SEMANTIC-GUIDED DEEP LEARNING FRAMEWORKS FOR SCENE RECOGNITION: A COMPARATIVE STUDY OF CNN AND TRANSFORMER MODELS

A Thesis Submitted
in Partial Fulfillment of the Requirements for the Award of the Degree of

MASTER OF TECHNOLOGY

in

Information Technology

by

Muheet Alam
(24/ISY/09)

Under the Supervision of

Dr. Seba Susan
Professor, Department of Information Technology
Delhi Technological University

To the
Department Of Information Technology
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

May, 2026

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, **Dr. Seba Susan (Professor, Department of Information Technology)**, for her constant guidance, valuable suggestions, and support throughout this research work. Her insights and encouragement played an important role in shaping the direction of this study.

I am also thankful to **Prof. Dinesh Kumar Vishwakarma (Head, Department of Information Technology)** and the Department of Information Technology, Delhi Technological University, for providing the necessary academic environment, resources, and infrastructure required for carrying out this work.

I would also like to thank my peers and colleagues for the discussions and exchanges of ideas that contributed to this research in many ways.

I gratefully acknowledge the contributions of the researchers and authors whose work has been referenced in this thesis. Their studies provided important foundations and motivation for this work.

Finally, I express my heartfelt gratitude to my family and friends for their continuous encouragement, patience, and support throughout this journey.

Muheet Alam

Muheet Alam
24/ISY/09
M.Tech(Information Technology)
Department of Information Technology
Delhi Technological University

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Muheet Alam (24/ISY/09)**, student of **M.Tech. (Information Technology)**, hereby certify that the work presented in the thesis entitled **"Semantic-Guided Deep Learning Frameworks for Scene Recognition: A Comparative Study of CNN and Transformer Models"**, submitted in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in the Department of Information Technology, Delhi Technological University, is an authentic record of my own work carried out during the period from **August 2024 to May 2026** under the supervision of **Dr. Seba Susan, Professor, Department of Information Technology, Delhi Technological University**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Muheet Alam

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Seba 29/5/26

Signature of Supervisor

- Not applicable.

Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that **Muheet Alam (24/ISY/09)** has carried out their research work presented in this thesis entitled entitled **"Semantic-Guided Deep Learning Frameworks for Scene Recognition: A Comparative Study of CNN and Transformer Models"** for the award of **Master of Technology** from the Department of Information Technology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other Institution.

29/5/26

Signature

Dr. Seba Susan

Professor, Department of Information Technology

Delhi Technological University

Main Bawana Road, Delhi-110042

Date: 29-05-2026

Place: Delhi Technological University, Delhi

Semantic-Guided Deep Learning Frameworks for Scene Recognition: A Comparative Study of CNN and Transformer Models

[Muheet Alam, 24/ISY/09]

ABSTRACT

Indoor scene recognition remains a challenging problem in computer vision due to large intra-class variation, strong inter-class similarity, and the complex contextual relationships that exist between objects and spatial layouts within indoor environments. Unlike object recognition, scene understanding requires the model to interpret not only the presence of semantic entities but also their spatial organization and contextual interactions. Conventional visual recognition approaches based solely on appearance features often struggle to capture these higher-level semantic relationships, particularly in scenes where multiple categories share similar visual structures. Semantic guidance has therefore emerged as an effective strategy for improving scene understanding by incorporating object-level contextual information into the recognition process.

This thesis investigates how semantic supervision interacts with different deep neural representation architectures for indoor scene recognition. Rather than focusing solely on improving classification accuracy through larger models or architectural complexity, the study examines how the underlying representation structure of a backbone influences the effectiveness of semantic-guided feature learning. The work is structured as a progressive investigation across convolutional and transformer-based architectures under a consistent semantic-aware learning framework.

The first phase of the study explores semantic-guided scene recognition using convolutional neural networks. A dual-branch framework consisting of an RGB branch and a semantic branch is employed, where semantic features derived from segmentation maps are integrated with visual representations through attention-based fusion. Within this framework, the effect of backbone architecture is analyzed by comparing ResNet-50 and ResNeXt-50 (32×4d) under identical training and fusion conditions. Experimental observations show that ResNeXt produces stronger scene representations and achieves improved recognition performance on the MIT Indoor-67 dataset. The results suggest that aggregated residual transformations and increased representational diversity enable more effective semantic-guided feature interaction than standard residual learning.

Building upon these observations, the second phase extends the investigation to transformer-based architectures in order to analyze how different representation formats respond to semantic supervision. The study evaluates Vision Transformers and hierarchical Swin Transformers within a representation-aligned semantic learning framework. Since transformer architectures organize visual information differently,

semantic representations are adapted to match the native structure of each backbone. Semantic maps are converted into token representations for Vision Transformers to enable token-level cross-attention, while hierarchical spatial semantic features are used for Swin Transformers to preserve locality and spatial alignment during fusion.

Experimental results indicate that hierarchical transformer representations achieve more effective semantic-guided scene understanding than token-only representations. In particular, Swin-Tiny demonstrates stronger performance and more stable semantic interaction behavior compared to ViT-based models despite lower model complexity.

Collectively, the findings of this thesis suggest that the effectiveness of semantic-aware scene recognition depends not only on the availability of semantic information, but also on how naturally the representation structure of the architecture supports semantic integration. Architectures that preserve spatial hierarchy and contextual locality appear to align more effectively with semantic scene cues than architectures relying purely on global token interactions. The study further highlights the importance of representation-aware semantic encoding when designing multimodal scene understanding systems.

Overall, this thesis presents a structured empirical investigation into semantic-guided representation learning across modern deep neural architectures for indoor scene recognition. The work establishes that semantic supervision becomes more effective when aligned with the native representation structure of the underlying backbone, and it provides insights that may guide future research in semantic-aware visual representation learning, multimodal scene understanding, and architecture-aware fusion design.

LIST OF PUBLICATIONS

[P1]. Muheet Alam, and Seba Susan. "ResNeXt-Based Multimodal Scene Recognition via Semantic-Guided Attention Fusion." Accepted as a full paper for Oral Presentation at the *2026 Kalinga Conference on Communication & Computing (KalingaConf 2026)*, with publication in IEEE Xplore and Scopus indexing.

[P2]. Muheet Alam, and Seba Susan. "Representation-Aligned Semantic Guidance in Transformer-Based Indoor Scene Recognition." Under review at the *2026 IEEE International Conference on Sustainable Technologies for Smart Development Goals (ICSTSDG 2026).*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. **CNN:** Convolutional Neural Network

2. CFA: Cross Feature Aggregation

3. CPU: Central Processing Unit

4. GPU: Graphics Processing Unit

5. HOG: Histogram of Oriented Gradients

6. SIFT: Scale-Invariant Feature Transform

7. SOSF: Semantic Object Spatial Fusion

8. SWIN: Shifted Window Transformer

9. ViT: Vision Transformer

10. ResNet: Residual Network

11. ViT-B/16: Vision Transformer Base with 16×16 Patch Size

12. AdamW: Adaptive Moment Estimation with Weight Decay

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Scene recognition is a fundamental problem in computer vision that focuses on understanding the overall environment represented in an image. Unlike object recognition, which aims to identify individual entities in isolation, scene recognition requires the interpretation of multiple visual elements together with their contextual and spatial relationships. A scene is often characterized not only by the objects present within it, but also by how those objects are arranged and interact with the surrounding environment. This broader understanding of visual context makes scene recognition an important component in intelligent systems that rely on environmental awareness for decision-making and interaction [1].

The ability to recognize scenes accurately plays a significant role in a wide range of real-world applications, including robotics, autonomous navigation, assistive systems, surveillance, multimedia retrieval, and human–computer interaction. In indoor environments particularly, contextual understanding becomes essential because many locations share visually similar structures while serving entirely different functional purposes. For example, a bookstore and a library may contain overlapping object categories such as shelves and books, yet differ in layout, arrangement, and contextual organization. Similarly, waiting rooms, classrooms, and auditoriums may exhibit comparable seating patterns despite belonging to distinct scene categories. These characteristics make indoor scene recognition substantially more challenging than conventional object-centric visual classification tasks.

One of the primary difficulties in indoor scene recognition arises from large intra-class variation and strong inter-class similarity. Images belonging to the same scene category may differ significantly in illumination, viewpoint, scale, object arrangement, and background clutter. At the same time, different scene categories frequently contain common objects and overlapping structural patterns. As a result, relying solely on isolated object detection or low-level appearance features is often insufficient for robust scene understanding. Effective recognition requires learning representations that can capture both semantic context and spatial organization simultaneously.

Traditional scene recognition approaches relied heavily on handcrafted descriptors designed to capture texture, gradients, color distributions, and local structural patterns [2–4]. Although these methods provided reasonable performance on simpler datasets, they struggled to generalize to complex environments containing rich contextual dependencies and large visual variability. The emergence of deep learning, particularly convolutional neural networks (CNNs), significantly improved scene recognition by enabling models to learn hierarchical feature representations directly from large-scale image data. Deep convolutional architectures demonstrated strong capability in capturing progressively

abstract visual patterns, ranging from low-level edges and textures to higher-level semantic structures [1].

Despite these advances, appearance-driven deep representations still face limitations in complex indoor environments. Visual features extracted from RGB images may not always encode the semantic relationships necessary for distinguishing scenes with similar layouts or shared object distributions. In many cases, scene understanding depends not only on recognizing visible objects but also on interpreting how those objects contribute collectively to the functional identity of the environment. For instance, the presence of a bed may suggest a bedroom, but its arrangement relative to surrounding objects and spatial structure often provides equally important contextual information. These observations have motivated increasing interest in semantic-aware scene recognition approaches that incorporate higher-level semantic cues into the recognition process [5].

Semantic guidance introduces additional contextual information by explicitly modeling object-level semantics within a scene. Instead of relying entirely on appearance-based representations, semantic-aware frameworks integrate auxiliary information such as segmentation maps or object-level annotations to guide feature learning. Such semantic representations provide structured contextual cues regarding object presence, spatial arrangement, and scene composition, which can complement conventional visual features. Prior studies have shown that semantic information can improve robustness in indoor scene recognition, particularly in environments where visual ambiguity and contextual overlap are common [5– 7] .

More recently, transformer-based visual architectures have emerged as powerful alternatives for image recognition tasks by modeling long-range contextual dependencies through self-attention mechanisms [6], [8]. Vision Transformers process images by converting image patches into sequential embedding representations and learning contextual relationships across these embeddings using attention-based interaction [8]. Hierarchical transformer architectures such as the Swin Transformer further extend this idea by combining attention-based contextual learning with structured multi-stage spatial feature hierarchies [9].

Token-oriented transformer architectures primarily model contextual relationships through attention-based interaction across distributed patch embeddings, whereas convolutional frameworks preserve contextual continuity through hierarchical spatial feature organization. Although semantic-aware scene recognition has gained considerable attention in recent years, much of the existing research primarily focuses on improving classification performance through stronger architectures or more sophisticated fusion strategies [5–7]. Comparatively less attention has been given to understanding how different representation paradigms respond to semantic supervision under consistent learning conditions. In particular, the interaction between semantic guidance and representation structure across convolutional and transformer-based architectures remains insufficiently explored.

Motivated by these observations, this thesis investigates semantic-guided scene recognition from a representation-learning perspective. Rather than treating semantic supervision merely as an auxiliary source of information, the study examines how different deep neural architectures utilize semantic context based on their inherent representation characteristics. The work progressively explores convolutional and transformer-based frameworks within a unified semantic-aware learning setting in order to analyze how architectural design influences semantic-guided feature interaction and scene understanding.

1.2 PROBLEM STATEMENT

Indoor scene recognition continues to remain a challenging visual understanding problem despite substantial progress in deep learning-based recognition systems. Unlike object-centric classification tasks, indoor scene understanding requires the model to capture complex contextual relationships among multiple visual entities while simultaneously interpreting global spatial structure. Scenes are often characterized not only by the presence of objects, but also by their arrangement, interaction, and functional association within the environment. As a result, visually similar scenes may belong to different semantic categories, while images from the same category can exhibit significant variations in appearance, illumination, layout, and viewpoint.

Deep convolutional neural networks have significantly improved scene recognition by learning hierarchical visual representations from large-scale image datasets [5,10]. More recently, transformer-based architectures have demonstrated strong capability in modeling long-range visual dependencies through self-attention mechanisms. At the same time, several semantic-aware approaches have shown that integrating semantic information, such as object-level context or segmentation-based representations, can further improve scene understanding by providing complementary contextual cues beyond raw visual appearance.

However, existing semantic-guided scene recognition methods primarily focus on improving classification performance through stronger architectures, larger models, or increasingly sophisticated fusion mechanisms. In many cases, semantic information is incorporated as an auxiliary modality without sufficiently analyzing how the internal representation structure of the backbone architecture influences semantic-guided learning. Consequently, the interaction between semantic supervision and representation format remains comparatively underexplored [5– 7].

This limitation becomes particularly important when considering the representational differences between modern deep architectures. Convolutional neural networks learn hierarchical spatial representations that preserve locality and contextual continuity through successive convolutional operations. In contrast, Vision Transformers organize images as sequences of tokens and rely on global self-attention to model relationships across image regions [8]. Hierarchical transformers such as the Swin Transformer combine attention-based learning with progressively aggregated spatial feature representations [9]. Since semantic information itself is inherently structured and spatially contextual, different representation paradigms may interact with semantic guidance in fundamentally different ways.

Another important issue is that several existing approaches apply similar semantic fusion strategies across different backbone architectures without sufficiently considering whether the semantic representation format aligns naturally with the internal representation structure of the underlying model. Therefore, a systematic investigation is required to analyze how semantic supervision behaves across different deep representation architectures under a controlled learning framework. Understanding this interaction is important not only for improving scene recognition performance, but also for gaining deeper insight into how modern deep networks learn semantic and contextual representations for complex visual understanding tasks.

Motivated by these challenges, this thesis investigates semantic-guided indoor scene recognition from a representation-centric perspective. The study examines how

convolutional and transformer-based architectures respond to semantic supervision, how representation structure influences semantic-guided feature interaction, and whether representation-aligned semantic encoding can improve semantic-aware scene understanding in indoor environments.

1.3 RESEARCH OBJECTIVES

The primary objective of this thesis is to investigate how semantic supervision interacts with different deep neural representation architectures for indoor scene recognition. The study focuses on understanding the role of representation structure in semantic-guided feature learning rather than solely improving classification accuracy through larger or more complex models.

To achieve this objective, the thesis aims to address the following specific goals:

1. To study the effectiveness of semantic-guided feature learning for indoor scene recognition using deep neural architectures.
2. To analyze the influence of convolutional backbone design on semantic-aware scene representation learning through a comparative investigation of ResNet and ResNeXt architectures.
3. To investigate how different transformer representation formats, particularly token-based and hierarchical spatial representations, interact with semantic guidance during scene understanding.
4. To examine the importance of representation-aligned semantic encoding for effective integration of semantic and visual information across transformer architectures.
5. To perform a comparative architectural analysis of convolutional and transformer-based semantic-aware frameworks under a consistent experimental setting using the MIT Indoor-67 dataset.
6. To derive representation-level insights regarding how spatial structure, contextual modeling, and representational diversity influence semantic-guided indoor scene recognition.

Through these objectives, the thesis seeks to contribute toward a deeper understanding of semantic-aware representation learning and the role of architecture design in modern scene understanding systems.

1.4 RESEARCH QUESTIONS

The central focus of this thesis is to understand how semantic supervision interacts with different deep representation architectures for indoor scene recognition. In particular, the study investigates whether the effectiveness of semantic-guided learning depends on how visual information is internally represented and organized within deep neural networks.

Based on the problem formulation and research objectives, the thesis is guided by the following research questions:

1. How does semantic supervision influence scene representation learning in deep neural architectures for indoor scene recognition?
2. To what extent does backbone architecture affect the effectiveness of semantic-guided feature interaction within convolutional scene recognition frameworks?
3. How do token-based transformer representations and hierarchical spatial transformer representations differ in their ability to utilize semantic context for scene understanding?
4. Does representation-aligned semantic encoding improve the integration of semantic and visual information across different transformer architectures?
5. Are hierarchical spatial representations more naturally suited for semantic-guided indoor scene recognition than purely token-based representations?
6. How do representational diversity, spatial locality, and contextual modeling contribute to semantic-aware scene understanding across convolutional and transformer-based architectures?

These research questions collectively guide the experimental design and architectural analysis presented throughout the thesis. The investigation is intended not only to evaluate recognition performance, but also to derive broader insights into how representation structure influences semantic-aware visual learning in modern deep neural systems.

1.5 SCOPE OF THE THESIS

This thesis focuses on semantic-guided indoor scene recognition using modern deep neural architectures. The study investigates how semantic supervision interacts with different representation structures in convolutional and transformer-based models, with particular emphasis on representation learning, semantic feature integration, and architectural behavior during scene understanding.

The experimental analysis is conducted using the MIT Indoor-67 dataset, which is a widely adopted benchmark for indoor scene recognition. The work is limited to indoor environments, where scene categories exhibit strong contextual dependencies, large intra-class variation, and significant overlap in object composition and spatial structure. The investigation does not extend to outdoor scene recognition or general-purpose visual classification tasks.

The thesis considers two major representation paradigms in deep visual learning. The first phase investigates convolutional architectures through a semantic-aware framework using ResNet-50 and ResNeXt-50 backbones. The second phase extends the analysis

toward transformer-based architectures, including Vision Transformers and hierarchical Swin Transformers. The study examines how these architectures respond to semantic supervision under controlled experimental settings while maintaining a consistent semantic-aware learning framework.

The scope of the work includes semantic feature extraction, semantic-guided fusion mechanisms, representation-aligned semantic encoding, and comparative architectural analysis. Semantic information is incorporated through segmentation-based contextual representations generated using pretrained semantic segmentation models [11]. However, the segmentation model itself is not optimized or modified as part of this research. The focus remains on understanding how semantic information interacts with different backbone representations during scene recognition.

This thesis does not aim to propose an entirely new state-of-the-art architecture or optimize large-scale benchmark performance through extensive hyperparameter tuning or massive model scaling. Instead, the work is intended as a structured empirical investigation into the relationship between semantic supervision and representation structure across deep neural architectures. Similarly, the study does not explore multimodal extensions involving depth information, video-based scene understanding, language supervision, or cross-domain adaptation.

The experimental evaluation is limited to selected convolutional and transformer architectures in order to maintain a controlled and interpretable comparative framework. While deeper or larger variants may provide additional performance improvements, the objective of the thesis is to analyze representation behavior and semantic interaction rather than exhaustively benchmark all possible architectures.

Overall, the scope of this thesis is centered on developing a representation-aware understanding of semantic-guided indoor scene recognition and analyzing how architectural design influences semantic feature learning in modern deep neural systems.

1.6 THESIS CONTRIBUTIONS

The primary contribution of this thesis lies in the systematic investigation of semantic-guided scene recognition from a representation-learning perspective across modern deep neural architectures. Rather than focusing solely on benchmark-oriented performance improvement, the study analyzes how different representation structures influence the integration and effectiveness of semantic supervision in indoor scene understanding.

The major contributions of the thesis are summarized as follows:

1. A semantic-aware indoor scene recognition framework is investigated using convolutional and transformer-based architectures under a consistent experimental setting to analyze the interaction between semantic guidance and visual representation learning.
2. A comparative architectural study is conducted between ResNet-50 and ResNeXt-50 within a semantic-guided convolutional framework, demonstrating the influence of aggregated residual transformations and representational diversity on semantic-aware scene recognition performance.
3. A representation-aligned semantic learning framework is explored for transformer-based scene recognition, where semantic encoding strategies are

adapted according to the representation format of the underlying transformer architecture.

4. The study provides an analysis of token-based transformer representations and hierarchical spatial transformer representations in the context of semantic-guided indoor scene understanding using Vision Transformers and Swin Transformers.
5. Experimental observations show that hierarchical spatial representations preserve semantic and contextual relationships more effectively than purely token-based representations for indoor scene recognition tasks.
6. The thesis establishes that semantic supervision becomes more effective when the semantic representation format is aligned with the native representation structure of the underlying architecture.
7. The work contributes broader representation-level insights into how spatial locality, contextual structure, and representational diversity influence semantic-aware feature learning across deep neural architectures.

Collectively, these contributions position the thesis as a structured empirical investigation into semantic-aware representation learning for indoor scene understanding rather than a purely benchmark-driven architectural study.

1.7 THESIS ORGANIZATION

This thesis is organized into six chapters, each progressively developing the investigation into semantic-guided representation learning for indoor scene recognition.

Chapter 1 introduces the background and motivation behind indoor scene recognition and discusses the importance of semantic context in complex visual understanding tasks. The chapter presents the problem statement, research objectives, research questions, scope of the study, major contributions, and the overall organization of the thesis.

Chapter 2 presents a detailed literature review covering conventional scene recognition methods, convolutional neural networks, semantic-aware scene recognition approaches, and transformer-based visual architectures. The chapter further discusses representation-centric analysis relevant to semantic-guided scene understanding and identifies the research gaps motivating the present work.

Chapter 3 investigates semantic-guided scene recognition using convolutional neural networks. A dual-branch semantic-aware framework is analyzed using ResNet-50 and ResNeXt-50 backbones in order to study the influence of representational diversity and aggregated residual learning on semantic-aware feature interaction.

Chapter 4 extends the investigation toward transformer-based architectures. The chapter analyzes how semantic supervision interacts with token-based and hierarchical transformer representations using Vision Transformers and Swin Transformers within a representation-aligned semantic learning framework.

