 achieved faster validation inference, likely due to its optimized feature propagation structure.

Overall, DenseNet-161 provides the strongest discriminative performance among the evaluated backbones. Although ResNet-50 remains a viable alternative for deployment scenarios with strict computational constraints, DenseNet-161 was selected as the final backbone for the optimized CN-CNN configuration due to its superior generalization performance in fine-grained AC indoor unit recognition.

Table 7. Comparison of CN-CNN performance using ResNet-50 and DenseNet-161 backbones.

Backbone	Val Top-1 (%)	Val Top-3 (%)	Precision	Recall	F1-score
ResNet-50	71.69	84.03	0.6444	0.7169	0.6621
DenseNet-161	**74.03**	**87.01**	**0.6959**	**0.7403**	**0.7054**

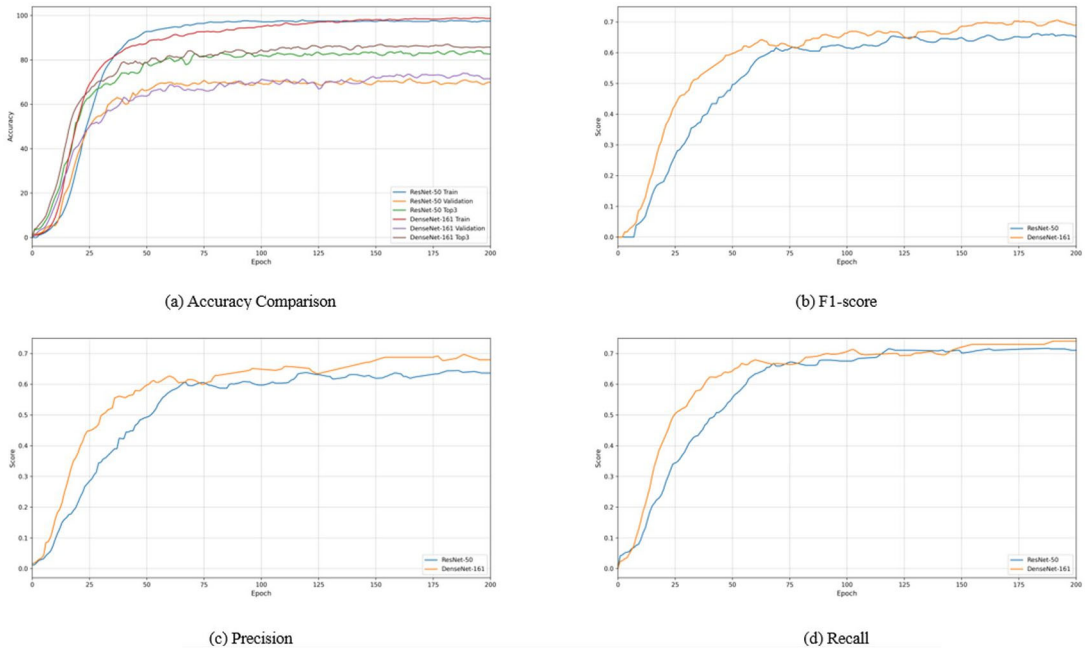


Figure 9. ResNet-50 vs DenseNet-161 backbones: (a) Accuracies, (b) f1-score, (c) Precision, (d) Recall.

Table 8. Computational cost comparison for CN-CNN using different backbones.

Backbone	Mean Training Time (s)	Mean Validation Time (s)	M_GPU (MB)
ResNet-50	112.43	16.19	10970
DenseNet-161	**121.45**	**11.83**	**10885**

4.5 Quantitative summary of ablation results

To systematically evaluate the relative contribution of each architectural refinement, the incremental performance gains across all ablation stages are summarized in Table 9.

Starting from the initial CN-CNN configuration (ReLU, Linear head, ResNet-50, ReduceLROnPlateau), successive refinements produced cumulative improvements in validation Top-1 accuracy from 62.34% to 74.03%, corresponding to a + 11.69 percentage point increase (18.8% relative improvement). The F1-score increased from 0.5683 to 0.7054, yielding a + 0.1371 gain (24.1% relative improvement), indicating substantially improved class-balanced discrimination.