Chapter 5 presents the experimental results, comparative analysis, and representation-level discussion of the investigated architectures. The chapter interprets the observed performance trends and examines how representation structure influences semantic-guided scene understanding across convolutional and transformer-based frameworks.

Chapter 6 concludes the thesis by summarizing the major findings and architectural insights derived from the study. The chapter also discusses future research directions and

the broader relevance of semantic-aware representation learning for intelligent visual understanding systems.

CHAPTER 2

LITERATURE REVIEW AND PROBLEM IDENTIFICATION

2.1 INTRODUCTION

Scene recognition has evolved into an important research problem in computer vision due to its central role in enabling machines to interpret complex visual environments. Unlike object recognition, which primarily focuses on identifying isolated entities, scene recognition requires understanding the overall semantic and spatial structure of an environment [12]. Indoor scene understanding is particularly challenging because scenes are often defined through a combination of object presence, contextual relationships, and spatial organization rather than by a single dominant visual cue. These challenges have motivated extensive research into learning more effective visual representations for scene-level understanding [13], [14].

Early scene recognition methods relied on handcrafted visual descriptors designed to capture texture, gradients, and local structural patterns. Although these approaches achieved moderate success on simpler datasets, they struggled to generalize to complex indoor environments characterized by large intra-class variation and strong inter-class similarity [5]. The emergence of deep convolutional neural networks significantly improved scene recognition performance by enabling hierarchical feature learning directly from image data. Convolutional architectures demonstrated strong capability in capturing progressively abstract visual representations, leading to substantial advances in large-scale visual recognition tasks.

Despite these improvements, purely appearance-based feature learning often remains insufficient for robust indoor scene understanding. Many indoor environments contain overlapping visual patterns and similar object distributions, making semantic context increasingly important for reliable recognition. Consequently, several studies have explored semantic-aware scene recognition frameworks that integrate object-level or segmentation-based semantic information with visual representations. These approaches demonstrate that semantic guidance can provide complementary contextual cues that improve scene understanding beyond conventional RGB feature learning alone.

More recently, transformer-based architectures have introduced new representation paradigms for visual learning. Vision Transformers model images as sequences of patch tokens and learn global relationships through self-attention mechanisms, while hierarchical transformers preserve spatial feature organization across multiple representation stages. These architectural differences suggest that semantic information may interact differently with various representation formats depending on how spatial structure and contextual relationships are internally modeled within the network.

Motivated by these developments, this chapter reviews existing literature related to scene recognition, semantic-aware visual learning, convolutional and transformer-based architectures, and modern representation learning approaches. The chapter further

examines how semantic guidance has been incorporated into deep scene recognition systems and identifies the research gaps that motivate the representation-centric investigation presented in this thesis.

2.2 EVOLUTION OF SCENE RECOGNITION

Scene recognition has undergone significant evolution over the past two decades, progressing from handcrafted visual descriptor-based methods to modern deep representation learning frameworks. This progression has largely been driven by the increasing complexity of visual understanding tasks and the growing need for models capable of capturing semantic context, spatial structure, and high-level environmental relationships. As indoor scene understanding involves interpreting multiple interacting visual components rather than isolated objects, advances in scene recognition have closely followed developments in visual representation learning.

Early scene recognition methods primarily relied on handcrafted feature descriptors designed to capture low-level appearance patterns such as texture, edges, gradients, and local spatial structures. Traditional approaches commonly used descriptors including Scale-Invariant Feature Transform (SIFT) [2], Histogram of Oriented Gradients (HOG) [3], GIST [4], and bag-of-visual-words representations to model scene characteristics. These methods attempted to encode global scene layout and local image statistics using manually designed features combined with conventional machine learning classifiers. Although such approaches demonstrated reasonable performance on constrained datasets, they often struggled to generalize to complex real-world environments containing large appearance variation, clutter, viewpoint changes, and overlapping object distributions.

The introduction of large-scale scene datasets such as MIT Indoor-67 significantly highlighted the limitations of handcrafted representations for indoor scene understanding. Indoor environments contain highly diverse object arrangements and contextual dependencies that are difficult to model through manually engineered descriptors alone. As a result, the focus gradually shifted toward learning feature representations directly from data rather than relying on fixed handcrafted visual patterns.

The emergence of deep convolutional neural networks marked a major transition in scene recognition research. CNN-based architectures enabled hierarchical representation learning, where lower layers captured local visual primitives while deeper layers learned increasingly abstract semantic structures. This hierarchical feature learning capability substantially improved scene recognition performance across multiple benchmarks. Architectures such as AlexNet [15], VGGNet [16], GoogLeNet [17], and later ResNet [18] demonstrated that deep visual representations could effectively capture both local appearance cues and broader contextual information necessary for scene understanding.

Residual learning further improved deep representation learning by addressing optimization difficulties associated with increasing network depth. ResNet introduced identity shortcut connections that enabled stable training of deeper convolutional networks, allowing models to learn richer and more discriminative visual representations [18]. Subsequent architectures such as ResNeXt [19] further expanded residual learning by combining multiple parallel grouped transformations within each residual block, improving representational diversity while maintaining computational efficiency [19]. These developments suggested that architectural design itself plays an important role in determining the quality and diversity of learned scene representations.

At the same time, researchers increasingly recognized that scene understanding depends not only on visual appearance but also on semantic relationships between objects and spatial context. This led to the development of semantic-aware scene recognition approaches that incorporated object-level cues, semantic segmentation information, contextual attention mechanisms, and multimodal feature fusion strategies [5], [20]. Semantic-guided methods demonstrated that integrating contextual object information could improve robustness in challenging indoor environments where scene categories exhibit strong visual overlap [7], [10].

More recently, transformer-based architectures have introduced a new direction for visual representation learning. Inspired by advances in natural language processing, Vision Transformers model images as sequences of patch embeddings and use self-attention mechanisms to capture long-range relationships between image regions [8]. Unlike convolutional networks, which rely primarily on localized receptive fields and hierarchical spatial feature extraction, transformers emphasize global relational modeling through attention-based interactions. This shift introduced fundamentally different representation structures for visual learning.

Hierarchical transformer architectures such as the Swin Transformer [9] further expanded transformer-based visual learning by combining self-attention with multi-stage spatial feature hierarchies. These models preserve locality and spatial continuity while enabling progressively larger receptive fields across network stages. As a result, modern scene recognition research now includes multiple representation paradigms, ranging from convolutional spatial hierarchies to token-based transformer representations and hybrid hierarchical attention models.

The evolution of scene recognition therefore reflects a broader transition in computer vision from handcrafted appearance modeling toward increasingly sophisticated representation learning strategies. While early research primarily focused on visual pattern extraction, recent developments emphasize contextual understanding, semantic reasoning, and representation structure. These advances have created new opportunities to investigate how semantic information interacts with different deep representation architectures during scene understanding, particularly in complex indoor environments where contextual and spatial relationships play a central role.

2.3 DEEP CNN-BASED SCENE RECOGNITION

The introduction of deep convolutional neural networks (CNNs) significantly transformed scene recognition research by enabling models to learn hierarchical visual representations directly from image data. Unlike handcrafted feature-based approaches, CNNs automatically learn discriminative representations through multiple layers of convolutional operations, allowing the network to progressively capture low-level textures, mid-level structural patterns, and higher-level semantic information. This ability to learn hierarchical representations proved particularly effective for scene understanding tasks, where recognition depends on capturing both local object details and broader contextual structure simultaneously.

Early deep learning architectures such as AlexNet demonstrated the effectiveness of deep convolutional feature learning for large-scale ImageNet visual recognition tasks, motivating the broader adoption of CNNs for scene classification problems [15]. Subsequent architectures including VGGNet and GoogLeNet further improved representation quality through deeper networks and computationally efficient

architectural designs [16–17]. These models established that deeper convolutional representations could capture increasingly abstract scene-level characteristics beyond simple object appearance.

One of the most important developments in deep CNN-based representation learning was the introduction of residual learning through ResNet architectures. Increasing network depth had previously led to optimization difficulties such as vanishing gradients and degradation in training performance. ResNet mitigated these limitations by incorporating residual skip pathways that allowed feature information to flow directly across layers. This residual formulation made it possible to optimize much deeper convolutional architectures more reliably while maintaining effective gradient propagation during training [18].

For scene recognition tasks, residual learning proved particularly beneficial because indoor environments often contain complex combinations of objects, textures, spatial layouts, and contextual patterns that require rich hierarchical representations. Deep residual networks demonstrated strong capability in learning discriminative scene-level features while maintaining computational feasibility [5]. As a result, ResNet-based architectures became widely adopted in semantic-aware and multimodal scene recognition frameworks.

Despite the strong performance of residual networks, increasing network depth alone does not always guarantee improved representation quality for complex scene understanding tasks. Scene recognition requires the model to simultaneously capture diverse visual cues, including object groupings, structural layouts, contextual arrangements, and background patterns. Standard residual blocks typically learn a single transformation pathway within each block, which may limit the diversity of learned feature representations at a fixed depth.

To overcome the limitations of conventional residual architectures, ResNeXt introduced grouped residual transformations as an additional design dimension referred to as cardinality. Rather than scaling only network depth or layer width, the architecture combines several parallel transformation branches within each residual block and aggregates their outputs into a unified representation [19]. This approach improves feature diversity while keeping computational cost relatively manageable.

The idea of cardinality is particularly relevant for scene recognition, where indoor environments often contain multiple semantic cues that appear simultaneously across different regions of the image. Parallel grouped transformations allow the network to model varied visual characteristics at the same time, including object configurations, spatial organization, and contextual scene patterns. Prior studies suggest that increasing cardinality can strengthen representation learning more effectively than relying solely on deeper or wider network structures, especially for visually complex recognition tasks [19].

At the same time, CNN-based scene recognition research increasingly incorporated contextual and semantic information to complement appearance-based feature learning. Several studies explored combining convolutional RGB features with object-level cues, semantic segmentation maps, and attention-guided contextual representations. These semantic-aware frameworks demonstrated that convolutional representations become more effective when semantic information is integrated during feature learning and fusion [5], [20].

Attention mechanisms further improved CNN-based scene recognition by enabling models to emphasize informative spatial regions and suppress less relevant visual responses [7]. In semantic-guided frameworks, attention-based fusion strategies often use semantic features to regulate convolutional feature activations, allowing contextual object information to guide scene representation learning [5]. Such approaches have shown improved robustness in distinguishing visually similar indoor environments where contextual semantics play an important role.

Although convolutional architectures have demonstrated strong capability in scene recognition, their representation behavior remains closely tied to architectural design choices such as residual connectivity, grouped transformations, receptive field growth, and feature aggregation strategy. These observations suggest that the effectiveness of semantic-guided learning may depend not only on the availability of semantic information, but also on how convolutional architectures internally organize and diversify visual representations.

Consequently, modern CNN-based scene recognition research has gradually shifted from purely improving classification accuracy toward understanding how architectural structure influences representation learning and contextual feature interaction. This perspective is particularly relevant for semantic-aware indoor scene recognition, where successful understanding depends on effective integration of visual appearance, contextual semantics, and spatial relationships within the learned representation space.

2.4 SEMANTIC-AWARE SCENE RECOGNITION

Traditional scene recognition approaches primarily relied on learning visual representations directly from RGB images. Although deep convolutional networks substantially improved recognition performance through hierarchical feature extraction, appearance-based learning alone often remains insufficient for robust understanding of complex indoor environments. Indoor scenes are frequently characterized by contextual relationships among multiple objects and spatial arrangements rather than by isolated visual patterns. As a result, visually similar environments may belong to different scene categories despite sharing common textures, structures, or object distributions. These limitations motivated the development of semantic-aware scene recognition approaches that incorporate higher-level contextual information into the recognition process [10].

Semantic-aware scene recognition aims to improve scene understanding by integrating semantic cues such as object presence, semantic segmentation maps, contextual relationships, or region-level scene attributes with conventional visual representations [5]. Instead of relying entirely on appearance information, these methods attempt to guide representation learning using structured semantic context that reflects the functional composition of the environment. Such contextual information becomes particularly valuable in indoor scenes where object arrangement and semantic relationships strongly influence scene identity.

Early semantic-aware approaches explored the use of object detection outputs and semantic attributes to complement visual feature learning. These methods demonstrated that object-level contextual cues could improve scene classification performance by providing information about functional scene composition [5,20]. For example, the presence of objects such as beds, desks, ovens, or bookshelves can provide important semantic indicators regarding the underlying scene category. However, object-centric semantic representations alone were often insufficient because scene understanding also

depends on spatial layout and contextual interaction among objects as depicted in the Fig 2.1 given below.

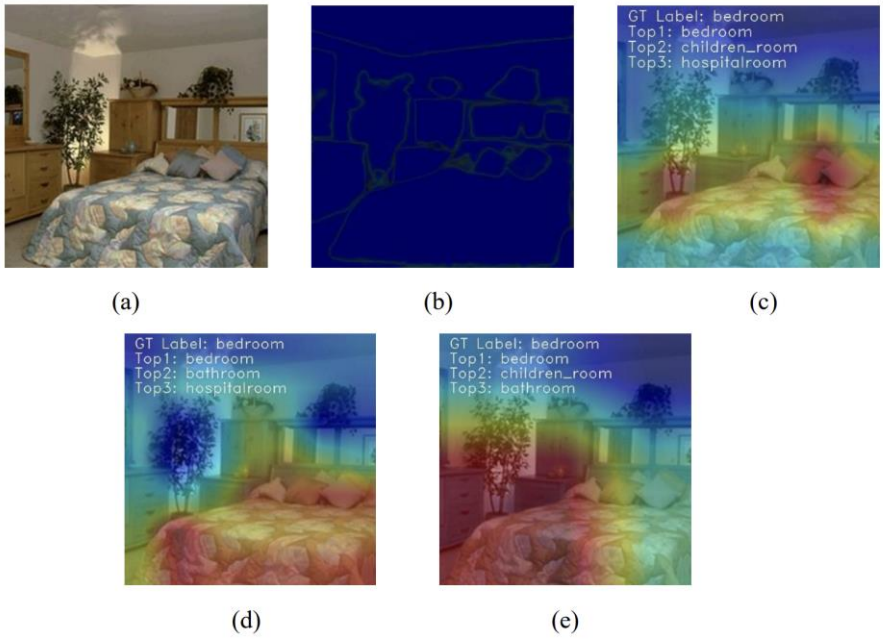

Fig 2.1: *Scene recognition outcomes based on input data:*

(a) An RGB image illustrating the "Bedroom" scene category.

(b) Semantic segmentation derived from (a).

(c) Class Activation Map (CAM) generated solely from the RGB image (a).

(d) CAM generated exclusively from the semantic segmentation (b).

(e) CAM created using the proposed method, which combines both (a) and (b).

The Top@3 predicted classes are displayed in the upper-left corner of images (c) to (e).

The growing availability of semantic segmentation models further expanded semantic-aware scene recognition research. Semantic segmentation provides dense pixel-level semantic annotations describing object categories and spatial structure throughout the image [11]. Several studies utilized segmentation-based semantic maps as auxiliary inputs for scene recognition frameworks, allowing models to capture richer contextual information beyond conventional RGB appearance features. By incorporating semantic segmentation representations, scene recognition systems became more capable of distinguishing visually ambiguous environments through contextual reasoning [5], [20].

Deep semantic-aware frameworks commonly employ multi-branch architectures consisting of separate RGB and semantic processing streams [5], [20]. In such frameworks, the RGB branch learns visual appearance representations while the semantic branch extracts object-level contextual information from segmentation maps or semantic annotations. The outputs of these branches are subsequently combined through feature

fusion mechanisms to produce semantically enriched scene representations. This multimodal learning strategy allows semantic context to guide or complement visual feature learning during scene classification .

Attention mechanisms have become increasingly important in semantic-aware scene recognition frameworks. Instead of relying solely on direct feature concatenation, attention-based fusion allows semantic cues to dynamically influence visual feature representations during learning. In many semantic-guided architectures, attention modules help the network assign greater importance to spatially relevant scene regions while reducing the influence of less useful background information [20]. This selective interaction between semantic and visual features has been shown to improve recognition performance in indoor environments, where fine contextual details often distinguish visually similar scene categories [5], [7].

Several studies have also explored contextual reasoning mechanisms to model relationships between objects, scene layout, and semantic structure more explicitly. Graph-based contextual learning, region interaction modeling, and semantic relation learning frameworks have demonstrated that scene understanding benefits from capturing interactions among semantic entities rather than treating objects independently. These approaches further highlight that scene recognition is fundamentally a contextual reasoning problem rather than a purely appearance-based classification task.

Despite the effectiveness of semantic-aware learning, most existing methods primarily focus on improving recognition accuracy through stronger semantic encoders, larger backbones, or increasingly sophisticated fusion modules. Comparatively less attention has been given to analyzing how the representation structure of the underlying architecture influences semantic-guided feature interaction. In many cases, identical semantic fusion strategies are applied across different backbone architectures without considering whether semantic representations align naturally with the internal organization of visual features.

This limitation becomes increasingly important with the emergence of modern representation paradigms beyond conventional convolutional architectures. Semantic information is inherently structured and spatially contextual, and its effectiveness may depend on how the underlying model represents spatial relationships, contextual continuity, and feature interactions. Convolutional architectures preserve locality through hierarchical spatial feature maps, whereas transformer-based architectures may organize information through token sequences or hierarchical attention representations. Consequently, semantic guidance may interact differently with different representation formats.

These observations suggest that understanding semantic-aware scene recognition requires more than evaluating classification accuracy alone. It also requires examining how semantic supervision interacts with the representation structure of deep neural architectures during feature learning and scene understanding. This perspective motivates the representation-centric investigation explored in the subsequent chapters of this thesis.

2.5 TRANSFORMER-BASED VISUAL REPRESENTATION LEARNING

Transformer-based architectures have recently emerged as an important direction in visual representation learning due to their ability to model long-range relationships through self-attention mechanisms. Originally developed for natural language processing

tasks, transformers demonstrated strong capability in learning contextual dependencies between sequential elements. Their extension to computer vision introduced a fundamentally different representation paradigm compared to conventional convolutional neural networks, shifting the focus from localized convolutional feature extraction toward global relational modelling [10].

Vision Transformer models an image as a sequence of visual tokens instead of maintaining conventional spatial feature maps throughout the network. The input image is first partitioned into non-overlapping patches, after which each patch is flattened and transformed into a learned token embedding. These token representations are processed using multiple transformer encoder layers composed of multi-head self-attention and feed-forward modules. Through the self-attention mechanism, each token can exchange information with all other tokens in the sequence, enabling the architecture to model long-range contextual dependencies across different image regions [8].

Unlike convolutional architectures, which rely on inductive biases such as locality and weight sharing, Vision Transformers learn visual relationships primarily through attention-based interactions. This design provides strong flexibility in modelling contextual dependencies and long-range semantic interactions. As a result, transformer-based models have demonstrated competitive performance across multiple visual recognition tasks, including image classification, object detection, segmentation, and scene understanding.

However, the representation structure of Vision Transformers differs substantially from that of convolutional networks. CNNs preserve spatial continuity through hierarchical feature maps, whereas ViTs organize visual information as token sequences [8]. Although positional embeddings provide implicit spatial information, explicit spatial locality is not directly maintained in the same manner as convolutional representations [1]. Consequently, transformer-based visual learning introduces a different balance between global contextual modeling and spatial structural preservation.

These representational differences become particularly important for scene recognition tasks. Indoor scene understanding depends heavily on both object relationships and spatial organization within the environment. While global self-attention enables flexible contextual interaction across image regions, purely token-based representations may weaken local spatial continuity that is often important for understanding scene layout and structural arrangement [8], [21]. This challenge motivated the development of hierarchical transformer architectures that combine attention-based learning with multi-scale spatial representation hierarchies.

Swin Transformer proposed a hierarchical transformer architecture that retains spatial feature structure while benefiting from attention-driven contextual modeling. Rather than computing attention across the entire image at every stage, the model restricts self-attention to localized windows and gradually combines information through successive hierarchical layers. To facilitate communication beyond individual windows, shifted window partitioning is introduced, enabling neighboring regions to exchange information without substantially increasing computational cost [9].

This hierarchical organization makes the Swin architecture more aligned with conventional convolutional feature pyramids, while still relying on transformer-based attention operations for representation learning. Intermediate feature representations continue to preserve spatial locality across stages, helping maintain structural consistency and contextual relationships during learning [9]. These characteristics are especially valuable for indoor scene recognition tasks, where accurate understanding often depends

not only on the presence of objects but also on how they are spatially arranged within the environment.

Transformer-based visual learning has also been increasingly explored in multimodal and semantic-aware recognition tasks [6], [22]. Several recent studies have incorporated semantic information, contextual embeddings, or cross-modal attention mechanisms into transformer frameworks to improve scene understanding and contextual reasoning [10], [21], [22]. Attention-based architectures naturally support flexible feature interaction between modalities, making them suitable for integrating semantic cues with visual representations [7].

Nevertheless, the effectiveness of semantic guidance within transformer architectures may depend strongly on the underlying representation format. Token-based transformers process information through sequential embeddings and global attention interactions, whereas hierarchical transformers maintain structured spatial feature organization across representation stages. Semantic information itself is inherently spatial and contextual, suggesting that semantic feature integration may behave differently across these representation paradigms.

Despite the growing adoption of transformer-based models for visual recognition, comparatively limited work has examined how representation structure influences semantic-guided feature interaction within transformer architectures. Existing studies primarily focus on performance improvement through larger models or stronger attention mechanisms [7], [21], while representation-level compatibility between semantic information and visual structure remains less explored.

These observations indicate that transformer-based scene recognition should not be viewed solely as a replacement for convolutional architectures, but rather as an alternative representation paradigm with distinct semantic interaction behavior. Understanding how semantic supervision interacts with token-based and hierarchical transformer representations is therefore important for developing more effective semantic-aware scene understanding systems. This representation-centric perspective forms an important foundation for the architectural analysis presented later in this thesis.

2.6 REPRESENTATION LEARNING PERSPECTIVE

Representation learning forms a fundamental component of modern scene recognition because the organization and quality of internal feature representations strongly influence how effectively a model can interpret spatial context and semantic relationships within an environment. In indoor scene understanding, the goal extends beyond recognizing isolated visual elements; models must also encode object arrangements, contextual dependencies, and overall scene composition within a unified representation space. As deep learning methods have progressed, multiple architectural paradigms have emerged, each structuring visual information differently. These differences in representation organization can influence how semantic cues interact with visual features during scene interpretation.

Convolutional neural networks primarily construct hierarchical spatial representations using localized convolutional filtering operations. Initial layers typically respond to low-level visual characteristics such as edges, gradients, and texture patterns, whereas deeper layers gradually encode more abstract semantic structures and contextual information. Throughout this hierarchy, CNNs retain spatial neighborhood relationships within feature

maps, allowing structural coherence to persist across successive representation stages. Such spatially organized feature hierarchies are particularly useful for scene recognition tasks in which object placement and environmental layout contribute significantly to scene identity [15].

Residual architectures further strengthened convolutional representation learning by improving information propagation across deep feature hierarchies. Architectures such as ResNeXt [19] expanded this idea through grouped residual transformations that encourage multiple parallel feature pathways within individual blocks. From a representation-learning standpoint, these architectural variations allow networks to encode a broader range of visual and contextual patterns simultaneously. In scene recognition, where environments often contain diverse object arrangements and overlapping semantic signals, increased representational diversity can support more effective semantic-feature interaction.

Transformer-based vision models introduced a contrasting representation paradigm built around token-wise interaction through self-attention operations. Vision Transformer processes images as sequences of visual tokens and models relationships globally through attention across the token set. Unlike convolutional representations, transformer features are not strictly limited by localized receptive fields, enabling direct contextual interaction between distant image regions. This ability to model long-range dependencies allows transformers to capture broad contextual relationships efficiently [8].

At the same time, token-oriented representations modify how spatial structure is maintained within the learned feature space. Because images are represented as ordered token sequences rather than continuous spatial feature maps, locality is preserved more indirectly through positional encoding and learned attention behavior. Although this formulation supports strong contextual reasoning at the global level, certain fine-grained spatial dependencies important for layout-sensitive scene understanding may become less explicitly represented.

Hierarchical transformer architectures aim to combine the advantages of attention-based contextual modeling with structured spatial representation learning. Models such as Swin Transformer preserve multi-stage spatial organization while progressively enlarging contextual receptive fields through localized attention windows. Consequently, hierarchical transformers retain stronger spatial continuity than purely token-based transformer models while continuing to benefit from attention-driven contextual interaction [9].

These representational characteristics become especially significant when semantic supervision is integrated into the learning pipeline. Semantic cues obtained from segmentation maps or object-level annotations are closely tied to both contextual relationships and spatial arrangement. As a result, the effectiveness of semantic-guided learning depends not only on the semantic information itself, but also on how naturally it fits within the representational organization of the backbone architecture.

In convolutional models, semantic guidance can interact directly with hierarchical spatial feature maps, allowing localized semantic modulation across visual regions and contextual structures [5]. Hierarchical transformers similarly maintain sufficient spatial organization for semantic features to align effectively with visual representations during fusion. By comparison, token-based transformer architectures often require semantic information to be reformulated into compatible token embeddings before effective interaction can occur. These observations indicate that semantic encoding strategies and

backbone representation formats should ideally remain structurally compatible during multimodal integration.

More broadly, representation learning in semantic-aware scene recognition involves balancing three interconnected aspects: contextual reasoning, preservation of spatial organization, and compatibility with semantic information. Architectures emphasizing global interaction may model wide contextual relationships effectively, but can reduce sensitivity to localized structural cues. Conversely, architectures that maintain stronger spatial hierarchy may better support structured semantic integration while preserving scene-level continuity.

Accordingly, understanding semantic-aware scene recognition requires examining not only classification accuracy, but also the internal organization of semantic and contextual information within learned representations. The success of semantic supervision may ultimately depend on whether the architecture naturally supports the type of semantic interaction required for scene understanding. This representation-centered viewpoint provides the conceptual basis for the comparative architectural analysis presented in the subsequent chapters of this thesis.

2.7 RESEARCH GAPS AND PROBLEM IDENTIFICATION

Recent advances in deep learning have substantially improved scene recognition performance through stronger visual representations, deeper architectures, semantic-aware learning strategies, and transformer-based contextual modeling. Convolutional neural networks demonstrated the effectiveness of hierarchical feature learning for capturing scene-level visual patterns, while semantic-aware frameworks showed that integrating contextual object information can further improve indoor scene understanding. More recently, transformer-based architectures introduced new representation paradigms capable of modeling long-range contextual relationships through self-attention mechanisms. Collectively, these developments have significantly advanced the state of scene recognition research.

At the same time, existing literature reveals several important limitations that remain insufficiently explored, particularly from a representation-learning perspective. Most semantic-aware scene recognition approaches primarily focus on improving classification performance through stronger backbones, larger models, enhanced semantic encoders, or increasingly sophisticated fusion mechanisms. Although these methods demonstrate that semantic guidance can improve scene understanding, comparatively less attention has been given to understanding how semantic supervision interacts with different representation architectures during feature learning.

One major limitation is that many existing semantic-guided frameworks treat semantic information as a generic auxiliary modality without considering whether the semantic representation format aligns naturally with the internal representation structure of the backbone architecture. In several studies, identical semantic fusion strategies are applied across different architectures despite substantial differences in how visual information is organized and processed internally. Such approaches often overlook the possibility that semantic guidance may behave differently depending on whether the underlying representation preserves spatial hierarchy, contextual locality, or token-level relational structure.

This issue becomes increasingly important with the emergence of transformer-based visual architectures. Vision Transformers organize images as token sequences and emphasize global contextual interaction through self-attention, whereas convolutional networks preserve locality through hierarchical spatial feature maps. Hierarchical transformers such as the Swin Transformer introduce yet another representation paradigm by combining spatial hierarchies with attention-based contextual modeling. Despite these architectural differences, existing studies rarely analyze how semantic supervision interacts with these distinct representation structures under controlled experimental conditions.

Another important research gap is the limited comparative analysis between convolutional and transformer-based semantic-aware scene recognition frameworks from a representation-centric perspective. Most existing comparisons focus primarily on benchmark performance and parameter scaling rather than examining how different architectures utilize semantic context during feature interaction. Consequently, important questions regarding representation compatibility, semantic alignment, and contextual integration remain insufficiently addressed.

Furthermore, several semantic-aware approaches rely heavily on stronger architectures or larger-scale models without isolating the influence of representation structure itself. As a result, it becomes difficult to determine whether observed performance improvements arise from increased model capacity, architectural design, representational diversity, or more effective semantic interaction mechanisms. A more controlled investigation is therefore required to understand how semantic guidance behaves across architectures while maintaining consistent semantic-aware learning settings.

The literature also indicates that relatively limited attention has been given to representation alignment between semantic features and backbone representations. Semantic information derived from segmentation maps is inherently spatial and context-dependent. However, token-based architectures may require semantic information to be represented differently from hierarchical spatial architectures in order to achieve meaningful interaction. Existing works seldom investigate whether adapting semantic encoding according to the representation format of the backbone can improve semantic-guided scene understanding.

Based on these observations, the central problem addressed in this thesis can be formulated as follows:

> "*Existing semantic-aware scene recognition methods insufficiently investigate how representation structure influences semantic-guided feature learning across convolutional and transformer-based architectures.*"

Motivated by this problem, the present work investigates semantic-guided indoor scene recognition from a representation-learning perspective. The study analyzes how different deep architectures respond to semantic supervision, how representation structure affects semantic interaction, and whether representation-aligned semantic encoding can improve scene understanding across convolutional and transformer-based frameworks. By conducting a controlled comparative investigation using semantic-aware CNN and transformer architectures, the thesis aims to provide deeper insight into the relationship between semantic supervision and representation learning in modern scene recognition systems.

CHAPTER 3

SEMANTIC-GUIDED CNN FRAMEWORK

3.1 INTRODUCTION

Deep convolutional neural networks have played a major role in advancing scene recognition by enabling hierarchical representation learning from large-scale visual data. Convolutional architectures progressively learn hierarchical visual patterns through stacked convolutional operations, beginning with low-level textures and edges, advancing toward structural configurations, and eventually capturing higher-level semantic abstractions. This hierarchical feature learning makes CNN-based models well suited for visual understanding tasks involving complex environmental and spatial structure [23]. In indoor scene recognition, where scenes are characterized by object co-occurrence, contextual relationships, and spatial organization, hierarchical convolutional representations provide an effective mechanism for capturing both local and global visual information [24], [25].

At the same time, several studies have shown that appearance-based convolutional representations alone are often insufficient for robust indoor scene understanding [7]. Indoor environments frequently contain visually overlapping structures and semantically similar object distributions, making contextual reasoning increasingly important for reliable classification. As discussed in the previous chapter, semantic-aware scene recognition approaches attempt to address this limitation by incorporating object-level semantic information alongside visual representations. Such semantic guidance allows the model to utilize contextual cues that may not be explicitly captured through RGB appearance learning alone.

Although semantic-aware learning has shown promising results, the effectiveness of semantic-guided feature interaction depends strongly on the quality and diversity of the learned visual representations. Different convolutional architectures organize and transform visual information differently, which may influence how semantic information interacts with learned feature maps during fusion. However, many existing semantic-aware frameworks adopt backbone architectures without systematically analyzing how architectural design affects semantic-guided scene understanding.

Motivated by this observation, the present chapter investigates semantic-guided indoor scene recognition within a convolutional representation learning framework. The study focuses on analyzing how backbone architecture influences semantic-aware feature learning under a controlled multimodal setting. In particular, a comparative investigation is conducted between ResNet-50 and ResNeXt-50 (32×4d), where the overall semantic-guided framework, fusion strategy, training procedure, and experimental setup remain unchanged while only the RGB backbone architecture is modified. This design enables a focused analysis of how representational diversity and aggregated residual transformations affect semantic-guided scene representation learning.

The investigation is performed using a dual-branch semantic-aware framework consisting of an RGB branch and a semantic branch connected through an attention-based fusion mechanism. Semantic information derived from segmentation maps is used to guide visual feature learning by modulating convolutional representations through semantic-aware attention. Within this setting, the chapter examines whether architectures with richer representational diversity provide more effective semantic-guided scene understanding for complex indoor environments.

The remainder of this chapter presents the proposed semantic-aware convolutional framework, discusses the investigated backbone architectures, explains the semantic-guided fusion strategy and training methodology, and analyzes the experimental observations obtained on the MIT Indoor-67 dataset.

3.2 FRAMEWORK OVERVIEW

The proposed framework is designed to learn complementary information from visual appearance and semantic scene cues for indoor scene recognition. The architecture follows a dual-branch formulation consisting of an RGB branch and a semantic branch, whose outputs are integrated through an attention-based fusion module prior to final scene classification. A conceptual overview of the framework is presented in Fig. 3.2. The overall design is inspired by semantic-aware scene recognition approaches that combine appearance features with object-level semantic information to improve scene understanding [5].

The RGB branch is responsible for extracting high-level visual features from the input image. This branch captures appearance-driven information including scene layout, texture variations, structural arrangements, and object-related visual patterns using a deep convolutional backbone. In the present study, the RGB stream serves as the primary component for examining how convolutional representation structure influences semantic-aware scene recognition. To maintain a controlled experimental setting, the overall framework remains unchanged while only the RGB backbone is varied between ResNet-50 and ResNeXt-50 (32×4d).

Alongside the RGB stream, the semantic branch processes semantic representations generated from precomputed segmentation maps. These semantic inputs provide additional information regarding object categories and spatial composition that may not be fully captured through RGB appearance features alone. The semantic branch employs a lightweight convolutional encoder with channel attention mechanisms to selectively emphasize informative semantic responses while reducing less useful activations. This design allows semantic information to contribute complementary scene-level cues during feature extraction.

The outputs of the RGB and semantic branches are integrated using an attention-based fusion strategy. Instead of performing direct feature concatenation, semantic features are used to regulate visual feature responses through adaptive modulation. This interaction allows the network to emphasize visually relevant scene regions based on semantic evidence and improves the integration of object-level and structural information within the fused feature space. Consequently, the framework aims to produce more discriminative scene representations for indoor environment recognition.

An important characteristic of the proposed framework is that all experiments are conducted under identical fusion and optimization settings. The semantic branch

architecture, fusion module, training strategy, and classification pipeline remain fixed throughout the study. As a result, observed performance differences can be attributed primarily to the choice of convolutional backbone rather than variations in fusion design or optimization methodology.

From an architectural perspective, the framework provides a controlled setting for analyzing how different convolutional backbones interact with semantic supervision during scene recognition. Since semantic information is integrated directly with high-level convolutional feature maps, the effectiveness of the overall model may depend on factors such as feature diversity, residual feature quality, and the ability of the backbone to capture complex spatial patterns. This design therefore provides a suitable basis for investigating the relationship between convolutional representation structure and semantic-aware indoor scene recognition performance.

3.3 RGB BACKBONE ARCHITECTURES

The RGB branch forms the primary visual representation component of the proposed semantic-guided framework. Its role is to extract high-level convolutional feature representations capable of capturing scene layout, structural composition, texture patterns, and contextual visual cues from indoor environments. Since the objective of this study is to analyze how convolutional representation structure influences semantic-guided scene understanding, particular emphasis is placed on the choice of backbone architecture used within the RGB branch.

To perform a controlled architectural investigation, two residual convolutional architectures are evaluated within the same semantic-aware framework: ResNet-50 and ResNeXt-50 (32×4d). Both architectures follow residual learning principles and maintain comparable overall network organization, allowing a fair comparison while isolating the influence of representational diversity and grouped residual transformations.

To improve optimization stability in deeper convolutional networks, ResNet employed residual mappings connected through identity shortcut pathways, allowing gradients to propagate more effectively across layers [18]. Instead of forcing each layer to learn a complete transformation directly, residual blocks focus on learning residual functions relative to the input representation. This formulation supports the training of substantially deeper architectures while reducing degradation problems commonly observed in deep CNN optimization. For scene recognition tasks, such hierarchical feature extraction is particularly important because indoor environments contain multiple levels of visual and semantic structure, ranging from local texture patterns and object-level details to broader spatial arrangements and scene layout information.

Within the proposed framework, ResNet-50 serves as the baseline convolutional architecture for semantic-guided scene understanding. The network progressively extracts hierarchical feature maps through stacked residual stages, producing high-level visual representations that are subsequently integrated with semantic information through attention-guided fusion. Although ResNet demonstrates strong capability in learning discriminative scene representations, each residual block primarily learns a single transformation pathway, which may limit representational diversity in visually complex environments.

Indoor scenes often contain heterogeneous visual structures, overlapping object distributions, and multiple contextual cues that must be interpreted simultaneously.

Capturing such diverse scene characteristics may require richer feature transformations capable of modeling multiple visual patterns in parallel. Motivated by this observation, the study further investigates ResNeXt-50 (32×4d) as an alternative convolutional backbone within the same semantic-aware framework.

ResNeXt extends the residual learning framework by incorporating grouped convolutional pathways within each residual block, allowing multiple feature mappings to be learned in parallel and combined at the block output [19]. Rather than relying on a single transformation stream, the architecture increases the number of parallel branches within the block structure, introducing cardinality as an additional architectural dimension alongside network depth and width. This multi-branch organization allows the network to capture a broader range of visual patterns while maintaining relatively efficient computational complexity. For scene recognition tasks, such feature diversity is beneficial because indoor environments often contain varied object arrangements, structural layouts, and contextual patterns that must be modeled simultaneously.

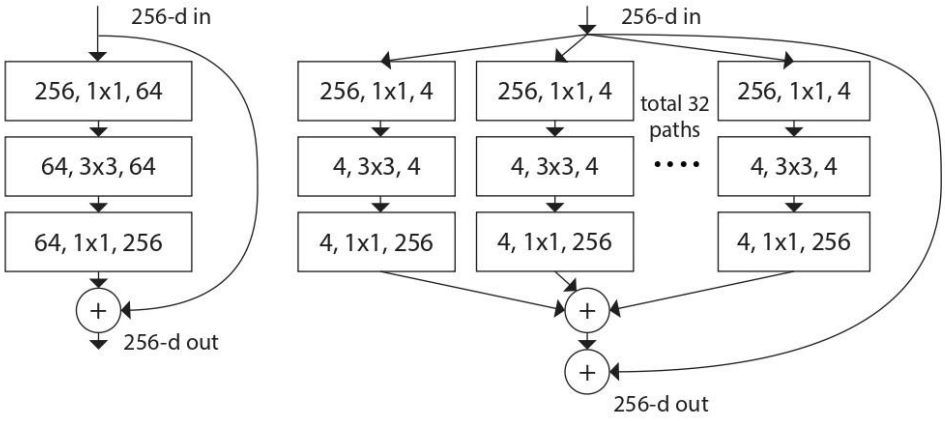

Fig 3.1. *Structural comparison of residual building blocks in ResNet and ResNeXt [19].*

From a representation-learning perspective, grouped residual transformations allow the network to capture multiple contextual patterns and structural characteristics in parallel [19]. This property is particularly relevant for scene recognition tasks, where understanding often depends on interpreting combinations of objects, layouts, background structures, and spatial relationships distributed across the scene. By increasing representational diversity, ResNeXt may provide richer visual feature maps for semantic-guided interaction during multimodal fusion.

Another important advantage of ResNeXt within semantic-aware learning is its ability to produce more diverse high-level convolutional responses while maintaining architectural simplicity. Since semantic guidance operates by modulating visual representations through attention-based interaction, the quality and diversity of the underlying RGB features directly influence the effectiveness of semantic fusion. Richer feature representations may allow semantic information to regulate a broader range of contextual patterns and scene structures during feature integration.

To ensure experimental fairness, both ResNet-50 and ResNeXt-50 are incorporated into the same semantic-guided framework without modifying downstream fusion modules, semantic encoders, or classification layers. The output dimensionality of the RGB branch

remains consistent across architectures, allowing the semantic-guided attention mechanism to operate identically during all experiments. Consequently, observed differences in recognition behavior can be attributed primarily to differences in convolutional representation structure rather than variations in training conditions or fusion strategy.

The comparative investigation between ResNet and ResNeXt therefore provides an opportunity to analyze how representational diversity influences semantic-guided scene understanding within convolutional architectures. Rather than evaluating backbone performance purely from a classification perspective, the study examines how architectural structure affects semantic-aware feature interaction and contextual representation learning in complex indoor environments.

3.4 SEMANTIC BRANCH AND CONTEXTUAL LEARNING

While the RGB branch captures appearance-based visual representations, indoor scene understanding also depends heavily on contextual semantic information related to object presence, spatial organization, and functional scene composition. Many indoor environments contain visually similar layouts and overlapping structural patterns, making semantic context increasingly important for distinguishing between scene categories. To incorporate such contextual information into the recognition process, the proposed framework employs a dedicated semantic branch designed to learn semantic-aware scene representations from segmentation-based inputs.

The semantic branch operates on precomputed semantic representations generated using an external semantic segmentation model. These semantic maps encode dense object-level contextual information across the scene and provide complementary cues beyond conventional RGB appearance features. Unlike raw visual inputs, semantic representations emphasize scene composition in terms of object categories and spatial arrangement, allowing the framework to incorporate higher-level contextual reasoning during scene understanding.

Within indoor environments, semantic context often plays a critical role in defining scene identity. For example, the presence and arrangement of objects such as beds, desks, shelves, appliances, or seating structures may provide strong indicators regarding the functional category of a scene. Semantic representations therefore help reduce ambiguity in situations where appearance-based features alone may not sufficiently distinguish visually similar environments.

The semantic branch follows a lightweight convolutional encoder design that progressively transforms the semantic input into high-level contextual feature representations. Initial convolutional operations capture local semantic structures and spatial object distributions, while deeper layers aggregate broader contextual relationships across the scene. Through this hierarchical encoding process, the semantic branch learns semantic feature maps that represent object co-occurrence patterns, contextual scene composition, and spatial semantic organization.

To further improve contextual feature learning, channel attention mechanisms are incorporated within the semantic encoder. Channel attention mechanisms allow the network to selectively strengthen semantically important feature channels while reducing the influence of less useful semantic activations [26], [27]. Through this adaptive weighting process, the semantic branch can prioritize object categories and contextual

cues that contribute more strongly to scene understanding during feature extraction. In indoor scene recognition, such selective feature refinement is particularly beneficial because different semantic entities contribute unevenly to scene discrimination. For example, certain objects and spatial cues may provide strong evidence for a specific indoor category, whereas others may introduce redundancy or ambiguity within the learned feature space.

Another important characteristic of the semantic branch is that it remains lightweight relative to the RGB backbone. The semantic encoder is not intended to replace visual representation learning, but rather to provide complementary contextual guidance during feature interaction. This design ensures that semantic information influences scene understanding without dominating the overall learning process. Instead, semantic features serve as contextual regulators that guide the emphasis of visual representations during semantic-aware fusion.