Among the evaluated components, classification head modification (Linear → CosFace) delivered the largest individual Top-1 improvement (+5.33 pp), accounting for approximately 45.6% of the total accuracy gain. Activation refinement (ReLU → GELU) contributed +4.02 pp (34.4% of total gain), while backbone enhancement (ResNet-50 → DenseNet-161) yielded +2.34 pp (20.0%). Learning-rate scheduling primarily improved optimization stability, contributing +3.90 pp relative to ReduceLROnPlateau under equivalent settings.

The results indicate that embedding-space regularization exerts the strongest influence on fine-grained AC discrimination, followed by nonlinear activation smoothing and dense feature reuse. Structural backbone improvements provide complementary but comparatively smaller gains once representation learning is stabilized through activation and margin-based supervision.

From a performance-efficiency perspective, the final configuration (DenseNet-161 + GELU + CosFace + OneCycleLR) achieves the highest discriminative accuracy at the cost of moderately increased computational demand. The observed improvements therefore reflect deliberate prioritization of representational capacity over minimal resource consumption in fine-grained recognition scenarios.

5. Discussion

This section interprets the empirical findings presented in Sections 4.3 and 4.5 in the broader context of fine-grained visual classification and appliance-level asset recognition. While the ablation study quantitatively identified the architectural components that most strongly influence performance, the following discussion examines the underlying reasons for these effects, their practical implications for energy-related applications, and the limitations inherent to the current dataset and modeling framework.

5.1 Interpretation of results

The experimental findings reveal that fine-grained recognition of residential AC indoor units is governed primarily by representation geometry and multi-scale feature integration rather than by architectural depth alone. The superior performance of the CN-CNN architecture in the baseline comparison underscores the importance of cross-layer feature aggregation in scenarios characterized by subtle inter-class differences.

Unlike conventional convolutional networks that rely predominantly on final-layer activations, CN-CNN progressively integrates intermediate representations from multiple depths. This design enables simultaneous exploitation of low-level structural cues-such as vent spacing, panel segmentation, and edge curvature-and higher-level semantic abstractions describing overall unit morphology. In the AC recognition setting, discriminative information is often weak, spatially localized, and distributed across scales. Cross-layer fusion therefore appears particularly well aligned with the structural characteristics of the dataset, explaining its advantage over residual, bilinear, and hybrid convolution-attention baselines.

Beyond the base architecture, the ablation study indicates that embedding-space supervision exerts the strongest influence on performance. The marked improvement obtained with CosFace demonstrates that explicit cosine-margin

Table 9. Incremental performance gains across ablation stages.

Modification	Top-1 Before (%)	Top-1 After (%)	Δ Top-1 (pp)	Δ F1
ReLU → GELU	62.34	66.36	+4.02	+0.0430
Linear → CosFace	66.36	71.69	+5.33	+0.0508
Plateau → OneCycle	67.79	71.69	+3.90	+0.0324
ResNet-50 → DenseNet-161	71.69	74.03	+2.34	+0.0433

regularization enhances inter-class separability in normalized feature space. Fine-grained AC categories frequently exhibit minimal intra-class variation and extremely small inter-class differences; consequently, the geometry of the embedding space becomes critical. CosFace encourages tighter intra-class clustering while preserving optimization stability, resulting in improved balanced classification performance.

In contrast, ArcMargin-based supervision did not generalize effectively under the present dataset constraints. The additive angular margin introduces stronger geometric constraints that may be sensitive to limited per-class sample sizes. When intra-class variability is modest and dataset scale is relatively small, aggressive angular enforcement can destabilize optimization and degrade generalization. The observed underperformance of ArcMargin therefore suggests that margin-based formulations must be carefully matched to dataset scale and distribution characteristics.

Activation refinement further contributed to performance gains. The transition from ReLU to GELU improved discrimination consistency, indicating that smoother nonlinear activation dynamics are advantageous in fine-grained contexts. Hard thresholding of negative activations may suppress weak but informative structural signals. In contrast, GELU preserves low-magnitude responses through probabilistic weighting, facilitating retention of subtle geometric cues essential for distinguishing visually similar AC models.

Backbone enhancement via DenseNet-161 provided complementary improvements. Dense connectivity promotes feature reuse and efficient propagation of information across layers, reinforcing multi-scale representation learning. Once embedding geometry and activation smoothness are optimized, enhanced feature reuse yields additional but comparatively smaller gains. This suggests that backbone selection becomes increasingly impactful after representation-level refinements stabilize discriminative learning.

Finally, learning-rate scheduling influenced convergence dynamics rather than representational structure. The improved stability observed with OneCycleLR indicates that dynamic learning-rate modulation supports effective optimization in the presence of margin-based supervision and smooth nonlinearities. However, its relative contribution remained secondary compared to classification head and activation design.

Collectively, these results suggest that fine-grained AC recognition performance is driven primarily by embedding-space regularization and cross-layer structural feature integration. Architectural depth and scheduling strategies provide supportive but less dominant contributions once representation geometry is properly structured.

5.2 Practical implications

The proposed vision-based AC recognition framework demonstrates practical potential for appliance-level asset identification within residential and commercial building environments. Accurate model-level recognition enables automated inventory generation, lifecycle tracking, maintenance scheduling, and integration with energy management systems. By reducing reliance on manual surveys and metadata-based inventories, computer vision-driven recognition can improve consistency, scalability, and operational efficiency in large building portfolios.

In particular, fine-grained AC model identification can support energy analytics workflows by linking visual inspection data with manufacturer specifications, efficiency ratings, refrigerant type, and operational characteristics. Such integration enables more precise energy performance assessment and retrofits prioritization. For facility managers and energy auditors, automated recognition systems may reduce inspection time while minimizing human labeling errors.

From a deployment perspective, the refined configuration (DenseNet-161 + GELU + CosFace + OneCycleLR) achieves strong discriminative performance under realistic installation conditions. Although this configuration introduces moderate computational overhead, inference time remains compatible with modern GPU-enabled edge devices and workstation-based inspection systems. For resource-constrained scenarios, alternative configurations (e.g., ResNet-50 backbone) provide a viable trade-off between performance and computational demand.

The results further suggest that embedding-space regularization and cross-layer feature aggregation are particularly important for appliance-level recognition tasks characterized by subtle geometric variations. These insights may generalize beyond AC indoor units to other HVAC components or visually similar building assets.

While the proposed framework shows promising applicability, its performance and generalization capability are inherently influenced by dataset composition and scale. These limitations are discussed in the following subsection.

5.3 Limitations

The findings of this study should be interpreted within several methodological and experimental constraints.

First, the dataset scale imposes inherent limitations on statistical robustness. With a relatively small number of samples per class, fine-grained discrimination relies on sparse supervision. Although macro-averaged metrics were employed, per-class variability and confidence intervals were not systematically quantified. Consequently, performance estimates may exhibit sensitivity to data partitioning and initialization conditions.

Second, the dataset construction process introduces structural biases. The combination of web-sourced catalog imagery and field-acquired photographs may result in heterogeneous data distributions. Controlled product images and real-world installation images differ in lighting, perspective, and background complexity. The current evaluation does not explicitly disentangle these domain effects.

Moreover, the experimental protocol follows a closed-set classification paradigm. All evaluation samples belong to categories observed during training. Real-world asset identification systems may encounter previously unseen models or visually similar variants not represented in the training set. The present study does not assess behavior under open-set or unknown-class conditions.

Also, the evaluation focuses on overall classification metrics without explicit analysis of decision consistency across runs or sensitivity to hyperparameter variations. While ablation studies isolate individual architectural components, robustness to stochastic variation was not systematically examined.

Finally, although the refined architecture achieves improved discriminative performance, its internal decision-making mechanisms remain partially opaque. No formal interpretability or feature-attribution analysis was conducted to determine which structural cues most strongly influence classification outcomes.

These considerations define the scope within which the reported improvements should be understood.

5.4 Future directions

Building upon the present findings, several structured research directions can further advance vision-based AC recognition and its integration within broader energy-informatics systems.

A primary extension involves expanding the recognition framework beyond closed-set classification toward open-set and incremental learning scenarios. Real-world asset management systems must accommodate newly introduced AC models and visually similar variants absent from the original training distribution. Incorporating mechanisms for unknown-class detection and adaptive model updating would significantly improve long-term deployment viability.

Robustness under distributional shift constitutes another important direction. Systematic investigation of domain-invariant feature learning, contrastive pretraining, and cross-domain validation protocols could strengthen generalization across heterogeneous imaging conditions, including varying lighting, installation contexts, and device aging effects.