The semantic branch also plays an important role in maintaining multimodal consistency within the framework. Since semantic inputs preserve spatial organization derived from segmentation maps, the learned semantic feature maps remain spatially aligned with the convolutional visual representations extracted by the RGB branch. This alignment enables meaningful semantic-guided interaction during fusion, allowing contextual semantic cues to modulate corresponding visual structures effectively.

Importantly, the semantic branch architecture and processing pipeline remain fixed across all experiments conducted in this chapter. The same semantic encoder, attention configuration, and preprocessing strategy are used for both ResNet-50 and ResNeXt-50 based frameworks. As a result, differences in semantic-guided recognition performance can be analyzed primarily in relation to the RGB backbone architecture and its representation characteristics rather than changes in semantic feature extraction itself.

From a broader perspective, the semantic branch enables the framework to move beyond purely appearance-driven scene recognition toward context-aware representation learning. By integrating semantic object-level information with hierarchical convolutional visual features, the framework establishes a multimodal learning setting in which semantic context actively contributes to scene representation formation and contextual scene understanding.

3.5 ATTENTION-BASED SEMANTIC FUSION

The RGB and semantic branches of the proposed framework are integrated through an attention-based semantic fusion mechanism designed to combine appearance information with contextual semantic cues in a structured and adaptive manner. Instead of treating visual and semantic features as independent representations, the fusion module allows semantic information to actively regulate the emphasis of visual feature responses during scene representation learning [5], [20]. This design enables the framework to incorporate contextual understanding directly into the visual representation space.

In conventional multimodal fusion approaches, features from different modalities are often combined through direct concatenation or simple feature aggregation operations [7]. Although such methods preserve information from multiple sources, they may not effectively model the contextual relationship between semantic and visual representations. In indoor scene recognition, not all visual regions contribute equally to scene understanding, and the importance of a visual feature often depends on its semantic

relevance within the scene context. Consequently, the proposed framework adopts an attention-guided fusion strategy in which semantic information selectively modulates convolutional visual features.

Let the high-level feature representation extracted from the RGB branch be denoted by:

$$F_{rgb} \in \mathbb{R}^{C*H*W}$$

and the semantic feature representation generated by the semantic branch be represented as:

$$F_{sem} \in \mathbb{R}^{C_s*H*W}$$

where C and C_s denote the channel dimensions of the RGB and semantic feature maps respectively, while H and W represent the spatial dimensions. Since the two branches may produce feature maps with different channel dimensionalities, lightweight projection layers are first used to transform both representations into a shared feature space before fusion.

The semantic feature representation is subsequently passed through a sigmoid activation function to generate attention weights that encode the contextual importance of semantic responses across spatial locations and channels. The semantic attention map is expressed as:

$$A_{sem} = \sigma(F_{sem})$$

where σ denotes the sigmoid activation function. The resulting attention weights lie within the range $[0, 1]$, allowing the framework to regulate the contribution of semantic information adaptively during feature interaction [26], [27].

The final semantic-guided fused representation is obtained through element-wise modulation between the RGB feature maps and the semantic attention representation:

$$F_{fused} = F_{rgb} \odot A_{sem}$$

where $\odot$ represents element-wise multiplication. Through this operation, semantic information selectively emphasizes informative visual regions while suppressing less relevant feature responses. As a result, the fused representation becomes more context-aware and semantically discriminative for scene understanding [5], [20].

From a representation-learning perspective, this fusion mechanism allows semantic context to influence visual representation formation directly rather than being incorporated only at the final classification stage. Semantic guidance therefore acts as a contextual regulator over convolutional feature activations, enabling the model to focus on scene structures and object relationships that contribute meaningfully to scene identity [5]. This behaviour is particularly important in indoor environments where visually similar layouts may require subtle contextual distinctions for reliable recognition.

Another important characteristic of the proposed fusion strategy is that semantic guidance operates at a high-level representation stage rather than at low-level visual layers. Performing fusion at deeper convolutional stages allows semantic information to interact with semantically meaningful visual abstractions instead of raw textures or local appearance patterns [5]. Consequently, the attention mechanism captures broader contextual relationships associated with scene composition and object arrangement [26], [27].

The effectiveness of this fusion process also depends on the quality and diversity of the underlying RGB representations. Since semantic attention modulates convolutional feature responses directly, richer and more diverse visual representations may provide a stronger foundation for semantic-guided interaction. This aspect becomes particularly relevant when comparing ResNet and ResNeXt architectures within the same framework, as differences in representational diversity may influence how effectively semantic cues regulate visual feature learning.

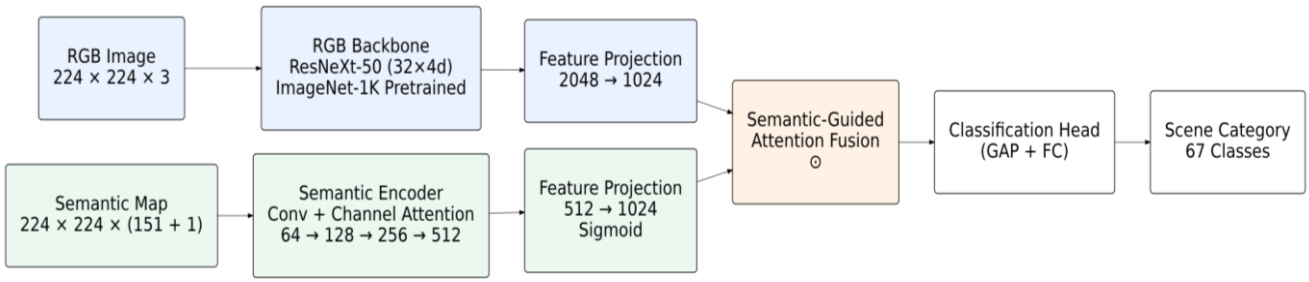

Fig 3.2. Overview of the proposed Multimodal Scene Classification Framework (adapted from the author's conference paper [P1]).

Importantly, the semantic fusion mechanism remains identical across all experiments conducted in this chapter. The same attention-guided fusion design is used for both ResNet-50 and ResNeXt-50 based frameworks, ensuring that observed performance differences arise primarily from differences in backbone representation structure rather than changes in multimodal fusion strategy. This controlled setup allows the study to isolate the influence of convolutional representation learning on semantic-guided scene understanding.

Overall, the attention-based semantic fusion mechanism provides an effective means of integrating contextual semantic information with convolutional visual representations. By enabling semantic guidance to selectively modulate visual feature activations, the framework establishes a context-aware representation learning process capable of capturing both appearance-based and semantic characteristics of complex indoor environments.

3.6 TRAINING STRATEGY AND EXPERIMENTAL SETUP

The proposed semantic-guided convolutional framework is evaluated using a controlled training strategy designed to ensure fair comparison between backbone architectures while maintaining stable multimodal optimization. Since the objective of this study is to analyze how convolutional representation structure influences semantic-guided scene understanding, all experiments are conducted under identical semantic fusion, optimization, and training conditions. The only architectural component varied during experimentation is the RGB backbone used within the visual branch.

3.6.1 Dataset Description

Experiments are conducted using the MIT Indoor-67 dataset, which is one of the most widely used benchmarks for indoor scene recognition. The dataset contains 15,620 RGB images distributed across 67 indoor scene categories, including environments such as

bookstores, libraries, classrooms, kitchens, corridors, auditoriums, and waiting rooms. Each category exhibits considerable variation in viewpoint, illumination, object arrangement, and scene composition, making the dataset particularly challenging for contextual scene understanding [24].

Following the standard evaluation protocol, each category is divided into 80 training images and 20 testing images. Performance is evaluated using Top-1 classification accuracy, which measures the percentage of test images for which the predicted scene category matches the ground-truth label.

The dataset is particularly suitable for semantic-aware scene recognition because indoor scenes often contain overlapping object distributions and strong contextual dependencies. Scene categories are frequently distinguished by object relationships and spatial organization rather than isolated visual patterns alone. Consequently, the benchmark provides an effective setting for analyzing semantic-guided representation learning.

3.6.2 Semantic Input Generation

Since the MIT Indoor-67 dataset does not provide semantic annotations directly, semantic representations are generated using a pretrained semantic segmentation model trained on the ADE20K dataset. The segmentation network remains fixed throughout all experiments and is used solely to produce dense semantic maps corresponding to each RGB image.

The generated semantic maps encode contextual object-level information across the scene and serve as inputs to the semantic branch of the framework. These representations provide complementary contextual cues regarding object categories, spatial structure, and semantic scene composition. Importantly, the segmentation model itself is not optimized during training, ensuring that the investigation focuses specifically on semantic-guided scene recognition rather than segmentation learning.

3.6.3 Data Preprocessing and Augmentation

To improve generalization and robustness during training, several preprocessing and augmentation strategies are applied to the RGB images. Input images are resized and randomly cropped to the required spatial resolution before being normalized using ImageNet mean and standard deviation values. Random horizontal flipping and appearance-based augmentations such as brightness variation, contrast adjustment, Gaussian blur, and multiplicative noise are applied during training to simulate realistic visual variability.

The semantic maps undergo the same spatial transformations as the RGB images in order to preserve alignment between visual and semantic representations. Additional coarse dropout augmentation is applied to the semantic inputs to improve robustness against noisy or incomplete semantic information.

During validation and testing, random augmentations are disabled. Images are resized and center-cropped using consistent evaluation settings, and final performance is reported using standard Top-1 accuracy evaluation.

3.6.4 Staged Training Strategy

The framework is trained using a staged transfer learning strategy to stabilize multimodal optimization and improve semantic-aware feature interaction. By initializing the RGB backbone with weights pretrained on large-scale visual datasets, the model can reuse generalized visual features while adapting them to the indoor scene recognition task [28].

The RGB branch is initialized using ImageNet-pretrained weights for both ResNet-50 and ResNeXt-50 architectures. The semantic branch and fusion modules are trained using task-specific semantic-aware scene recognition data.

Training is performed in three stages:

Stage 1: RGB Branch Pretraining

During the initial stage, only the RGB backbone and its auxiliary classifier are trained while the semantic branch and fusion layers remain frozen. This stage allows the convolutional backbone to adapt pretrained visual representations to the indoor scene recognition task.

Stage 2: Semantic Branch Pretraining

The semantic branch is subsequently trained independently while keeping the RGB backbone frozen. This stage enables the semantic encoder to learn contextual semantic representations from segmentation-based inputs without interference from multimodal fusion dynamics.

Stage 3: Joint Fusion Training and Fine-Tuning

After branch-level pretraining, the complete framework is trained jointly. During this phase, semantic-guided attention fusion modules and classification layers are optimized together with the pretrained backbone representations. Lower learning rates are used for pretrained backbone parameters, while relatively higher learning rates are applied to newly initialized fusion layers and classification components.

This staged optimization strategy helps stabilize multimodal interaction during training and reduces the risk of noisy semantic gradients disrupting pretrained visual representations during early learning stages.

3.6.5 Optimization Settings

The framework is optimized using the AdamW optimizer with weight decay regularization to improve training stability and generalization. Mini-batch training is performed using a batch size of 32. Learning rate scheduling is applied during optimization to ensure gradual convergence of the multimodal framework.

The model is trained using categorical cross-entropy loss for scene classification:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i denotes the ground-truth scene label and $\hat{y}_i$ represents the predicted probability for the corresponding scene category [29].

Mixed-precision training is employed to reduce memory consumption and improve computational efficiency during optimization. Gradient clipping is additionally applied to maintain stable optimization behavior during multimodal training.

3.6.6 Experimental Consistency

An important aspect of the experimental design is that all configurations remain identical across backbone comparisons. The semantic branch architecture, semantic-guided fusion mechanism, training schedule, optimization settings, preprocessing pipeline, and evaluation protocol remain unchanged for both ResNet-50 and ResNeXt-50 experiments.

This controlled setup ensures that any observed differences in recognition performance can be attributed primarily to differences in convolutional representation structure rather than variations in training methodology or multimodal fusion design. Consequently, the experimental framework provides a reliable basis for analyzing how representational diversity and aggregated residual transformations influence semantic-guided scene understanding in indoor environments.

3.7 RESULTS AND DISCUSSION

This section presents the experimental results obtained using the proposed semantic-guided convolutional framework on the MIT Indoor-67 dataset. The primary objective of the experiments is not only to evaluate recognition performance, but also to analyze how convolutional representation structure influences semantic-guided feature learning under identical multimodal training conditions. To ensure a controlled investigation, all experiments are conducted using the same semantic branch, attention-based fusion strategy, optimization settings, and evaluation protocol while varying only the RGB backbone architecture.

3.7.1 Quantitative Performance Comparison

Table 3.1 presents the comparative performance of the investigated convolutional architectures along with selected related approaches evaluated on the MIT Indoor-67 dataset.

Table 3.1

Comparison of Scene Recognition Performance on MIT Indoor-67 Dataset

Category	Architecture	Configuration	Parameters	Top-1 Accuracy
Baseline	ResNet-50	*resnet50*	85M	87.01
	ResNet-18	*resnet18*	47M	85.58

Ours	ConvNeXtV2	*convnextv2_base*	115M	85.97
	ResNeXt-50	*resnext50_32x4d*	60M	**91.30**

The experimental results show that the proposed ResNeXt-50 based semantic-guided framework achieves the highest Top-1 accuracy of 91.30% on the MIT Indoor-67 dataset. Compared to the ResNet-50 baseline operating under identical semantic-aware learning conditions, the ResNeXt-based model achieves an absolute improvement of more than 4%. This improvement is particularly significant because the overall semantic branch, fusion mechanism, training procedure, and optimization strategy remain unchanged across experiments.

The results indicate that the observed performance gains arise primarily from differences in convolutional representation structure rather than changes in semantic processing or multimodal fusion design. In particular, the increased representational diversity introduced by aggregated residual transformations appears to improve semantic-guided scene representation learning substantially.

Another important observation is that the ResNeXt-50 based framework achieves superior performance while maintaining lower parameter complexity compared to several larger architectures. Although ConvNeXtV2 and ViT-B/16 contain substantially larger parameter counts, their performance remains below that of the proposed ResNeXt-based semantic-guided framework. This suggests that architectural organization and representation quality may influence semantic-aware scene recognition more strongly than model scale alone.

3.7.2 Effect of Representational Diversity

One of the central observations emerging from the experiments is the importance of representational diversity in semantic-guided scene understanding. Indoor scene recognition requires the simultaneous interpretation of multiple contextual components, including object arrangements, background structures, spatial layout, and semantic relationships among scene elements. Standard residual architectures primarily learn a single transformation pathway within each residual block, which may limit the variety of contextual patterns captured at a fixed network depth.

In contrast, ResNeXt introduces grouped residual transformations that allow multiple parallel feature transformations to be learned simultaneously. From a representation-learning perspective, this design increases the diversity of visual patterns captured within the convolutional representation space. Such diversity appears particularly beneficial for semantic-guided fusion because semantic information interacts directly with high-level visual representations during attention-based modulation.

The results suggest that richer convolutional representations provide a stronger foundation for semantic-guided feature interaction. Since semantic attention selectively emphasizes contextual visual patterns, architectures capable of producing more diverse feature activations may allow semantic information to regulate scene understanding more effectively. This behavior may explain the consistent improvement observed when replacing ResNet-50 with ResNeXt-50 under otherwise identical conditions.

Another possible explanation is that grouped residual transformations improve the network's ability to capture heterogeneous scene structures distributed across the image. Indoor environments frequently contain multiple coexisting semantic regions, object clusters, and contextual substructures. Learning parallel transformations may therefore help the network represent diverse scene components more effectively before semantic modulation occurs.

3.7.3 Semantic-Guided Feature Interaction

The experiments further demonstrate that semantic guidance contributes meaningfully to scene understanding when integrated with high-quality convolutional representations. The attention-based semantic fusion mechanism allows semantic context to regulate visual feature activations adaptively, enabling the framework to emphasize semantically informative scene regions during representation learning.

This behavior becomes particularly important in indoor environments where visually similar layouts may correspond to different semantic scene categories. Contextual semantic information derived from segmentation maps helps the framework distinguish between scenes that share overlapping textures or structural appearance but differ in functional composition and object arrangement.

The stronger performance observed with the ResNeXt backbone also suggests that semantic-guided interaction depends partly on the structure and quality of the visual representation space itself. Semantic supervision does not operate independently of the backbone architecture; rather, its effectiveness appears closely tied to the richness and contextual abstraction capability of the underlying convolutional features.

3.7.4 Comparative Architectural Interpretation

The comparison between ResNet-50 and ResNeXt-50 highlights an important architectural insight regarding semantic-aware scene recognition. Increasing representational diversity through cardinality appears more effective for semantic-guided contextual learning than simply increasing network depth or parameter count alone.

This observation aligns with the broader representation-learning perspective developed throughout the thesis. Scene understanding requires capturing multiple contextual relationships simultaneously, and architectures capable of modeling diverse feature transformations may provide more effective support for semantic-aware interaction. In this setting, semantic guidance acts not merely as auxiliary information, but as a contextual regulator whose effectiveness depends strongly on the representation characteristics of the backbone architecture.

The experiments therefore suggest that semantic-aware scene recognition should not be viewed solely as a multimodal fusion problem. Instead, the structure of the learned representation space itself plays an important role in determining how effectively semantic context can influence scene understanding.

3.7.5 Confusion Analysis and Failure Cases

Although the proposed framework demonstrates strong overall performance, certain scene categories continue to exhibit noticeable confusion. Most misclassifications occur between environments that share similar semantic composition or spatial organization, such as bookstores and libraries, waiting rooms and auditoriums, or toy stores and children's rooms.

These failure cases indicate that even with semantic guidance, distinguishing highly correlated indoor environments remains challenging when scenes contain overlapping object distributions and similar structural layouts. In some cases, semantic segmentation errors may also propagate into the fusion stage, reducing the effectiveness of semantic-guided modulation.

Another limitation is that convolutional representations, despite strong hierarchical learning capability, may still struggle to model very long-range spatial dependencies across large indoor environments. While semantic guidance improves contextual understanding, the framework remains constrained by the locality-oriented nature of convolutional feature extraction. This observation becomes particularly important when considering scene categories involving large spatial extent or complex global structure.

3.7.6 Discussion Summary

Overall, the experimental results demonstrate that convolutional backbone architecture plays a substantial role in semantic-guided indoor scene recognition. The findings suggest that increasing representational diversity through grouped residual transformations improves semantic-aware feature learning more effectively than standard residual representations under the investigated framework.

More importantly, the experiments reveal that the effectiveness of semantic supervision depends not only on the availability of semantic information, but also on the representation quality and structural characteristics of the backbone architecture itself. Richer convolutional representations enable stronger semantic-guided interaction and more effective contextual scene understanding.

These observations motivate the subsequent investigation presented in the next chapter, where the analysis is extended beyond convolutional architectures toward transformer-based representation paradigms in order to examine how semantic guidance interacts with fundamentally different visual representation structures.

3.8 LIMITATIONS AND OBSERVATIONS

Although the proposed semantic-guided convolutional framework demonstrates strong performance on indoor scene recognition, several limitations remain that provide important insight into the behavior of semantic-aware convolutional representation learning. Analyzing these limitations is useful not only for understanding the constraints of the current framework, but also for identifying broader architectural challenges associated with semantic-guided scene understanding.

One important limitation arises from the dependence on precomputed semantic segmentation maps. Since the semantic representations are generated using an external segmentation model that remains fixed during training, errors in semantic prediction directly influence the quality of semantic-guided feature interaction. Incorrect object

labels, incomplete segmentation boundaries, or missing contextual regions may reduce the effectiveness of semantic attention during fusion. In scenes containing cluttered layouts or visually ambiguous structures, such segmentation inaccuracies can propagate into the multimodal representation space and affect final scene classification performance.

Another limitation is related to the difficulty of distinguishing scene categories with highly overlapping semantic composition. Several indoor environments share common object distributions and similar structural organization, making fine-grained contextual discrimination challenging even under semantic-aware learning. For example, bookstores and libraries often contain visually similar shelves and book arrangements, while waiting rooms and auditoriums may share comparable seating structures and spatial layouts. Although semantic guidance improves contextual representation learning, the framework still exhibits confusion in scenes where object presence alone is insufficient for reliable discrimination.

The experiments also suggest that convolutional architectures remain influenced by locality-oriented representation learning despite the use of semantic-guided attention. Hierarchical convolutional feature extraction is highly effective for capturing structured spatial patterns and local contextual information; however, modeling long-range scene relationships across large indoor environments remains comparatively difficult. Scene understanding often requires reasoning about spatial dependencies distributed across distant image regions, particularly in environments with complex layout organization. While semantic guidance improves contextual emphasis, convolutional representations may still have limited ability to model broad global relationships compared to architectures designed explicitly for long-range interaction.

An additional observation emerging from the experiments is that semantic supervision does not contribute equally across all backbone architectures. The comparative analysis between ResNet-50 and ResNeXt-50 indicates that the effectiveness of semantic-guided learning depends strongly on the representational quality and diversity of the underlying visual features. Richer convolutional representations appear to provide a more suitable foundation for semantic modulation and contextual feature interaction. This observation reinforces the broader thesis perspective that semantic-aware scene recognition is closely connected to representation structure rather than semantic fusion alone.

The results further indicate that increasing representational diversity through grouped residual transformations improves semantic-guided scene understanding more effectively than simply increasing model size or parameter count. The stronger performance of ResNeXt-50 compared to several larger architectures suggests that architectural organization and feature transformation diversity may play a more important role in semantic-aware representation learning than scaling network complexity alone.