As dataset scale increases, transformer-based and hybrid CNN-Transformer architectures warrant systematic evaluation. Vision Transformers and hierarchical attention models have demonstrated strong performance in large-scale fine-grained benchmarks due to their capacity to capture long-range dependencies and global structural relationships. Assessing their applicability to appliance-level recognition under expanded data regimes may further enhance embedding robustness.

Model efficiency optimization also remains critical for large-scale deployment. Exploring lightweight backbones, structured pruning, quantization, or knowledge distillation strategies could enable comparable discriminative performance under stricter computational constraints, facilitating implementation in edge-based inspection systems or mobile auditing platforms.

In addition, integrating interpretability techniques-such as saliency visualization, embedding-space clustering analysis, and feature attribution mapping- would provide deeper insight into the structural cues driving classification decisions. Enhanced interpretability may improve system transparency and support practical validation by energy auditors or facility managers.

Finally, a broader research direction involves coupling visual asset recognition with agnostic non-intrusive load monitoring (NILM) frameworks. Linking appliance-level visual identification with energy disaggregation models could enable hybrid systems that combine visual inspection and electrical consumption signatures. Extending the framework to additional household devices beyond AC units may support comprehensive, vision-assisted energy disaggregation and automated building inventory systems.

Collectively, these directions aim to enhance scalability, robustness, interpretability, and cross-domain integration of fine-grained visual asset recognition within next-generation smart building ecosystems.

6. Conclusion

This work presented an end-to-end vision-based framework for fine-grained recognition of residential AC indoor units, addressing the challenge of appliance-level asset identification in energy and building management contexts. Unlike coarse object detection tasks, model-level AC recognition requires discrimination among visually similar categories with subtle geometric differences, positioning the problem within the domain of fine-grained visual classification.

A new curated dataset comprising 720 images across 216 AC model categories was constructed by integrating field-acquired and web-sourced imagery. To the best of our knowledge, this is the first documented dataset specifically designed for model-level visual recognition of residential AC indoor units, establishing a foundation for appliance-level computer vision research.

Beyond dataset creation, the study introduced a structured methodological pipeline encompassing baseline benchmarking, controlled architectural ablation, and task-specific optimization. Through systematic analysis, we demonstrated that cross-layer feature aggregation, cosine-margin embedding supervision, smooth nonlinear activation, and dynamic learning-rate scheduling jointly drive performance improvements in fine-grained AC recognition. The final optimized configuration-CN-CNN with DenseNet-161 backbone, GELU activation, CosFace classification head, and OneCycleLR scheduler-achieved a validation Top-1 accuracy of 74.03% and an F1-score of 0.7054, corresponding to a cumulative improvement of +11.69 percentage points over the initial baseline.

The ablation results highlight that embedding-space geometry and multi-scale feature integration exert a stronger influence on discriminative performance than architectural depth alone. These findings provide empirical evidence that a task-adapted configuration of existing architectural components can significantly enhance appliance-level recognition without introducing new structural modules.

Overall, this work establishes a principled end-to-end framework for fine-grained visual asset identification and demonstrates the feasibility of integrating deep learning-based recognition into automated building inventory, energy auditing, and smart energy management systems.

Ethics and consent

Formal ethical approval was not required for this study. Verbal informed consent was obtained from all individuals who voluntarily provided photographs of air-conditioning units located in premises under their control. Verbal rather than written consent was chosen because participation involved only a brief, one-time provision of appliance photographs and consent could be obtained directly and promptly at the time of contribution. This approach was more practical for the distributed image-collection process and avoided collecting additional personal identifiers, such as names and signatures, that were unnecessary for the study. Before providing the photographs, contributors were informed that the images would be used for research, model training, evaluation, and publication of the study results. The contributors did not provide consent for the unrestricted public release of the complete raw image dataset.

List of Figures

1: Framework overview PAGEREF _Toc238033465 \h 10

2: Architecture Principles PAGEREF _Toc238033466 \h 13

3: De-identified representative samples from the AC indoor unit dataset PAGEREF _Toc238033467 \h 15

4: Classification metrics of initial testing of architectures: (a) f1-score, (b) precision, (c) recall PAGEREF _Toc238033468 \h 17

5: Comparison of validation metrics between ReLU and GELU activation functions: (a) f1-score, (b) precision, (c) recall PAGEREF _Toc238033469 \h 19

6: Linear vs ArcMargin head: (a) Accuracy, (b) Top-3 score, (c) f1-score, (d) Precision, (e) Recall PAGEREF _Toc238033470 \h 20