At the same time, the experiments reveal that semantic-guided convolutional learning benefits from preserving structured spatial representations throughout the feature hierarchy. Semantic attention interacts naturally with hierarchical convolutional feature maps because both representations maintain explicit spatial organization across intermediate stages. This observation becomes particularly important when considering alternative representation paradigms where spatial continuity may be represented differently.

Collectively, these findings motivate a broader investigation into how semantic supervision behaves across fundamentally different representation architectures. While convolutional networks provide strong hierarchical spatial representations for semantic-

guided scene understanding, the observed limitations related to long-range contextual reasoning and representation diversity suggest the need to explore architectures capable of modeling contextual interaction differently.

Motivated by these observations, the next chapter extends the investigation toward transformer-based architectures in order to analyze how semantic guidance interacts with token-based and hierarchical transformer representations. In particular, the study examines whether representation alignment between semantic information and visual representation structure influences semantic-guided scene understanding more fundamentally across modern deep learning architectures.

CHAPTER 4

REPRESENTATION-ALIGNED TRANSFORMER FRAMEWORK

4.1. INTRODUCTION

The experimental observations presented in the previous chapter demonstrated that semantic-guided scene recognition is influenced not only by the availability of semantic information, but also by the representation characteristics of the underlying backbone architecture. In particular, the comparative analysis between ResNet-50 and ResNeXt-50 showed that increased representational diversity improves semantic-guided feature interaction and contextual scene understanding within convolutional frameworks. These findings suggest that semantic supervision and representation structure are closely interconnected during multimodal scene representation learning.

Although convolutional neural networks provide strong hierarchical spatial representations for indoor scene recognition, they remain fundamentally locality-oriented in their representation behavior. Convolutional feature extraction progressively aggregates local contextual information through stacked receptive fields, which may limit direct modeling of long-range dependencies across large indoor environments. Indoor scene understanding often requires interpreting relationships between spatially distant regions, object groupings, and contextual structures distributed throughout the scene. Consequently, architectures capable of modeling broader contextual interactions may provide alternative advantages for semantic-guided learning.

4.2. TRANSFORMER REPRESENTATION STRUCTURES

Transformer-based architectures introduce a fundamentally different representation-learning paradigm compared to conventional convolutional neural networks. While CNNs primarily learn hierarchical spatial representations through localized convolutional operations, transformers model visual information through attention-based interactions between image tokens. These differences influence not only how contextual relationships are captured within the network, but also how semantic information interacts with learned visual representations during scene understanding.

Vision Transformers process visual inputs by converting image patches into sequential embedding representations [8]. The input image is partitioned into fixed-size non-overlapping patches, and each patch is linearly mapped into a token embedding vector. Positional embeddings are then incorporated to retain spatial ordering information before the token sequence is passed through transformer encoder layers composed of multi-head self-attention and feed-forward operations.

Let the sequence of input patch embeddings be represented as:

$$X = [x_1, x_2, \ldots\ldots x_N]$$

where N denotes the number of image patches and x_i represents the embedding corresponding to the i^{th} image patch. Through self-attention, each token can interact with all other tokens in the sequence, enabling the model to capture long-range contextual relationships across spatially distant image regions [30].

The self-attention operation can be expressed as:

$$Attention(Q, K, V) = Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V denote the query, key, and value matrices derived from the token embeddings, and d_k represents the dimensionality of the key vectors. Through this mechanism, transformer architectures dynamically model contextual interactions between image regions irrespective of spatial distance [30].

From a representation-learning perspective, Vision Transformers emphasize global contextual interaction rather than localized spatial continuity [8]. Unlike convolutional feature maps, token representations do not explicitly preserve neighborhood structure across intermediate representation stages. Spatial relationships are instead modeled implicitly through positional embeddings and learned attention patterns [1], [8]. This design provides strong flexibility in contextual modeling but may reduce preservation of fine-grained spatial locality important for certain scene understanding tasks [9].

Indoor scene recognition often depends heavily on spatial arrangement and structural continuity within the environment [24]. Relationships among objects, furniture layouts, and scene organization patterns contribute significantly to scene identity. Consequently, purely token-based representations may face challenges in preserving locality-sensitive contextual information required for detailed semantic scene understanding.

To address this limitation, hierarchical transformer architectures such as the Swin Transformer incorporate spatially structured feature learning within the transformer framework [9]. Rather than applying global self-attention across all image tokens at every layer, the Swin architecture computes attention within smaller spatial windows and progressively builds multi-level feature representations across successive stages.

This formulation preserves spatial organization more effectively than standard Vision Transformers. Window-based attention maintains interaction among neighboring visual regions, while the shifted-window mechanism facilitates communication between adjacent windows in deeper layers of the network. Consequently, Swin Transformer combines the long-range modeling capability of transformers with a progressively organized spatial feature hierarchy resembling convolutional feature extraction.

From a semantic-aware learning perspective, these representation differences are highly significant. Semantic information derived from segmentation maps is inherently spatial and context-dependent. Hierarchical spatial representations may therefore provide more natural alignment for semantic interaction because semantic structure and visual feature organization remain spatially consistent throughout the representation hierarchy. In contrast, token-based transformer architectures may require semantic information to be converted into compatible token-level representations before meaningful interaction can occur.

Another important distinction lies in the balance between locality preservation and global contextual modeling. Vision Transformers prioritize flexible long-range interaction across all tokens, whereas hierarchical transformers maintain stronger spatial continuity while still enabling contextual aggregation across representation stages. These

representational properties may influence how effectively semantic guidance regulates scene understanding within different transformer architectures.

Consequently, transformer-based scene recognition should not be viewed as a single unified representation paradigm. Different transformer architectures organize visual information differently, preserve contextual structure differently, and may therefore interact with semantic supervision differently. Understanding these representation characteristics is essential for developing semantic-aware transformer frameworks capable of effectively integrating contextual semantic information with visual representations during indoor scene understanding.

4.3. UNIFIED SEMANTIC-AWARE TRANSFORMER FRAMEWORK

To investigate how semantic supervision interacts with different transformer architectures, a unified semantic-aware transformer framework is developed for indoor scene recognition. The framework maintains a consistent multimodal learning setting across transformer models while adapting semantic encoding and fusion operations according to the internal feature organization of the underlying backbone. This controlled setup allows systematic analysis of how token-based and hierarchical transformers respond to semantic feature interaction during scene understanding.

The overall framework follows a dual-stream architecture consisting of a visual transformer branch and a semantic encoding branch. The visual branch learns scene-related features from RGB images, while the semantic branch processes contextual information derived from segmentation maps. Unlike the convolutional framework presented in the previous chapter, semantic integration within the transformer setting is explicitly adapted to match the feature organization used by each backbone architecture.

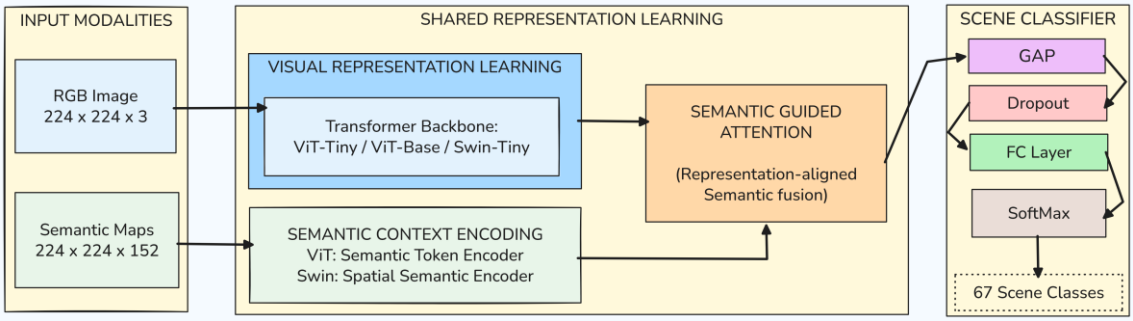

Fig 4.1. *Overview of the proposed Unified Semantic-aware Transformer Framework(adapted from the author's conference paper [P2]).*

The framework is evaluated using two representative transformer paradigms: Vision Transformers (ViTs) and hierarchical Swin Transformers. Vision Transformers process images as sequences of token embeddings and primarily model global contextual interaction through self-attention. In contrast, Swin Transformers preserve multi-level spatial organization through localized window attention and progressive feature aggregation. Since these architectures organize visual information differently, a single semantic fusion strategy may not produce equally meaningful interaction across both paradigms.

To address this issue, the framework introduces architecture-aware semantic encoding strategies tailored to each transformer type. For Vision Transformers, semantic

information is converted into token-compatible embeddings that interact naturally with patch-token features through attention-based fusion. For Swin Transformers, semantic representations preserve spatial feature organization to maintain compatibility with hierarchical visual feature maps across multiple stages.

The overall transformer framework consists of four major components:

1. RGB transformer backbone
2. Semantic representation encoder
3. Architecture-aware semantic fusion module
4. Scene classification head

The RGB transformer backbone extracts scene-related visual features from the input image using either a Vision Transformer or Swin Transformer architecture. The semantic branch processes segmentation-derived semantic inputs and generates contextual semantic features adapted to the organization of the corresponding backbone architecture. These features are subsequently integrated through semantic-aware fusion modules designed specifically for each transformer paradigm.

An important characteristic of the proposed framework is that the overall semantic learning strategy remains consistent across all experiments. Semantic supervision is not treated merely as auxiliary information appended at the classification stage. Instead, semantic information participates directly in intermediate feature interaction within the visual feature space. This design allows the study to focus on how architectural organization influences semantic integration behavior across transformer models.

Another important aspect of the framework is experimental consistency. The dataset, semantic input generation pipeline, optimization strategy, training schedule, and evaluation protocol remain fixed across all transformer experiments. Only the backbone architecture and the corresponding semantic encoding strategy are varied. As a result, differences in recognition performance and semantic interaction behavior can be analyzed primarily with respect to architectural organization rather than unrelated training variations.

From a broader architectural perspective, the unified framework establishes a controlled setting for examining whether semantic supervision becomes more effective when semantic information is organized in a manner compatible with the internal feature structure of the visual backbone. This question forms the central conceptual focus of the transformer-based investigation presented in this chapter.

4.4. SEMANTIC TOKEN ENCODING FOR VISION TRANSFORMERS

Vision Transformers process visual information as sequences of patch-level token embeddings rather than hierarchical spatial feature maps [8]. Consequently, semantic information represented in conventional spatial form cannot interact directly with transformer token representations without an appropriate representation conversion mechanism. To enable meaningful semantic-guided learning within the Vision Transformer framework, semantic representations must therefore be transformed into token-compatible embeddings aligned with the structure of the visual token space.

In the proposed framework, semantic maps generated from the segmentation model are first partitioned into non-overlapping patches using the same spatial partitioning strategy applied to the RGB image. Each semantic patch is subsequently projected into a semantic

token embedding through a learnable linear projection layer. This process converts spatial semantic information into a sequential token representation compatible with the transformer architecture.

Let the semantic token sequence be represented as:

$$S = [s_1, s_2, \ldots\ldots s_N]$$

where N denotes the number of semantic patches and s_i represents the embedding corresponding to the i^{th} semantic patch. Similarly, the visual token sequence extracted from the RGB image is represented as:

$$V = [v_1, v_2, \ldots\ldots\ldots v_N]$$

where v_i denotes the visual embedding associated with the corresponding image patch.

By converting semantic maps into token representations, semantic information becomes structurally compatible with the transformer representation space. This alignment allows semantic and visual tokens to interact naturally during attention-based fusion without violating the sequential processing paradigm of the Vision Transformer architecture.

The semantic token embeddings are integrated with visual tokens through cross-attention-based semantic interaction. In this mechanism, visual tokens attend to semantic token representations in order to incorporate contextual semantic information during feature learning. The cross-attention operation can be formulated as:

$$CrossAttention(Q_v, K_s, V_s) = Soft\max\left(\frac{Q_v K_s^T}{\sqrt{d_k}}\right) V_s$$

where Q_v represents queries generated from visual tokens, while K_s and V_s denote semantic token keys and values respectively. Through this operation, semantic context influences visual representation learning by guiding attention toward semantically informative relationships within the token space [1].

From a representation-learning perspective, semantic token encoding allows semantic supervision to operate directly within the transformer's native representation format. Instead of forcing spatial semantic maps into incompatible feature structures, the framework adapts semantic information according to the architectural organization of the backbone representation. This representation alignment is important because Vision Transformers process contextual relationships primarily through token-level interaction rather than explicit spatial feature hierarchies [8].

Another important advantage of semantic token encoding is that it preserves global contextual flexibility within the transformer representation space. Since each visual token can attend to all semantic tokens through cross-attention, the framework enables semantic interaction across distant image regions. This property allows the model to capture long-range semantic dependencies and contextual relationships that may span multiple parts of the scene.

However, token-based semantic interaction also introduces certain representational challenges. Unlike hierarchical spatial representations, token sequences do not explicitly preserve local spatial continuity throughout intermediate representation stages. Although positional embeddings encode spatial ordering information, semantic structure becomes represented more implicitly through learned attention relationships. As a result, fine-

grained locality-sensitive semantic interaction may become more difficult to preserve consistently during deeper transformer processing.

Despite these challenges, semantic token encoding provides a principled mechanism for integrating semantic supervision into Vision Transformer architectures while maintaining representation compatibility. The framework therefore enables investigation of how token-based transformer representations utilize semantic context during indoor scene understanding and how semantic-guided interaction behaves within globally contextual but structurally sequential representation spaces.

4.5. SPATIAL SEMANTIC ENCODING FOR SWIN TRANSFORMERS

Unlike standard Vision Transformers, Swin Transformers preserve hierarchical spatial organization throughout intermediate representation stages while simultaneously incorporating attention-based contextual modelling [9]. This hierarchical structure provides an important advantage for semantic-aware scene recognition because semantic information derived from segmentation maps is inherently spatial and context-dependent. Consequently, semantic representations can be integrated more naturally when spatial continuity and feature locality are preserved within the backbone representation space.

To leverage this property, the proposed framework employs a spatial semantic encoding strategy specifically designed for hierarchical transformer architectures. Instead of converting semantic information into token sequences as performed for Vision Transformers, semantic representations are maintained in structured spatial form to preserve alignment with the hierarchical feature organization of the Swin Transformer backbone.

The semantic branch processes segmentation-derived semantic maps using a lightweight hierarchical encoder that progressively extracts semantic feature representations across multiple spatial stages. These semantic feature maps maintain explicit spatial organization and remain aligned with the corresponding Swin Transformer feature hierarchy throughout the network.

Let the hierarchical semantic feature representation at stage ℓ be represented as:

$$S^{(l)} \in \mathbb{R}^{C_l \cdot H_l . W_l}$$

where $C_l,\ H_l,\ and\ W_l$ denote the channel dimension, height, and width of the semantic feature map at the corresponding hierarchical stage. Similarly, the visual feature representation extracted by the Swin Transformer backbone at stage l is represented as:

$$V^{(l)} \in \mathbb{R}^{C_l \cdot H_l \cdot W_l}$$

Because both semantic and visual representations preserve compatible hierarchical spatial organization, semantic interaction can occur directly across aligned feature structures without requiring conversion into token-only representations.

Semantic guidance is incorporated through spatial attention modulation applied at intermediate representation stages. The semantic feature maps are transformed into semantic attention masks that regulate visual feature responses across spatial locations. The semantic attention representation is formulated as:

$$A_{sem}^{(l)} = \sigma\left(S^{(l)}\right)$$

where σ denotes the sigmoid activation function applied element-wise to generate normalized semantic attention weights.

The final semantically modulated visual representation is obtained through spatially aligned feature interaction:

$$F_{fused}^{(l)} = V^{(l)} \bigodot A_{sem}^{(l)}$$

where ⊙ denotes element-wise multiplication between visual features and semantic attention masks. This formulation enables semantic context to regulate visual feature activations while preserving spatial continuity throughout the representation hierarchy. Since semantic and visual representations remain spatially aligned at every stage, the fusion mechanism supports locality-sensitive contextual interaction that is particularly important for indoor scene understanding tasks involving spatial layout reasoning and object arrangement interpretation [26], [27].

Another important advantage of spatial semantic encoding is that semantic structure remains explicitly preserved during representation learning. Unlike token-based transformer representations where spatial relationships become encoded implicitly through attention interactions, hierarchical spatial representations maintain structured locality across intermediate feature stages. This property allows semantic guidance to operate more directly on scene layouts, contextual regions, and object configurations during multimodal fusion.

The shifted window attention mechanism of Swin Transformer further enhances semantic-guided interaction by enabling contextual communication across neighboring spatial regions while maintaining computational efficiency. As feature hierarchies deepen, the model progressively captures broader contextual dependencies without completely discarding local spatial continuity. This balance between locality preservation and contextual aggregation provides a representation environment well-suited for semantic-aware scene understanding [5], [9].

From a representation-learning perspective, spatial semantic encoding demonstrates how semantic supervision can be adapted according to the representation structure of the backbone architecture. Rather than enforcing a generic semantic fusion strategy across all transformer paradigms, the proposed framework aligns semantic interaction with the native organization of the representation space itself. This representation compatibility becomes particularly important for indoor scene recognition, where semantic structure and spatial organization are closely interconnected.

Overall, the spatial semantic encoding strategy enables Swin Transformers to integrate semantic context while preserving hierarchical spatial continuity throughout feature learning. This design provides an effective mechanism for studying how semantic supervision interacts with hierarchical transformer representations during contextual scene understanding.

4.6 REPRESENTATION-ALIGNED SEMANTIC FUSION

The central principle underlying the proposed transformer framework is that semantic supervision should be aligned with the native representation structure of the underlying

architecture. Rather than treating semantic information as a generic auxiliary modality processed identically across all models, the framework adapts semantic encoding and fusion according to how visual information is internally organized within the backbone representation space. This concept forms the core representation-learning perspective of the present thesis.

Existing semantic-aware scene recognition approaches commonly employ similar semantic fusion strategies across different architectures without explicitly considering representation compatibility. However, Vision Transformers and hierarchical Swin Transformers organize visual information fundamentally differently. Vision Transformers process images as sequential token embeddings emphasizing global contextual interaction, whereas Swin Transformers preserve hierarchical spatial structure throughout intermediate representation stages. Since semantic information itself is inherently structured and spatially contextual, the effectiveness of semantic-guided learning may depend strongly on how naturally semantic representations align with these representation paradigms.

The proposed framework therefore introduces representation-aligned semantic fusion, where semantic interaction mechanisms are tailored specifically to the representation structure of the backbone architecture. For token-based Vision Transformers, semantic maps are transformed into token-compatible semantic embeddings that interact through cross-attention within the sequential representation space. For hierarchical Swin Transformers, semantic information is maintained in structured spatial form to preserve locality-sensitive alignment with hierarchical feature maps.

From a conceptual perspective, representation alignment seeks to minimize structural incompatibility between semantic information and visual representation organization during multimodal fusion. If semantic structure and visual representation format remain naturally compatible, semantic supervision can regulate contextual learning more effectively and consistently throughout the network.

For Vision Transformers, semantic-guided fusion operates primarily through token-level contextual interaction. Visual tokens attend globally to semantic token embeddings, allowing semantic information to influence representation learning through attention-based relational modeling. This mechanism enables broad contextual semantic interaction across distant image regions. However, since spatial continuity is represented implicitly within token space, semantic structure may become less explicitly localized during deeper representation stages.

In contrast, Swin Transformer fusion preserves spatial semantic alignment directly within hierarchical feature maps. Semantic attention masks operate on spatially organized visual representations across multiple stages, enabling semantic context to regulate localized scene structures while still supporting broader contextual aggregation through shifted window attention. This representation alignment preserves both semantic locality and hierarchical contextual continuity simultaneously.

The distinction between token-space fusion and spatial semantic fusion can therefore be interpreted as a broader difference between implicit and explicit structural semantic interaction. Token-based representations rely primarily on learned relational attention patterns to capture semantic relationships, whereas hierarchical spatial representations maintain direct structural correspondence between semantic and visual feature organization.

Another important implication of representation-aligned fusion is that semantic supervision becomes architecture-dependent rather than universally transferable through identical fusion operations. The experiments presented later in this chapter demonstrate that semantic interaction quality depends not only on the availability of semantic information, but also on whether the representation structure of the architecture naturally supports the type of semantic organization being introduced during multimodal learning.

This perspective extends beyond simple multimodal fusion and contributes toward a more general understanding of semantic-aware representation learning. Semantic supervision should not be viewed solely as additional information appended to a backbone architecture. Instead, effective semantic-guided learning requires considering how semantic structure interacts with contextual modeling behavior, locality preservation, feature hierarchy organization, and representation continuity within the architecture itself.

From a broader scene-understanding perspective, indoor environments contain complex contextual dependencies involving both long-range semantic relationships and fine-grained spatial organization. Architectures that preserve hierarchical spatial continuity while supporting contextual interaction may therefore provide more suitable representation environments for semantic-guided scene understanding than architectures relying purely on globally interacting token sequences.

The representation-aligned semantic fusion framework proposed in this thesis therefore establishes a conceptual foundation for analyzing semantic-guided learning across modern deep neural architectures. By adapting semantic interaction according to representation structure, the framework enables systematic investigation of how different representation paradigms influence semantic-aware indoor scene recognition performance and contextual feature learning behavior.

4.7 TRAINING STRATEGY AND EXPERIMENTAL SETUP

The proposed representation-aligned transformer framework is evaluated using a controlled experimental setup designed to analyze how semantic supervision interacts with different transformer representation structures under consistent learning conditions. To ensure fair comparison across architectures, all experiments are conducted using the same dataset, semantic input generation process, preprocessing pipeline, optimization strategy, and evaluation protocol. The primary variables investigated are the transformer backbone architecture and the corresponding semantic alignment strategy employed during multimodal fusion.