7: Linear vs CosFace head: (a) Accuracies, (b) f1-score, (c) Precision, (d) Recall PAGEREF _Toc238033471 \h 21

8: OneCycleLR vs ReduceLROnPlateau schedulers: (a) Accuracies, (b) f1-score, (c) Precision, (d) Recall PAGEREF _Toc238033472 \h 22

9: ResNet-50 vs DenseNet-161 backbones: (a) Accuracies, (b) f1-score, (c) Precision, (d) Recall PAGEREF _Toc238033473 \h 23

Data availability

Underlying and extended data

Zenodo: Supporting Numerical Data for Fine-Grained Visual Recognition of Residential Air Conditioner Units:

https://doi.org/10.5281/zenodo.21413007.[55]

The repository contains the aggregate numerical values underlying the tables and figures reported in this study, including classification metrics, computational-efficiency measurements, ablation-study comparisons, and the Python scripts used to reproduce the corresponding plots. The deposited materials are available under the terms of the Creative Commons Attribution 4.0 International license (CC BY 4.0).

Restricted image data

The complete raw image dataset is not publicly available. The field-acquired photographs were voluntarily provided by individuals and were captured in private premises. Although the photographs primarily depict air-conditioning units, some images may contain contextual details relating to the premises or their occupants. The contributors consented to the use of the images for research, model training, evaluation, and publication of the study results, but not to their unrestricted public dissemination under an open license.

Access to eligible field-acquired images may be requested from the corresponding author at rdaskos@epu.ntua.gr. Requests will be considered on a case-by-case basis and access will be provided only where the proposed use is compatible with the consent obtained, applicable data-protection requirements, and any necessary data-use and confidentiality agreement.

The dataset also includes images obtained from publicly accessible manufacturer and retail websites. These images are subject to third-party copyright and cannot be redistributed by the authors. Researchers may obtain comparable images directly from the original online sources described in the manuscript.

References

1. Strbac G: **Demand side management: Benefits and challenges.** *Energy Policy.* 2008; **36**(12): 4419–4426.
Publisher Full Text

2. Siano P: **Demand response and smart grids—a survey.** *Renew. Sust. Energ. Rev.* 2014; **30**: 461–478.
Publisher Full Text

3. Papias I, Michalakopoulos V, Sarmas E, *et al.*: **Aggregated flexibility dynamic forecasting of residential energy prosumers for DR programs.** *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA).* 2024; 1–9.
Publisher Full Text

4. Pérez-Lombard L, Ortiz J, Pout C: **A review on buildings energy consumption information.** *Energ. Buildings.* 2008; **40**(3): 394–398.
Publisher Full Text

5. International Energy Agency: *World Energy Outlook 2019.* Paris: IEA; 2019.
Publisher Full Text | **Reference Source**

6. Papias I, Michalakopoulos V, Sarmas E, *et al.*: **A data-driven framework for estimating residential energy flexibility for aggregated demand-side management.** *Sustainable Energy, Grids and Networks.* 2025b; **43**: 101783.
Publisher Full Text

7. Papias I, Sarantinopoulos E, Michalakopoulos V, *et al.*: **FLEXiT: A platform for estimating residential energy flexibility for aggregated & disaggregated demand-side management.** *SoftwareX.* 2025d; **32**: 102433.
Publisher Full Text

8. International Energy Agency: *The Future of Cooling: Opportunities for Energy Efficient Air Conditioning.* Paris: IEA; 2018.
Publisher Full Text | **Reference Source**

9. Arteconi A, Hewitt NJ, Polonara F: **Domestic demand-side management (dsm): Role of heat pumps and thermal energy storage (tes) systems.** *Appl. Therm. Eng.* 2013; **51**(1–2): 155–165.
Publisher Full Text

10. Hwang H, Yoon A, Yoon Y, *et al.*: **Demand response of HVAC systems for hosting capacity improvement in distribution networks: A comprehensive review and case study.** *Renew. Sust. Energ. Rev.* 2023; **187**: 113751.
Publisher Full Text