4.7.1 Dataset and Evaluation Protocol

Experiments are conducted using the MIT Indoor-67 dataset, which contains 67 indoor scene categories characterized by significant contextual complexity, object overlap, and structural variation [24]. The dataset includes a wide variety of indoor environments such as bookstores, libraries, laboratories, classrooms, corridors, kitchens, and waiting rooms. Since many categories share similar visual appearance and object composition, the benchmark provides a challenging evaluation setting for semantic-aware scene understanding.

The standard dataset split is adopted, with 80 training images and 20 testing images per category [24]. Model performance is evaluated using Top-1 classification accuracy on the

test set. All transformer architectures are evaluated under identical data splits and training conditions to ensure experimental consistency.

4.7.2. Semantic Representation Generation

Semantic inputs are generated using a pretrained semantic segmentation model trained on the ADE20K dataset [31]. The segmentation model remains fixed throughout all experiments and is used only to produce dense semantic maps corresponding to the RGB images. These semantic representations encode object-level contextual information and spatial scene composition, providing complementary semantic guidance during transformer-based scene representation learning.

The semantic generation pipeline remains identical across all transformer experiments. However, the semantic encoding strategy differs according to the representation structure of the backbone architecture. For Vision Transformers, semantic maps are converted into semantic token embeddings compatible with sequential token processing. For Swin Transformers, semantic information is preserved in hierarchical spatial form to maintain alignment with structured feature hierarchies.

4.7.3. Transformer Architectures

The experimental investigation includes both token-based and hierarchical transformer architectures in order to analyze representation-dependent semantic interaction behavior.

4.7.3.1 Vision Transformer Architectures

Two Vision Transformer variants are evaluated:

- ViT-Tiny
- ViT-Base/16

These architectures process images as patch-token sequences and model contextual interaction through global self-attention mechanisms [8]. Semantic guidance is incorporated through semantic token encoding and cross-attention-based semantic fusion.

4.7.3.2. Hierarchical Swin Transformer Architecture

The Swin-Tiny architecture is used as the representative hierarchical transformer model. Swin Transformer preserves multi-stage spatial feature organization while enabling contextual interaction through localized shifted-window attention mechanisms [9]. Semantic guidance is incorporated using hierarchical spatial semantic encoding aligned with intermediate feature hierarchies.

All transformer backbones are initialized using ImageNet-pretrained weights to leverage large-scale visual representation learning prior to task-specific fine-tuning on indoor scene recognition.

4.7.4. Data Preprocessing and Augmentation

RGB images are resized and normalized using standard ImageNet preprocessing statistics. During training, random cropping and horizontal flipping are applied to

improve generalization and reduce overfitting. Additional appearance-based augmentations including brightness variation, contrast adjustment, and Gaussian blur are used to improve robustness against illumination and appearance variation common in indoor environments.

Semantic maps undergo the same spatial transformations as the RGB inputs to preserve semantic alignment during multimodal fusion. For token-based transformer models, semantic maps are partitioned into patches using the same patch extraction configuration employed by the Vision Transformer backbone.

During evaluation, random augmentations are disabled, and all images are processed using consistent center-cropping and resizing settings.

4.7.5. Optimization Strategy

The transformer frameworks are optimized using the AdamW optimizer with weight decay regularization. Due to the increased optimization sensitivity of transformer architectures, learning rates are carefully controlled during fine-tuning to preserve stable convergence behavior.

Categorical cross-entropy loss is used for scene classification:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i denotes the ground-truth label and $\hat{y}_i$ represents the predicted scene probability.

Learning rate scheduling is applied throughout optimization to improve training stability and convergence. Mixed-precision training is additionally employed to improve computational efficiency and reduce memory consumption during transformer optimization.

4.7.6. Representation-Aligned Training Strategy

An important aspect of the proposed framework is that semantic interaction is introduced in a representation-aligned manner rather than through identical multimodal fusion across all architectures. Consequently, training behavior differs slightly depending on the representation format of the backbone architecture.

For Vision Transformers, semantic token embeddings are integrated through cross-attention modules operating within token space. During training, the model learns semantic-aware contextual relationships between visual tokens and semantic tokens simultaneously.

For Swin Transformers, semantic guidance is introduced progressively through hierarchical spatial semantic modulation across multiple representation stages. Since semantic and visual feature maps remain spatially aligned throughout the hierarchy, semantic interaction occurs more directly through localized feature regulation.

Despite these architectural differences, all frameworks are trained under equivalent optimization schedules and evaluation settings to ensure fair comparative analysis.

4.7.7. Experimental Consistency

To maintain experimental reliability, all architectures are evaluated using identical dataset splits, semantic generation procedures, preprocessing pipelines, training epochs, optimization strategies, and evaluation metrics. The semantic-aware learning philosophy also remains consistent across all experiments.

As a result, observed differences in scene recognition performance can be attributed primarily to differences in representation structure and semantic alignment behavior rather than unrelated implementation or optimization variations. This controlled setup therefore provides a suitable basis for analyzing how token-based and hierarchical transformer representations respond to semantic-guided learning in indoor scene recognition.

4.8. RESULTS AND DISCUSSION

This section presents the experimental observations obtained from the proposed representation-aligned transformer framework on the MIT Indoor-67 dataset. The experiments are designed not only to evaluate transformer-based scene recognition performance, but also to analyze how semantic supervision interacts with different transformer representation structures under consistent semantic-aware learning conditions. The investigation compares token-based Vision Transformer architectures and hierarchical Swin Transformer architectures using representation-aligned semantic fusion strategies. Since all experiments are performed under identical dataset splits, optimization settings, and semantic-aware learning conditions, observed differences can be interpreted primarily in relation to representation structure and semantic interaction behavior.

4.8.1. Quantitative Performance Comparison

Table 4.1 presents the comparative performance of the investigated transformer architectures.

Table 4.1

Comparison of Transformer-Based Semantic-Aware Frameworks on MITIndoor67

Architecture	Representation Type	Configuration	Parameters	Top-1 Accuracy (%)
ViT-Tiny	Token Transformer	*vit_tiny_patch16_224*	14.30M	77.15
ViT-Base	Token Transformer	*vit_base_patch16_224*	133.30M	81.37
Swin-Tiny	Hierarchical Transformer	*swin_tiny_patch4_window7_224*	46.85M	**83.46**

The results demonstrate that the hierarchical Swin Transformer framework achieves the strongest performance among the investigated transformer architectures. Swin-Tiny

significantly outperforms both ViT-Tiny and ViT-Base despite having lower parameter complexity than ViT-Base. These observations suggest that representation structure plays an important role in semantic-guided transformer learning beyond model scale alone.

Although Vision Transformers demonstrate competitive recognition capability, their performance remains consistently below that of the hierarchical Swin Transformer framework under identical semantic-aware learning conditions. This gap indicates that semantic supervision may interact more effectively with hierarchical spatial transformer representations than with purely token-based representations for indoor scene understanding tasks.

4.8.2. Token-Based Semantic Interaction Analysis

Vision Transformers process visual information through globally interacting patch-token embeddings. Semantic token encoding enables semantic information to operate within the same sequential representation space, allowing contextual semantic interaction through cross-attention mechanisms. This design successfully introduces semantic guidance into token-space representation learning and enables broad contextual semantic interaction across distant image regions.

The experiments show that semantic token fusion improves contextual representation learning within Vision Transformers, particularly in scene categories requiring broad contextual reasoning across spatially distant objects. Global token interaction allows the model to establish semantic relationships between different scene regions efficiently.

However, several limitations emerge within purely token-based semantic interaction. Since token representations do not explicitly preserve hierarchical spatial continuity across intermediate stages, semantic structure becomes represented more implicitly through attention relationships. As a result, certain locality-sensitive scene characteristics may become more difficult to preserve consistently during deeper transformer processing.

Indoor scene recognition frequently depends on subtle spatial arrangements and localized contextual relationships involving furniture organization, scene layout, and object positioning. While Vision Transformers capture broad contextual interaction effectively, purely token-based representations may weaken some of the structural continuity necessary for fine-grained semantic scene understanding.

These observations suggest that semantic supervision alone is insufficient unless the underlying representation structure supports stable contextual organization throughout feature learning.

4.8.3. Hierarchical Spatial Representation Analysis

The Swin Transformer framework demonstrates stronger semantic-guided scene understanding performance primarily due to its hierarchical spatial representation structure. Unlike Vision Transformers, Swin Transformer preserves multi-stage spatial continuity throughout intermediate representation hierarchies while simultaneously supporting contextual interaction through shifted-window attention.

This representation structure appears highly compatible with semantic supervision because semantic information itself is inherently spatial and context-dependent. Spatial semantic encoding allows semantic maps to remain directly aligned with hierarchical

visual feature representations across multiple stages of the network. Consequently, semantic guidance can regulate localized scene structures more naturally and consistently during feature learning.

Another important advantage of Swin Transformer is its ability to balance locality preservation and contextual aggregation simultaneously. Local window-based attention maintains structured spatial continuity, while shifted-window interaction progressively expands contextual receptive fields across the representation hierarchy. This balance

appears particularly beneficial for indoor scene recognition, where scene identity depends on both local semantic arrangement and broader contextual organization.

The stronger performance of Swin-Tiny compared to ViT-Base also indicates that preserving structured spatial representation may be more important for semantic-guided indoor scene understanding than increasing global attention capacity alone. This observation reinforces the broader representation-learning perspective developed throughout the thesis.

4.8.4. Representation Alignment Perspective

One of the most important findings emerging from the transformer investigation is that semantic supervision becomes more effective when semantic representation format aligns naturally with the underlying backbone representation structure.

For Vision Transformers, semantic information must be transformed into token-compatible embeddings before meaningful interaction can occur. Although this enables semantic-guided token interaction, semantic structure becomes represented indirectly within the attention space.

In contrast, Swin Transformers maintain direct spatial correspondence between semantic and visual representations throughout the representation hierarchy. This alignment allows semantic supervision to interact more naturally with contextual scene structure during multimodal learning.

These observations suggest that representation compatibility plays a critical role in semantic-aware scene recognition. Effective semantic-guided learning depends not only on the presence of semantic information, but also on whether the architecture preserves a representation structure capable of supporting the required semantic interaction behavior.

4.8.5. Comparative Interpretation with CNN Frameworks

Comparing the transformer investigation with the convolutional experiments presented in Chapter 3 reveals several important architectural insights.

Convolutional architectures demonstrated that representational diversity strengthens contextual feature integration, while transformer experiments indicate that representation alignment and spatial continuity strongly influence semantic-aware learning behavior. Hierarchical spatial transformers appear to combine several advantages of both paradigms by preserving structured feature organization while enabling broader contextual interaction through attention mechanisms.

These findings suggest that semantic-aware scene recognition depends fundamentally on representation organization rather than solely on semantic fusion complexity or model scale. Architectures that maintain contextual continuity and spatially meaningful representation hierarchies appear better suited for semantic-guided indoor scene understanding.

4.8.6. Discussion Summary

Overall, the experimental results demonstrate that transformer representation structure significantly influences semantic-guided scene understanding. Token-based Vision Transformers provide strong global contextual modeling capability but exhibit weaker locality preservation during semantic interaction. Hierarchical Swin Transformers achieve stronger semantic-aware learning behavior by preserving structured spatial continuity while supporting contextual attention-based aggregation.

Most importantly, the investigation establishes that semantic supervision becomes more effective when semantic representation format aligns naturally with the native representation structure of the backbone architecture. This representation-aligned learning perspective forms the central conceptual contribution of the transformer investigation and provides the foundation for the unified comparative analysis presented in the next chapter.

CHAPTER 5

EXPERIMENTAL RESULTS AND DISCUSSION

5.1. INTRODUCTION

The previous chapters investigated semantic-guided indoor scene recognition across convolutional and transformer-based deep learning architectures under controlled multimodal learning settings. Chapter 3 analyzed how convolutional representation structure influences semantic-aware scene understanding through comparative evaluation of ResNet-50 and ResNeXt-50 within an attention-guided semantic fusion framework. Chapter 4 extended this investigation toward transformer-based architectures and examined how semantic supervision interacts with token-based and hierarchical transformer representations using representation-aligned semantic encoding strategies.

Collectively, these investigations revealed that semantic-guided scene recognition depends not only on the availability of semantic information, but also on how visual representations are internally organized within the underlying architecture. The experiments demonstrated that representational diversity improves semantic-aware convolutional learning, while representation alignment strongly influences semantic interaction quality within transformer frameworks. In particular, architectures preserving structured spatial continuity consistently exhibited stronger semantic-guided contextual learning behaviour than architectures relying purely on globally interacting token representations.

Motivated by these observations, the present chapter provides a unified experimental discussion and comparative representation-level analysis across all investigated architectures. Rather than treating convolutional and transformer experiments as isolated investigations, the chapter synthesizes the broader architectural insights emerging from semantic-guided scene understanding across multiple representation paradigms.

The chapter first presents a consolidated summary of the experimental results obtained throughout the thesis. Subsequently, a comparative analysis is performed focusing on representation behavior, semantic interaction mechanisms, contextual modeling characteristics, locality preservation, and semantic alignment across convolutional and transformer architectures. The discussion further examines how different representation structures influence semantic-guided feature learning and contextual scene understanding in complex indoor environments.

The objective of this chapter is therefore not merely to compare numerical performance, but to derive broader representation-learning insights regarding semantic-aware scene recognition across modern deep neural architectures.

5.2. UNIFIED EXPERIMENTAL SUMMARY

To provide a unified overview of the investigations conducted throughout this thesis, Table 5.1 summarizes the performance of all major architectures evaluated under semantic-aware learning settings on the MIT Indoor-67 dataset.

Table 5.1

Unified Experimental Comparison Across Investigated Architectures

Architecture	**Representation Type**	**Semantic Fusion Strategy**	**Top-1 Accuracy (%)**	**Key Observation**
ResNet-50	Hierarchical CNN	Spatial Attention Fusion	87.01	Strong hierarchical representation learning
ResNeXt-50 (32×4d)	Hierarchical CNN with Aggregated Transformations	Spatial Attention Fusion	**91.30****	Improved representational diversity enhances semantic interaction
ViT-Tiny	Token-Based Transformer	Semantic Token Cross-Attention	77.15	Strong global contextual modeling but weaker locality preservation
ViT-Base/16	Token-Based Transformer	Semantic Token Cross-Attention	81.37	Improved contextual interaction with larger token representation space
Swin-Tiny	Hierarchical Spatial Transformer	Spatial Semantic Modulation	83.46	Strong semantic alignment through hierarchical spatial representation

The consolidated results reveal several important trends regarding semantic-guided scene understanding across representation architectures.

First, semantic-aware learning consistently improves scene recognition capability across both convolutional and transformer paradigms, indicating that semantic context provides valuable complementary information for indoor scene understanding. Semantic

supervision helps the models capture contextual relationships, object composition, and scene structure beyond conventional appearance-based visual learning alone.
Second, the strongest overall performance is achieved by the ResNeXt-50 based semantic-guided convolutional framework. The results suggest that representational diversity introduced through grouped residual transformations substantially strengthens contextual feature integration within hierarchical convolutional architectures.
Third, hierarchical representation structures consistently outperform purely token-based transformer representations under semantic-aware learning settings. Although Vision Transformers demonstrate effective global contextual modeling capability, hierarchical Swin Transformer representations achieve stronger semantic-guided scene understanding despite lower model complexity compared to ViT-Base.
Another important observation is that increasing parameter count alone does not guarantee improved semantic-aware performance. Several architectures with larger parameter complexity achieve lower recognition accuracy than comparatively smaller but structurally better-aligned models. This behavior suggests that representation organization and semantic compatibility may influence semantic-guided learning more strongly than model scale alone.
Overall, the unified experimental summary indicates that semantic-aware scene recognition is closely connected to representation structure, representational diversity, and semantic alignment across deep neural architectures.

5.3. REPRESENTATION-LEVEL COMPARATIVE ANALYSIS

One of the central observations emerging from the experimental investigations is that semantic-guided scene recognition depends strongly on how visual information is represented internally within the backbone architecture. Although all investigated frameworks incorporate semantic supervision through multimodal fusion, the effectiveness of semantic-guided learning varies substantially depending on representation structure, contextual modeling behavior, and locality preservation characteristics.

Convolutional architectures organize visual information through hierarchical spatial feature maps constructed progressively using localized convolutional operations. These hierarchical representations preserve explicit spatial continuity throughout intermediate layers while gradually expanding contextual abstraction capability. Within semantic-aware learning, this representation structure allows semantic information to interact naturally with localized visual regions during feature modulation.

The comparison between ResNet-50 and ResNeXt-50 further demonstrates that representational diversity plays an important role in semantic-guided feature learning. ResNeXt introduces multiple parallel residual transformations through grouped convolutions, enabling richer contextual feature representations compared to standard residual architectures. The experiments indicate that semantic guidance becomes more effective when operating over more diverse and expressive convolutional feature spaces.

Transformer architectures, however, introduce fundamentally different representation paradigms. Vision Transformers process images as sequences of patch embeddings and model contextual interaction globally through self-attention mechanisms. This token-based representation enables broad contextual communication across distant image regions but weakens explicit preservation of spatial continuity within intermediate representations [8].

From a semantic-aware learning perspective, this distinction becomes highly significant. Semantic information derived from segmentation maps is inherently spatial and context-dependent. Consequently, semantic supervision may interact differently depending on whether the architecture preserves structured spatial organization explicitly or encodes contextual relationships primarily through token-level attention interactions.

The experimental results show that token-based transformer representations successfully capture broad contextual semantic relationships but exhibit comparatively weaker locality-sensitive semantic interaction behavior. Although semantic token encoding enables representation compatibility within Vision Transformers, semantic structure becomes represented more implicitly within the token attention space.

In contrast, hierarchical Swin Transformer representations preserve spatial continuity throughout multiple representation stages while simultaneously enabling contextual interaction through shifted-window attention mechanisms [9]. This representation organization appears more naturally compatible with semantic supervision because semantic structure and visual representation hierarchy remain spatially aligned during feature learning.

Another important representation-level observation concerns contextual abstraction behavior. Convolutional architectures aggregate contextual information progressively through receptive field expansion, whereas transformers establish contextual interaction directly through attention-based communication between representation elements. Hierarchical transformers combine both principles by preserving structured locality while enabling contextual aggregation across hierarchical stages.

The experiments therefore suggest that effective semantic-aware scene understanding requires balancing three important representation characteristics simultaneously:

1. Contextual modeling capability
2. Spatial continuity preservation
3. Semantic representation compatibility

Architectures emphasizing only global contextual interaction may weaken locality-sensitive semantic structure, while architectures preserving hierarchical spatial continuity appear better suited for semantic-guided indoor scene understanding.
These observations collectively support the broader thesis perspective that semantic-aware scene recognition is fundamentally a representation-learning problem rather than solely a multimodal fusion problem.

5.4. CNN VS TRANSFORMER SEMANTIC INTERACTION

The comparative investigation between convolutional and transformer-based frameworks reveals several important differences in how semantic supervision interacts with visual representations during scene understanding.

Within convolutional frameworks, semantic interaction occurs primarily through modulation of hierarchical spatial feature maps. Since convolutional representations preserve locality explicitly across intermediate stages, semantic attention mechanisms can regulate visual activations corresponding to specific spatial structures and contextual regions directly. Semantic guidance therefore operates as localized contextual modulation over structured visual feature hierarchies.

The experiments indicate that convolutional semantic interaction benefits strongly from representational diversity and hierarchical contextual abstraction. In particular, ResNeXt-based representations provide richer semantic-guided feature interaction than standard residual representations due to increased transformation diversity within grouped residual pathways.

Transformer-based semantic interaction behaves differently because transformers organize visual information through attention-based representation structures rather than strictly localized spatial hierarchies. Vision Transformers establish semantic interaction primarily through token-level contextual attention between semantic and visual embeddings [8]. This design enables broad semantic communication across distant image regions and supports flexible contextual reasoning within the representation space.

However, purely token-based semantic interaction also introduces challenges related to structural continuity and locality preservation. Since token embeddings represent image patches independently within sequential attention space, semantic relationships become encoded implicitly through attention patterns rather than explicit hierarchical spatial structure. As a result, fine-grained semantic organization may become less stable across deeper representation stages.

Hierarchical Swin Transformers demonstrate a more balanced semantic interaction mechanism by combining attention-based contextual learning with structured spatial continuity [9]. Semantic guidance operates directly on hierarchical spatial representations while shifted-window attention progressively expands contextual interaction across representation stages. This design enables the model to preserve locality-sensitive semantic organization while still capturing broader contextual relationships.

Another important difference concerns how semantic information propagates throughout the representation hierarchy. In convolutional architectures, semantic guidance primarily influences localized contextual feature refinement through spatial modulation. In Vision Transformers, semantic interaction propagates globally through token-space relational attention. Swin Transformers combine both localized semantic regulation and progressive contextual aggregation simultaneously.

These observations suggest that semantic-aware scene recognition behavior differs fundamentally across representation paradigms. Convolutional frameworks emphasize locality-preserving semantic modulation, token-based transformers emphasize globally distributed semantic interaction, and hierarchical transformers balance both contextual flexibility and structured semantic continuity.

From a broader architectural perspective, the experiments indicate that semantic supervision is not universally architecture-independent. Instead, semantic interaction quality depends strongly on how contextual information, spatial structure, and semantic organization are represented internally within the backbone architecture.