11. Hart GW: **Nonintrusive appliance load monitoring.** *Proc. IEEE.* 1992; **80**(12): 1870–1891.
Publisher Full Text

12. Zoha A, Gluhak A, Imran MA, *et al.*: **Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey.** *Sensors.* 2012; **12**(12): 16838–16866.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. Serepas F, Papias I, Christakis K, *et al.*: **Lightweight embedded IoT gateway for smart homes based on an ESP32 microcontroller.** *Computers.* 2025; **14**(9): 391.
Publisher Full Text

14. Serepas F, Papias I, Bellos N, *et al.*: **A comprehensive approach to real-time and batch processing for energy-efficient IoT homes: Leveraging lambda architecture and data lakes.** *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA).* 2024; 1–6.
Publisher Full Text

15. Kolter JZ, Johnson MJ: **REDD: A public data set for energy disaggregation research.** *Proceedings of the Workshop on Data Mining Applications in Sustainability (SustKDD).* 2011; 59–62.
Reference Source

16. Parson O, Ghosh S, Weal M, *et al.*: **Non-intrusive load monitoring using prior models of general appliance types.** *In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12).* 2012; 356–362.
Publisher Full Text

17. Papias I, Michalakopoulos V, Karoutsou D, *et al.*: **Non-intrusive load monitoring using cluster-optimized denoising autoencoders.** *2025 16th International Conference on Information, Intelligence, Systems & Applications (IISA).* 2025a; 1–8.
Publisher Full Text

18. Papias I, Papapetrou L, Michalakopoulos V, *et al.*: **Denoising autoencoder for appliance-specific energy disaggregation in residential settings.** *2025 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv).* 2025c; 384–389.
Publisher Full Text

19. Deng J, Dong W, Socher R, *et al.*: **ImageNet: A large-scale hierarchical image database.** *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2009; 248–255.
Publisher Full Text

20. Krizhevsky A, Sutskever I, Hinton GE: **ImageNet classification with deep convolutional neural networks.** *Advances in Neural Information Processing Systems 25 (NeurIPS 2012).* 2012; **25**: 1097–1105.
Reference Source .

21. He K, Zhang X, Ren S, *et al.*: **Deep residual learning for image recognition.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016; 770–778.
Publisher Full Text

22. Wah C, Branson S, Welinder P, *et al.*: *The caltech-ucsd birds-200-2011 dataset Technical Report CNS-TR-2011-001.* California Institute of Technology; 2011.
Reference Source

23. Krause J, Stark M, Deng J, *et al.*: **3d object representations for fine-grained categorization.** *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW).* 2013; 554–561.
Publisher Full Text

24. Zheng H, Fu J, Zha Z-J, *et al.*: **Deep learning for fine-grained image analysis: A survey.** 2019. arXiv preprint arXiv:1907.03069.
Reference Source

25. Jiang Z, Messner JI: **Computer vision applications in construction and asset management phases: A literature review.** *Journal of Information Technology in Construction.* 2023; **28**: 176–199.
Publisher Full Text

26. Katipamula S, Brambley MR: **Methods for fault detection, diagnostics, and prognostics for building systems—a review, part I.** *HVAC&R Research.* 2005; **11**(1): 3–25.
Publisher Full Text

27. Kim W, Katipamula S: **A review of fault detection and diagnostics methods for building systems.** *Sci. Technol. Built Environ.* 2018; **24**(1): 3–21.
Publisher Full Text

28. Beghi A, Brignoli R, Cecchinato L, *et al.*: **Data-driven fault detection and diagnosis for HVAC water chillers.** *Control. Eng. Pract.* 2016; **53**: 79–91.
Publisher Full Text

29. Gupta S, Girshick R, Arbeláez P, *et al.*: **Learning rich features from RGB-D images for object detection and segmentation.** *Computer Vision–ECCV 2014, volume 8695 of Lecture Notes in Computer Science.* 2014; 345–360.
Publisher Full Text

30. Zhang N, Donahue J, Girshick R, *et al.*: **Part-based R-CNNs for fine-grained category detection.** *Computer Vision–ECCV 2014, volume 8689 of Lecture Notes in Computer Science.* 2014; 834–849.
Publisher Full Text

31. Krause J, Jin H, Yang J, *et al.*: **Fine-grained recognition without part annotations.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015; 5546–5555.
Publisher Full Text

32. Lin T-Y, RoyChowdhury A, Maji S: **Bilinear CNN models for fine-grained visual recognition.** *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 2015; 1449–1457.
Publisher Full Text