5.5. SPATIAL VS TOKEN REPRESENTATION ANALYSIS

One of the most important conceptual observations emerging from this thesis is the distinction between spatially organized representations and purely token-based representations in semantic-aware scene understanding.

Indoor scene recognition depends heavily on spatial arrangement, contextual organization, and structural continuity within the environment. Scene categories are often defined not only by object presence, but also by how objects are positioned relative to one another across the scene layout. Consequently, preserving meaningful spatial structure during representation learning becomes highly important for semantic-guided contextual reasoning [5].

Hierarchical convolutional representations and hierarchical Swin Transformer representations both maintain explicit spatial continuity throughout intermediate representation stages [9]. This structured organization allows semantic supervision to interact directly with localized contextual structures and object arrangements during feature learning. Semantic maps remain naturally aligned with visual feature organization, enabling consistent locality-sensitive semantic interaction.

In contrast, Vision Transformers represent visual information primarily through token sequences. Although positional embeddings encode relative spatial ordering information, the representation hierarchy itself does not explicitly preserve spatial continuity in the same structured manner as hierarchical architectures [8]. Semantic interaction therefore occurs indirectly through attention relationships between token embeddings.

This distinction influences how semantic guidance behaves during multimodal learning. Spatially organized representations preserve explicit contextual neighborhoods and structural locality throughout feature learning, allowing semantic modulation to operate directly on semantically meaningful scene regions. Token-based representations, however, rely on learned attention distributions to recover contextual relationships dynamically.

The experiments suggest that preserving hierarchical spatial continuity provides important advantages for semantic-guided indoor scene understanding. Swin Transformer achieves substantially stronger semantic-aware performance than Vision Transformer architectures despite lower parameter complexity, indicating that structured spatial organization contributes more effectively to semantic interaction than purely global token attention alone.

Another important observation is that semantic information itself possesses strong spatial structure. Segmentation-derived semantic maps inherently describe localized object distributions, contextual boundaries, and scene organization patterns. Consequently, semantic supervision becomes more naturally compatible with architectures preserving explicit spatial hierarchy throughout representation learning.

This compatibility can be interpreted as representation alignment between semantic structure and visual representation organization. Architectures maintaining spatial continuity allow semantic guidance to operate in a structurally consistent manner, whereas token-based architectures require semantic information to be transformed into less spatially explicit representation formats before interaction can occur.

The distinction between token-space semantic interaction and hierarchical spatial semantic interaction therefore extends beyond architectural implementation details. It

reflects a broader difference in how semantic context is represented, propagated, and preserved during scene understanding.

Collectively, these observations support the thesis-level conclusion that hierarchical spatial representations provide a more effective representation environment for semantic-guided indoor scene recognition than purely token-based representations.

5.6. LIMITATIONS OF THE PRESENT STUDY

Although the proposed investigations provide important insights into semantic-guided representation learning across convolutional and transformer architectures, several limitations remain that should be acknowledged.

One limitation arises from the dependence on externally generated semantic segmentation maps. Since semantic supervision is derived from pretrained segmentation models, inaccuracies in semantic prediction can influence multimodal fusion quality and scene recognition performance. Errors in segmentation boundaries, object labeling, or contextual region prediction may propagate into the semantic-aware representation space.

Another limitation concerns the scope of the investigated architectures. The study evaluates selected representative convolutional and transformer models in order to maintain a controlled comparative framework. However, additional transformer variants, hybrid architectures, or larger-scale multimodal models may exhibit different semantic interaction behavior.

The experiments are also limited to indoor scene recognition using the MIT Indoor-67 dataset. While the dataset provides substantial contextual complexity, the findings may not generalize fully to outdoor scene understanding, large-scale multimodal datasets, or domain-adaptive scene recognition settings.

An additional limitation is the dependence on ImageNet pretraining for backbone initialization. Pretrained representations strongly influence downstream semantic-aware learning behavior, and alternative pretraining strategies may produce different representation characteristics during semantic interaction.

Finally, the present investigation focuses primarily on visual-semantic interaction and does not explore multimodal extensions involving language supervision, depth information, temporal reasoning, or video-based scene understanding. Such modalities may provide additional contextual structure beyond segmentation-derived semantic guidance.

Despite these limitations, the study provides a controlled representation-level investigation into semantic-guided scene understanding across modern deep neural architectures.

5.7. KEY FINDINGS AND DISCUSSION SUMMARY

The investigations presented throughout this thesis demonstrate that semantic supervision significantly improves indoor scene understanding when integrated effectively with deep visual representations. More importantly, the experimental observations reveal that the effectiveness of semantic-guided learning depends strongly on representation structure, representational diversity, and semantic alignment within the underlying architecture.

The convolutional experiments showed that richer representational diversity strengthens contextual feature integration within hierarchical spatial representations. In particular,

grouped residual transformations in ResNeXt enabled stronger contextual semantic learning compared to standard residual architectures.

The transformer investigations further revealed that semantic interaction behavior differs substantially across token-based and hierarchical transformer representations. Vision Transformers demonstrated strong global contextual modeling capability but weaker locality-sensitive semantic interaction, whereas Swin Transformers preserved hierarchical spatial continuity while supporting contextual attention-based aggregation simultaneously.

Collectively, these findings establish that semantic supervision becomes more effective when semantic representation format aligns naturally with the native representation structure of the backbone architecture. Architectures preserving structured spatial continuity consistently demonstrated stronger semantic-aware scene understanding behavior than architectures relying purely on globally interacting token representations.

From a broader representation-learning perspective, the thesis demonstrates that semantic-aware scene recognition should not be viewed solely as a multimodal fusion problem or a benchmark optimization task. Instead, effective semantic-guided scene understanding depends fundamentally on how architectures organize contextual structure, preserve spatial continuity, and support semantically meaningful feature interaction during representation learning.

These observations provide the conceptual foundation for the final conclusions and future research directions presented in the next chapter.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

6.1. INTRODUCTION

Semantic-guided scene recognition has emerged as an important research direction in computer vision due to the increasing need for contextual understanding in complex visual environments. Indoor scene recognition, in particular, presents significant challenges because scene categories are often defined not only by appearance patterns, but also by semantic composition, object relationships, and spatial organization. Motivated by these challenges, this thesis investigated semantic-guided indoor scene recognition from a representation-learning perspective across convolutional and transformer-based deep neural architectures.

The study focused on understanding how semantic supervision interacts with different representation structures during scene understanding. Rather than treating semantic information as a generic auxiliary modality, the thesis examined whether the effectiveness of semantic-guided learning depends on the internal organization of visual representations within the backbone architecture. To explore this problem, controlled semantic-aware learning frameworks were investigated using convolutional residual networks, token-based Vision Transformers, and hierarchical Swin Transformers.

This chapter summarizes the major findings of the thesis, presents the key conclusions derived from the experimental investigations, discusses possible future research directions, and highlights the broader societal relevance of semantic-aware scene understanding systems.

6.2. THESIS SUMMARY

The thesis began by examining the fundamental challenges associated with indoor scene recognition. Unlike conventional object classification tasks, indoor scene understanding requires learning contextual relationships among multiple objects, spatial layouts, and semantic structures distributed across the environment. Traditional appearance-based recognition methods often struggle in such scenarios because visually similar environments may differ primarily in contextual composition rather than isolated visual patterns. These observations motivated the incorporation of semantic supervision into deep scene recognition frameworks.

The literature review presented the evolution of scene recognition from handcrafted feature-based methods to deep representation learning approaches involving convolutional neural networks and transformer architectures. The review further highlighted the growing importance of semantic-aware learning and identified an important research gap regarding the relationship between semantic supervision and representation structure across modern deep architectures.

The first experimental investigation focused on semantic-guided convolutional representation learning. A dual-branch semantic-aware framework consisting of RGB and

semantic branches connected through attention-based semantic fusion was developed using ResNet-50 and ResNeXt-50 backbones. The experiments demonstrated that semantic supervision substantially improves scene understanding and that representational diversity introduced through grouped residual transformations enhances semantic-guided feature interaction. The investigation further revealed that semantic-aware learning effectiveness depends partly on the richness and diversity of the underlying convolutional representation space.

Building upon these observations, the thesis extended the investigation toward transformer-based representation paradigms. Since transformer architectures organize visual information differently from convolutional networks, the study introduced a representation-aligned semantic framework in which semantic encoding strategies were adapted according to the representation structure of the backbone architecture. Vision Transformers employed semantic token encoding aligned with token-based representations, while Swin Transformers utilized hierarchical spatial semantic encoding aligned with structured feature hierarchies.

The transformer experiments demonstrated important differences in semantic interaction behavior across representation paradigms. Vision Transformers showed strong capability in modeling global contextual relationships through token-level interaction but exhibited weaker locality preservation during semantic-guided learning. In contrast, hierarchical Swin Transformers achieved stronger semantic-aware scene understanding by preserving spatial continuity while simultaneously enabling contextual aggregation through attention mechanisms.

The unified comparative analysis further established that semantic supervision becomes more effective when semantic representation format aligns naturally with the native representation structure of the backbone architecture. Hierarchical spatial representations consistently provided stronger semantic-guided interaction than purely token-based representations for indoor scene understanding tasks.

Overall, the thesis demonstrated that semantic-aware scene recognition should be viewed fundamentally as a representation-learning problem involving contextual modeling, semantic compatibility, and representation alignment across modern deep neural architectures.

6.3. MAJOR CONCLUSIONS

Based on the experimental investigations and comparative representation-level analysis conducted throughout this thesis, several important conclusions can be drawn.

First, semantic supervision significantly improves indoor scene recognition performance by incorporating contextual object-level information alongside appearance-based visual representations. Semantic guidance enables the models to capture scene composition, object relationships, and contextual structure more effectively than conventional RGB-based learning alone.

Second, the effectiveness of semantic-guided learning depends strongly on the representation structure of the backbone architecture. Semantic supervision does not operate independently of representation organization; instead, semantic interaction quality is closely influenced by how visual information is internally represented and processed within the network.

Third, representational diversity plays an important role in semantic-aware convolutional learning. The experimental comparison between ResNet-50 and ResNeXt-50 demonstrated that grouped residual transformations improve semantic-guided feature interaction by enabling richer and more diverse contextual representations within the convolutional feature space.

Fourth, transformer-based semantic interaction behavior differs substantially across representation paradigms. Token-based Vision Transformers provide strong global contextual modeling capability through self-attention mechanisms, but purely token-oriented representations exhibit weaker preservation of locality-sensitive semantic structure during scene understanding.

Fifth, hierarchical spatial representations provide more effective semantic-guided interaction for indoor scene recognition than purely token-based representations. The Swin Transformer framework consistently demonstrated stronger semantic-aware learning behavior because hierarchical feature organization preserves structured spatial continuity while still enabling contextual aggregation through attention-based interaction.

Sixth, semantic representation alignment emerges as a central factor in semantic-aware learning. Semantic supervision becomes more effective when semantic information is encoded in a manner naturally compatible with the representation structure of the underlying architecture. Representation-aligned semantic fusion therefore provides a more effective multimodal learning strategy than applying identical semantic interaction mechanisms across fundamentally different representation paradigms.

Finally, the thesis establishes that semantic-aware indoor scene recognition is fundamentally a representation-learning problem rather than solely a multimodal fusion or benchmark optimization problem. Effective scene understanding depends on the balance between contextual modeling capability, spatial continuity preservation, representational diversity, and semantic compatibility within the learned representation space.

6.4. FUTURE SCOPE

Although the present work provides important insights into semantic-guided representation learning for indoor scene recognition, several directions remain open for future research.

One promising direction involves investigating larger and more advanced hierarchical transformer architectures for semantic-aware scene understanding. Recent developments in foundation-scale vision models and multimodal transformer systems may provide stronger contextual reasoning capabilities and richer semantic interaction behavior for complex visual environments.

Another important direction is the development of adaptive semantic alignment mechanisms capable of dynamically adjusting semantic fusion strategies according to representation characteristics during training. Instead of using predefined semantic interaction schemes, future systems may learn representation-aware semantic adaptation automatically through end-to-end optimization.

Hybrid architectures combining convolutional inductive biases with transformer-based contextual modeling also represent an important area for future exploration. Such

architectures may benefit from both structured spatial continuity and flexible long-range contextual reasoning simultaneously.

Future work may additionally investigate semantic-aware learning using multimodal supervision beyond segmentation-derived semantic maps. Language-guided scene understanding, scene graph reasoning, depth-aware contextual learning, and multimodal vision-language representations may provide richer semantic context for indoor scene recognition systems.

Another promising research direction involves self-supervised and contrastive semantic representation learning. Large-scale multimodal pretraining strategies may help models learn semantic-contextual relationships more effectively without relying heavily on manually generated semantic annotations.

The present work is also limited to indoor scene recognition using the MIT Indoor-67 dataset. Future studies may extend the investigation toward larger-scale scene understanding benchmarks such as Places365, ADE20K scene-level tasks, and real-world multimodal indoor navigation datasets to evaluate the generalization capability of representation-aligned semantic learning frameworks.

Finally, future research may explore semantic-aware representation learning for embodied AI systems, robotics, autonomous navigation, and interactive intelligent environments where contextual scene understanding plays a critical role in decision-making and environmental interaction.

6.5. SOCIAL IMPACT

Semantic-aware scene understanding has significant potential to contribute toward the development of intelligent systems capable of interacting more effectively with complex human environments. By enabling machines to interpret contextual relationships, object arrangements, and environmental structure more accurately, semantic-guided scene recognition can support several important real-world applications.

In robotics and autonomous indoor navigation, semantic scene understanding can improve environmental awareness and contextual decision-making for service robots, assistive robots, and autonomous navigation systems operating within indoor spaces such as hospitals, offices, airports, and homes. Context-aware visual understanding may help intelligent systems navigate safely and interact more naturally with human-centered environments.

Semantic-aware scene recognition can also support assistive technologies for visually impaired individuals by enabling context-sensitive indoor navigation and environmental interpretation systems. Intelligent scene understanding may help provide more meaningful spatial and semantic information regarding surrounding environments.

In smart surveillance and security systems, semantic scene understanding may improve contextual event analysis and environmental monitoring by enabling systems to reason about scene composition and spatial context rather than relying solely on isolated object detection.

The broader implications of this work also extend toward context-aware intelligent environments, human-computer interaction systems, and future multimodal AI systems requiring deeper semantic understanding of real-world spaces.

At the same time, responsible development of intelligent visual systems remains important. Issues related to privacy, fairness, transparency, and ethical deployment must be carefully considered when applying semantic-aware scene understanding technologies in real-world environments.

6.6. CLOSING REMARKS

This thesis investigated semantic-guided indoor scene recognition across convolutional and transformer-based architectures from a representation-learning perspective. Through controlled experimental analysis and comparative architectural investigation, the study demonstrated that semantic supervision effectiveness depends strongly on representation structure, representational diversity, and semantic alignment within deep neural architectures.

The findings showed that hierarchical spatial representations provide more natural and effective environments for semantic-guided contextual learning than purely token-based representations in indoor scene understanding tasks. More broadly, the thesis established that semantic-aware scene recognition is fundamentally influenced by how contextual structure and semantic information are organized within the learned representation space itself.

It is hoped that the representation-centric perspective presented in this work contributes toward a deeper understanding of semantic-aware visual learning and motivates future research into more context-aware, semantically aligned, and structurally intelligent visual understanding systems.

REFERENCES

[1]. Susan, Seba, and Maduri Tuteja. "Feature engineering versus deep learning for scene recognition: a brief survey." *International Journal of Image and Graphics* 25, no. 06 (2025): 2550054.

[2]. Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3]. Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In 2005 *IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886-893. IEEE, 2005.

[4]. Oli, Aude, and Antonio Torralba. "Building the gist of a scene: The role of global image features in recognition." *Progress in brain research* 155 (2006): 23-36.

[5]. López-Cifuentes, Alejandro, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. "Semantic-aware scene recognition." *Pattern Recognition* 102 (2020): 107256.

[6]. Sarada and V. Jyothsna, "Multi-Model Deep Learning Framework for Scene Recognition in Indoor and Outdoor Environments: A Comparative Study," in *Proceedings of the 2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2025, pp. 1375–1383.

[7]. Parseh, Mohammad Javad, Mohammad Rahmanimanesh, Parviz Keshavarzi, and Zohreh Azimifar. "Scene representation using a new two-branch neural network model." *The Visual Computer 40*, no. 9 (2024): 6219-6244.

[8]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[9]. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF international conference on computer vision,* pp. 10012-10022. 2021.

[10]. Ahmed, Muhammad Waqas, and Ahmad Jalal. "Indoor scene classification using RGB-D data: A vision transformer and conditional random field approach." In 2024 *5th International Conference on Innovative Computing (ICIC)*, pp. 1-6. IEEE, 2024.

[11]. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

[12]. Xie, Lin, Feifei Lee, Li Liu, Koji Kotani and Qiu Chen. "Scene recognition: A comprehensive survey." *Pattern Recognit. 102* (2020): 107205.

[13]. Hu, Boyang, Yiping Gao, Xinyu Li, and Zerui Xi. "Relational Integrated Cross-modal Scene Fusion with CLIP for Fine-grained Indoor Scene Recognition in Intelligent Manufacturing Systems." *Chinese Journal of Mechanical Engineering (2026):* 100230.

[14]. Coelho, Thamiris, Leo SF Ribeiro, João Macedo, Jefersson A. dos Santos, and Sandra Avila. "Transformers-based few-shot learning for scene classification in child sexual abuse imagery." In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 8-14. SBC, 2024.

[15]. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems 25* (2012).

[16]. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[17]. Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the *IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015.

[18]. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the *IEEE conference on computer vision and pattern recognition,* pp. 770-778. 2016.

[19]. Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500. 2017.

[20]. Sun, Ning, Wenli Li, Jixin Liu, Guang Han, and Cong Wu. "Fusing object semantics and deep appearance features for scene recognition." *IEEE Transactions on Circuits and Systems for Video Technology 29,* no. 6 (2018): 1715-1728.

[21]. Xu, Kejie, Peifang Deng, and Hong Huang. "Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification." *IEEE Transactions on Geoscience and Remote Sensing 60 (2022)*: 1-15.

[22]. Ahmed, Muhammad Waqas, and Ahmad Jalal. "RGB-D Scene Classification: A Unified Framework with Vision Transformers and Contextual Models." In 2024 *3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, pp. 1-6. IEEE, 2024.

[23]. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature 521*, no. 7553 (2015): 436-444

[24]. Quattoni, Ariadna, and Antonio Torralba. "Recognizing indoor scenes." In 2009 *IEEE conference on computer vision and pattern recognition*, pp. 413-420. IEEE, 2009.

[25]. Zhou, Bolei, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. "Learning deep features for scene recognition using places database." *Advances in neural information processing systems 27* (2014).

[26]. Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141. 2018.

[27]. Woo, Sanghyun, Jongchan Park, Joon-Young Lee, and In So Kweon. "Cbam: Convolutional block attention module." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19. 2018.

[28]. Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. "Do better imagenet models transfer better?." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661-2671. 2019.

[29]. Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning. Vol. 1, no. 2. Cambridge: MIT press, 2016.*

[30]. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems 30 (2017)*.

[31]. Zhou, Bolei, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. "Semantic understanding of scenes through the ade20k dataset." *International journal of computer vision 127*, no. 3 (2019): 302-321.

LIST OF PUBLICATIONS & PROOFS

[P1]. Muheet Alam, and Seba Susan. "ResNeXt-Based Multimodal Scene Recognition via Semantic-Guided Attention Fusion." Accepted as a full paper for Oral Presentation at the *2026 Kalinga Conference on Communication & Computing (KalingaConf 2026)*, with publication in IEEE Xplore and Scopus indexing.

[P2]. Muheet Alam, and Seba Susan. "Representation-Aligned Semantic Guidance in Transformer-Based Indoor Scene Recognition." Under review at the *2026 IEEE International Conference on Sustainable Technologies for Smart Development Goals (ICSTSDG 2026).*

Revisiting Scene Recognition via Semantic-Guided Attention Fusion Using ResNeXt

Muheet Alam
Department of Information Technology
Delhi Technological University
Delhi-110042, India

Seba Susan
Department of Information Technology
Delhi Technological University
Delhi-110042, India

***Abstract*—Scene recognition remains a challenging visual classification problem because scenes vary widely in layout, object composition, and contextual structure. While combining visual appearance with semantic information has improved robustness, an important question remains underexplored: how much does the choice of visual backbone influence performance when such cues are learned together? In this work, we investigate this question by studying the impact of using a ResNeXt-50 backbone within a multimodal scene recognition framework and comparing it against a commonly adopted ResNet-50 baseline. ResNeXt introduces aggregated residual transformations through grouped convolutions, which allow multiple feature patterns to be learned in parallel without increasing computational cost. This property is particularly relevant for scene recognition, where diverse visual cues must be captured simultaneously. Motivated by these properties, we introduce a dual-branch multimodal framework that integrates ResNeXt-based visual features with semantic representations for scene classification. Semantic-guided attention modulates visual responses, producing discriminative fused features for accurate recognition. Experimental results show that the ResNeXt-50 based model achieves a Top-1 accuracy of 91.30%, compared to 87.10% obtained with ResNet-50 under identical settings. These findings suggest that backbone architecture plays a more critical role in multimodal scene recognition than is often assumed, and that increasing representational diversity through cardinality can lead to meaningful performance gains without additional model complexity.**

Keywords— Scene Recognition, Semantic Feature Fusion, ResNeXt-50, Deep Learning, Multimodal learning

I. Introduction

Autonomous intelligent systems are increasingly required to function reliably across a wide range of environments without constant human supervision. Achieving this capability depends on the system's ability to recognize and interpret the sur rounding scene, since environmental context provides essential cues for decision making, situational awareness, and long-term planning. For example, assistive vision systems used in smart buildings or wearable devices must distinguish between indoor settings such as offices, corridors, or public spaces in order to provide context-aware guidance and feedback. Similarly, large-scale image management and retrieval platforms rely on accurate scene understanding to organize, index, and search visual data collected from diverse sources. Scene recognition addresses this need by assigning a high-level semantic label to an image that captures the overall environment rather than focusing solely on individual objects. This comprehensive understanding enables systems to reason about spatial layout, functional usage, and contextual relationships within a scene. As a result, scene recognition has become a key component in many real-world applications, including human–computer interaction, video surveillance, autonomous systems, and multimedia retrieval. Moreover, it often serves as a foundational building block for more advanced visual tasks such as object detection, activity analysis, and image retrieval, where reliable contextual understanding can significantly improve overall system performance [1].