33. Fu J, Zheng H, Mei T: **Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017; 4438–4446.
Publisher Full Text

34. Wei X-S, Song Y-Z, Mac Aodha O, *et al.*: **Fine-grained image analysis with deep learning: A survey.** *IEEE Trans. Pattern Anal. Mach. Intell.* 2022; **44**(12): 8927–8948.
PubMed Abstract | **Publisher Full Text**

35. Gao Y, Beijbom O, Zhang N, *et al.*: **Compact bilinear pooling.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016; 317–326.
Publisher Full Text

36. Kong S, Fowlkes CC: **Low-rank bilinear pooling for fine-grained classification.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017; 7025–7034.
Publisher Full Text

37. Yu C, Zhao X, Zheng Q, *et al.*: **Hierarchical bilinear pooling for fine-grained visual recognition.** *Computer Vision–ECCV 2018, volume 11220 of Lecture Notes in Computer Science.* 2018; 595–610.
Publisher Full Text

38. Ge Z, Bewley A, McCool C, *et al.*: **Fine-grained classification via mixture of deep convolutional neural networks.** *2016 IEEE Winter Conference on Applications of Computer Vision (WACV).* 2016; 1–6.
Publisher Full Text

39. Zhao B, Wu X, Feng J, *et al.*: **Diversified visual attention networks for fine-grained object classification.** *IEEE Trans. Multimed.* 2017; **19**(6): 1245–1256.
Publisher Full Text

40. Zheng H, Fu J, Mei T, *et al.*: **Learning multi-attention convolutional neural network for fine-grained image recognition.** *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 2017; 5209–5217.
Publisher Full Text

41. Zhu H, Ke W, Li D, *et al.*: **Dual cross-attention learning for fine-grained visual categorization and object re-identification.** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2022; 4692–4702.
Publisher Full Text

42. Guo C, Xie J, Liang K, *et al.*: **Cross-layer navigation convolutional neural network for fine-grained visual classification.** *Proceedings of the 3rd ACM International Conference on Multimedia in Asia.* Association for Computing Machinery; 2021; 1–5.
Publisher Full Text

43. Dai Z, Liu H, Le QV, *et al.*: **CoAtNet: Marrying convolution and attention for all data sizes.** *Advances in Neural Information Processing Systems.* 2021; **34**: 3965–3977.
Reference Source

44. Goodfellow I, Bengio Y, Courville A: *Deep Learning.* MIT Press; 2016.
Reference Source

45. Wang H, Wang Y, Zhou Z, *et al.*: **CosFace: Large margin cosine loss for deep face recognition.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2018; 5265–5274.
Publisher Full Text

46. Cui S, Hui B: **Dual-dependency attention transformer for fine-grained visual classification.** *Sensors.* 2024; **24**(7): 2337.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

47. Nair V, Hinton GE: **Rectified linear units improve restricted boltzmann machines.** *Proceedings of the 27th International Conference on Machine Learning (ICML).* 2010; 807–814.
Publisher Full Text

48. Hendrycks D, Gimpel K: **Gaussian error linear units (GELUs).** 2016. arXiv preprint arXiv:1606.08415.
Reference Source .

49. Kingma DP, Ba J: **Adam: A method for stochastic optimization.** *Proceedings of the 3rd International Conference on Learning Representations (ICLR).* 2015.
Reference Source

50. Smith LN, Topin N: **Super-convergence: Very fast training of neural networks using large learning rates.** *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, volume 11006 of Proceedings of SPIE.* 2019; 369–386.
Publisher Full Text

51. Paszke A, Gross S, Massa F, *et al.*: **PyTorch: An imperative style, high-performance deep learning library.** *Advances in Neural Information Processing Systems.* 2019; **32**: 8024–8035.
Reference Source

52. Shorten C, Khoshgoftaar TM: **A survey on image data augmentation for deep learning.** *J. Big Data.* 2019; **6**: 60.
Publisher Full Text

53. Micikevicius P, Narang S, Alben J, *et al.*: **Mixed precision training.** *Proceedings of the International Conference on Learning Representations (ICLR).* 2018.
Reference Source

54. Skroutz online marketplace:
Reference Source

55. Daskos R, Dimitropoulos N, Papias I, *et al.*: **Supporting numerical data for fine-grained visual recognition of residential air conditioner units.** *Zenodo.* 2026.
Publisher Full Text