The core objective of scene recognition is to assign meaningful semantic labels to images, where each label corresponds to a type of environment defined by humans, such as natural scenes, indoor spaces, or outdoor settings. Unlike object-centric recognition, scene recognition focuses on understanding the overall environment depicted in an image rather than identifying individual objects in isolation. This requires capturing global layout, spatial structure, and contextual relationships among visual elements. Due to this requirement, scene recognition has become a fundamental problem in visual understanding, especially in applications that demand accurate interpretation of complex and diverse environments, such as robotics, surveillance, and autonomous systems. Unlike object recognition, this task requires understanding global spatial structure, object co-occurrence, and contextual relationships, which often exhibit large intra-class variation and strong inter class similarity. These characteristics make scene recognition particularly sensitive to the quality of learned visual representations and the model's ability to capture complementary contextual cues [2].

Recent studies have shown that combining visual features with semantic information can improve scene recognition by providing additional cues about object presence and spatial context [3]. However, the performance of such approaches is influenced not only by how different features are combined, but also by the quality of the visual representations learned before fusion. In particular, the choice of the visual backbone plays an important role in determining how informative and complementary these features are. Despite this, backbone

Investigating How Transformer Representations Influence Semantic Guidance in Indoor Scene Recognition: A Representation-Centric View of Semantic Scene Understanding

Muheet Alam
Department of Information Technology
Delhi Technological University
Delhi-110042, India

Seba Susan
Department of Information Technology
Delhi Technological University
Delhi-110042, India

Abstract— **Indoor scene recognition relies heavily on semantic context, as scenes are defined not only by objects but also by their spatial arrangement. While transformer-based architectures have shown strong performance in visual recognition, it remains unclear how their internal representation structures interact with semantic guidance. In this work, we investigate this interaction through a semantic-aware indoor scene recognition framework that integrates RGB features with semantic segmentation information. A key aspect of the framework is representation-aligned semantic encoding, where semantic features are structured to match the representation format of the underlying transformer backbone. Specifically, semantic maps are converted into tokens for Vision Transformers to enable cross-attention with RGB features, while spatial semantic feature maps are used for the hierarchical Swin Transformer to preserve spatial alignment during fusion. We evaluate three transformer backbones—ViT-Tiny, ViT-Base, and Swin-Tiny—on the MIT Indoor 67 dataset under identical training conditions. Results show that Swin-Tiny achieves the highest accuracy (83.46%), outperforming ViT-Base (81.37%) and ViT-Tiny (77.15%) despite using fewer parameters. These findings suggest that hierarchical spatial representations align more naturally with semantic scene cues, enabling more effective semantic-guided feature integration. Overall, this study highlights the importance of representation structure in designing semantic-aware transformer models for scene recognition.**

Keywords— Scene Recognition, Semantic Feature Fusion, Hierarchical Vision Transformers, Swin Transformer, Multimodal learning

I. INTRODUCTION

Indoor scene recognition requires understanding the semantic structure of visual environments rather than simply detecting individual objects. Unlike object recognition, which focuses on identifying isolated entities, scene recognition involves interpreting how multiple visual elements collectively define a place. Indoor environments are often characterized by the objects they contain and the relationships among them—for example, the co-occurrence of a stove and refrigerator suggests a kitchen, while a bed and wardrobe indicate a bedroom. At the same time, scenes are also defined by spatial layout: corridors are recognizable by their elongated structure, whereas bookstores and libraries are more strongly defined by the objects they contain. Consequently, effective scene recognition requires representations that capture both object-level semantic cues and global spatial structure. These challenges, along with large intra-class variability in indoor scenes, have motivated benchmarks such as the MIT Indoor 67 dataset [3], which is widely used for evaluating indoor scene recognition systems [1,2].

Deep learning approaches, particularly convolutional neural networks (CNNs), have significantly improved scene recognition by learning hierarchical visual features from large-scale datasets. However, purely visual feature learning does not always capture the semantic relationships that characterize complex environments. As a result, several works have explored incorporating object-level or semantic information into scene recognition pipelines, demonstrating that semantic context can provide complementary cues for understanding scene categories [4–8].

More recently, transformer-based architectures have emerged as powerful alternatives for visual representation learning [9]. Vision Transformers (ViT) model images as sequences of patch tokens and capture relationships between them through global self-attention [10]. In contrast, hierarchical transformers such as the Swin Transformer maintain spatial feature maps and progressively aggregate information across multiple stages [11]. As illustrated in Fig. 1, these architectures differ fundamentally in how visual representations are structured: token-based transformers treat images as unordered token sequences, whereas hierarchical transformers preserve spatial structure while expanding the receptive field across layers.

These representational differences may influence how semantic information interacts with visual features [12–14]. Token-based representations enable flexible global interactions between image regions, while spatial hierarchical representations maintain explicit spatial relationships between objects. Since semantic cues in indoor environments often depend on both object presence and spatial arrangement [6], the representation format of the backbone architecture may play a

PROOF OF ACCEPTANCE FOR 1ST PAPER

1. REGISTRATION FEE RECEIPT

EVENTPEDIA

Razorpay
Invoicing and payments
powered by Razorpay

Payment Receipt Transaction Reference: pay_StwG41wlinPKeU

This is a payment receipt for your transaction on KALINGACONF 2026

AMOUNT PAID **₹ 9,500.00**

ISSUED TO
muheetalam784@gmail.com
+919027460390

PAID ON
26 May 2026

Registered Author Name
Muheet Alam

DESCRIPTION	UNIT PRICE	QTY	AMOUNT
Non IEEE Member Full Paper Registration	₹ 9,500.00	1	₹ 9,500.00
	Total		**₹ 9,500.00**
	Amount Paid		₹ 9,500.00

ACCEPTANCE MAIL

Fw: IEEE KALINGACONF 2026 - ACCEPTANCE NOTIFICATION

1 message

Seba Susan <seba_406@yahoo.in>
To: Muheet Alam <muheetalam_24isy09@dtu.ac.in>

Thu, 28 May, 2026 at 3:33 pm

---------- Forwarded message ---------
From: Microsoft CMT <noreply@msr-cmt.org>
To: Seba Susan <seba_406@yahoo.in>
Sent: Tuesday, 26 May 2026 at 12:55:59 pm IST
Subject: IEEE KALINGACONF 2026 - ACCEPTANCE NOTIFICATION

Dear Seba Susan

Paper ID / Submission ID : 721

Title : ResNeXt-Based Multimodal Scene Recognition via Semantic-Guided Attention Fusion

We are pleased to inform you that your paper has been accepted for the Oral Presentation as a full paper for the- " 2026 Kalinga Conference on Communication & Computing (Kalingaconf) , will be held in Kalinga University , Raipur , Chhattisgarh , India.

All accepted and presented papers will be submitted to IEEE Xplore for the further publication and will be indexed by Ei Compendex and Scopus Indexing.

Complete the Registration Process (The last date of payment is 29 MAY 2026)

Registration Link for Payment :

For Indian Authors : https://rzp.io/rzp/QU1DToGV

For Foreign Authors : https://www.explara.com/e/kalingaconf-2026
(Click on Stripe payment to make payment and then pay)

After the last date of registration IEEE PDF Express and E copyright information will be given for registered authors.

Note :

1. Any changes with the Author name, Affiliation and content of paper will not be allowed after acceptance. (if not added then can add and update in CMT)
2.This is Hybrid Conference, both online and physical presentation mode is available,

Reviewer Comments (Summary)

- Reviewer feedback indicates good technical contribution.
- Suggested minor improvements for clarity and formatting.
- Methodology and results are appreciated.

Please revise the paper as per comments and submit your camera-ready version along with author registration.

Further instructions will be communicated soon after registration.

Congratulations once again.

Thanks & Regards
TPC Chair
IEEE KALINGACONF 2026
Email Support : kalingaconf@gmail.com
Mob No : +91 8805435941
Website : https://kalingaconf.in/

Please do not reply to this email as it was generated from an email account that is not monitored.

SEMANTIC-GUIDED DEEP LEARNING FRAMEWORKS FOR SCENE RECOGNITION: A COMPARATIVE STUDY OF CNN AND TRANSFORMER MODELS

by Seba Susan

29/5/26

Submission date: 28-May-2026 03:32PM (UTC+0530)
Submission ID: 2971175606
File name: Thesis_for_Plag.docx (2.68M)
Word count: 23154
Character count: 172080

SEMANTIC-GUIDED DEEP LEARNING FRAMEWORKS FOR SCENE RECOGNITION: A COMPARATIVE STUDY OF CNN AND TRANSFORMER MODELS

ORIGINALITY REPORT

8%	7%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%
2	www.jianshu.com Internet Source	1%
3	www.ijcaonline.org Internet Source	<1%
4	www.mdpi.com Internet Source	<1%
5	ebin.pub Internet Source	<1%
6	staging.preprints.org Internet Source	<1%
7	export.arxiv.org Internet Source	<1%
8	Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, Álvaro García-Martín. "Semantic-aware scene recognition", Pattern Recognition, 2020 Publication	<1%
9	deepai.org Internet Source	<1%
10	Zhao, Shiyu. "Enhancing Visual Understanding With Large Foundational Models.", Rutgers	<1%

	The State University of New Jersey, School of Graduate Studies Publication	
11	"Communication and Intelligent Systems", Springer Science and Business Media LLC, 2025 Publication	<1 %
12	Fida Hussain Dahri, Ghulam E Mustafa Abro, Nisar Ahmed Dahri, Asif Ali Laghari, Zain Anwar Ali. "Advancing Robotic Automation with Custom Sequential Deep CNN-Based Indoor Scene Recognition", ICCK Transactions on Intelligent Systematics, 2024 Publication	<1 %
13	stax.strath.ac.uk Internet Source	<1 %
14	Boyang Hu, Yiping Gao, Xinyu Li, Zerui Xi. "Relational Integrated Cross-modal Scene Fusion with CLIP for Fine-grained Indoor Scene Recognition in Intelligent Manufacturing Systems", Chinese Journal of Mechanical Engineering, 2026 Publication	<1 %
15	fbscience.com Internet Source	<1 %
16	ir.vistas.ac.in Internet Source	<1 %
17	Shikha Gupta, Krishan Sharma, Dileep Aroor Dinesh, Veena Thenkanidiyoor. "Visual Semantic-Based Representation Learning Using Deep CNNs for Scene Recognition", ACM Transactions on Multimedia Computing, Communications, and Applications, 2021 Publication	<1 %

18	ijsrem.com Internet Source	<1 %
19	www.medrxiv.org Internet Source	<1 %
20	Xinyu Zhou, Xinru Tang, Bin Chen. "A novel multimodal semantic–spatial representation method for embodied perception and spatial reasoning", Pattern Recognition, 2026 Publication	<1 %
21	pure.uva.nl Internet Source	<1 %
22	www.iaeme.com Internet Source	<1 %
23	spiral.imperial.ac.uk Internet Source	<1 %
24	Yueqi Duan, Jiwen Lu, Jianjiang Feng, Jie Zhou. "Deep Localized Metric Learning", IEEE Transactions on Circuits and Systems for Video Technology, 2018 Publication	<1 %
25	journalppw.com Internet Source	<1 %
26	ijses.com Internet Source	<1 %
27	Jingxin Liang, Yangyang Xu, Haorui Song, Yuqin Lu, Yuhui Deng, Yiyi Long, Yan Huang, Shengxin Liu, Jianbo Jiao, Shengfeng He. "ContX: Scene context prediction via context bank and layout perception", Pattern Recognition, 2025 Publication	<1 %
28	psasir.upm.edu.my Internet Source	<1 %

29	d197for5662m48.cloudfront.net Internet Source	<1 %
30	DeSmet, Chance Nicholas. "Generative Adversarial Networks for Multi-Objective Synthetic Data Generation", Washington State University, 2024 Publication	<1 %
31	Hongje Seong, Junhyuk Hyun, Euntai Kim. "FOSNet: An End-to-End Trainable Deep Neural Network for Scene Recognition", IEEE Access, 2020 Publication	<1 %
32	Lingping Kong, Ponnuthurai Nagaratnam Suganthan, Václav Snášel, Varun Ojha, Jeng-Shyang Pan. "Enhancing Sampling Performance in XGBoost by Ensemble Feature Engineering", Pattern Recognition, 2026 Publication	<1 %
33	www.grafiati.com Internet Source	<1 %
34	Abhilasha Sharma, Vishwas Rathi, Anupam Biswas, Anil Singh, Omer Rana. "Multimodal Artificial Intelligence in Precision Agriculture - Practices, Challenges, and Applications", CRC Press, 2026 Publication	<1 %
35	Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniques and Smart Manufacturing", CRC Press, 2026 Publication	<1 %
36	encyclopedia.arabpsychology.com Internet Source	<1 %

37	Aysha Naseer, Moneerah Alotaibi, Haita F. Alharron, Nouf Abdullah Almujally, Shroo T. Alhambi, Ahmad Jalal, Bumshik Lee. "Enhancing scene understanding using RGB-D visuals and deep learning segmentation models", ETRI Journal, 2026 Publication	<1 %
38	Ye Xu, Lihua Duan, Conggui Huang, Chongpeng Huang. "Deep feature voting: a semantic-driven and local context-aware approach for image classification", Multimedia Tools and Applications, 2023 Publication	<1 %
39	repository.iaa.ac.tz:8080 Internet Source	<1 %
40	theses.hal.science Internet Source	<1 %
41	Manahardas, Chavda Sagarkumar. "DeepScenes: Scene Level Image Classification.", Gujarat Technological University Publication	<1 %
42	Rui Ma, Xuegang Dai, Zuochao Yang, Zhixiong Wei, Bin Zhang. "CAFR-Net: A transformer-contrastive framework for robust spinal MRI segmentation via global-local synergy", PLOS One, 2025 Publication	<1 %
43	visionbook.mit.edu Internet Source	<1 %
44	Amorim, Manuel João Gomes Alves. "Exploring the Efficiency and Interpretability of Kolmogorov-Arnold Networks Across	<1 %

Diverse Domains", Universidade do Porto (Portugal), 2025
Publication

45	Zamanidoost, Yadollah. "Early-Stage Lung Cancer Detection Using Deep Learning Algorithms.", Ecole Polytechnique, Montreal (Canada) Publication	<1 %
46	ijcrt.org Internet Source	<1 %
47	www.ijraset.com Internet Source	<1 %
48	Jie Liu, Chuangwei Xu, Shiyuan Han, Linye Song, Peixiao Wang, Tong Zhang. "Uncertainty-aware precipitation nowcasting with diffusion model simulating precipitation evolution processes", Engineering Applications of Artificial Intelligence, 2026 Publication	<1 %
49	S. P. Godlin Jasil, V. Ulagamuthalvi. "Deep learning architecture using transfer learning for classification of skin lesions", Journal of Ambient Intelligence and Humanized Computing, 2021 Publication	<1 %
50	Yongsheng Pan, Yong Xia, Yang Song, Weidong Cai. "Locality constrained encoding of frequency and spatial information for image classification", Multimedia Tools and Applications, 2018 Publication	<1 %
51	dokumen.pub Internet Source	<1 %

dspace.bits-pilani.ac.in:8080

52	Internet Source	<1 %
53	dspace.mit.edu Internet Source	<1 %
54	papers.neurips.cc Internet Source	<1 %
55	Chen, Zhimin. "Multi-Modal Data-Efficient Learning for 3D Machine Vision.", Clemson University Publication	<1 %
56	Mohammad Javad Parseh, Mohammad Rahmanimanesh, Parviz Keshavarzi, Zohreh Azimifar. "Scene representation using a new two-branch neural network model", The Visual Computer, 2023 Publication	<1 %
57	Xing-Rong Fan, Dahai Cai, Xiujuan Wang, Mengzhen Kang. "SGS-DETR: A lightweight transformer for real-time strawberry growth stage detection in smart agriculture systems", Smart Agricultural Technology, 2026 Publication	<1 %
58	etd.aau.edu.et Internet Source	<1 %
59	Abbas Moallem, Helmut Degen, Stavroula Ntoa. "Artificial Intelligence and Large Language Models - A Scientific Perspective", CRC Press, 2026 Publication	<1 %
60	Kandakji, Lynn. "Probabilistic Modelling and Deep Learning for Multi-Dimensional Keratoconus Detection.", University of London, University College London (United Kingdom)	<1 %

	Publication	
61	Lecture Notes in Computer Science, 2012. Publication	<1 %
62	assets-eu.researchsquare.com Internet Source	<1 %
63	thesesjournal.com Internet Source	<1 %
64	www.eurecom.fr Internet Source	<1 %
65	www.thetalkingmachines.com Internet Source	<1 %
66	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018 Publication	<1 %
67	"Intelligent Computing & Optimization", Springer Science and Business Media LLC, 2022 Publication	<1 %
68	"Smart Technologies, Systems and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %
69	Ali Jamali, Swalpa Kumar Roy, Pedram Ghamisi. "WetMapFormer: A unified deep CNN and vision transformer for complex wetland mapping", International Journal of Applied Earth Observation and Geoinformation, 2023 Publication	<1 %
70	Dashuai Wang, Minghu Zhao, Zhuolin Li, Sheng Xu, Xiaohu Wu, Xuan Ma, Xiaoguang Liu. "A survey of unmanned aerial vehicles	<1 %

and deep learning in precision agriculture", European Journal of Agronomy, 2025
Publication

71	Leng, Ziyang. "Representation Learning for Image-Based Autonomous Driving Perception.", University of California, Los Angeles Publication	<1 %
72	Lin Mao, Xuemeng Li, Dawei Yang, Rubo Zhang. "Convolutional Feature Frequency Adaptive Fusion Object Detection Network", Neural Processing Letters, 2021 Publication	<1 %
73	M. Dhavamani, K. Arun Prakash, P. Vadivel. "Applied Mathematics, Automation and Computing", CRC Press, 2026 Publication	<1 %
74	Peng Tang, Jin Zhang, Xinggang Wang, Bin Feng, Fabio Roli, Wenyu Liu. "Learning extremely shared middle-level image representation for scene classification", Knowledge and Information Systems, 2016 Publication	<1 %
75	Shuang Bai, Huadong Tang. "Softly combining an ensemble of classifiers learned from a single convolutional neural network for scene categorization", Applied Soft Computing, 2018 Publication	<1 %
76	Xinliang Xu, Qiming Liu, Xin Wen, Heng Zhao, Zhenhao Wang, Chong Wang. "An Adaptive Audiovisual Fusion Method Based on Prediction Confidence for Fine Granularity Bird Species Recognition", Applied Sciences, 2026 Publication	<1 %

77	ecole-itn.eu Internet Source	<1 %
78	ijrcar.com Internet Source	<1 %
79	oa.upm.es Internet Source	<1 %
80	www.frontiersin.org Internet Source	<1 %
81	www.ijprems.com Internet Source	<1 %
82	www.preprints.org Internet Source	<1 %
83	Lecture Notes in Computer Science, 2015. Publication	<1 %
84	Yang, Jinfu, Jizhao Zhang, Guanghui Wang, and Mingai Li. "Contour Detection-based Discovery of Mid-level Discriminative Patches for Scene Classification", International Journal of Advanced Robotic Systems, 2016. Publication	<1 %
85	Aytekin, Çaglar, A. Aydin Alatan, and Tien-Hsin Chao. "", Optical Pattern Recognition XXII, 2011. Publication	<1 %
86	Bal Virdee, Ummer Iqbal, Ashish Khanna, Moolchand Sharma, Roman Danel. "Transforming Healthcare with AI and IoT - Intelligent Solutions for a Digital Future", CRC Press, 2026 Publication	<1 %
87	Meiqiao Bi, Minghua Wang, Zhi Li, Danfeng Hong. "Vision Transformer with Contrastive	<1 %

Learning for Remote Sensing Image Scene Classification", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022
Publication

88 Xiaofeng Yang, Tonghe Wang. "AI-driven Medical Image Analysis in Precision Radiation Therapy", CRC Press, 2026 <1%
Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Seba Susan

SEMANTIC-GUIDED DEEP LEARNING FRAMEWORKS FOR SCENE RECOGNITION: A COMPARATIVE STUDY OF CNN AND ...

Quick Submit

Quick Submit

Delhi Technological University

29/5/26

Document Details

Submission ID
trn:oid:::1:3580954015

Submission Date
May 28, 2026, 3:31 PM GMT+5:30

Download Date
May 28, 2026, 3:38 PM GMT+5:30

File Name
Thesis_for_Plag.docx

File Size
2.7 MB

68 Pages

23,154 Words

172,080 Characters

Submission ID trn:oid:::1:3580954015

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.