School of
Electronics and Communication Engineering

Research Experience for Undergraduates

on

Hybrid Equivariant-Invariant Feature Learning for Visual perception in Dynamic visual environments

By:

1. **Vineet V Desai** USN: 01FE22BEC176

Semester: VII, 2025-2026

Under the Guidance of

Prof Preeti Pillai

Dr. Ujwala Patil

Dr. Nalini Iyer

K.L.E SOCIETY'S
KLE Technological University,
HUBBALLI-580031
2025-2026

SCHOOL OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that the project entitled **"Hybrid Equivariant-Invariant Feature Learning for Visual perception in Dynamic visual environments "** is a bonafide work carried out by the student **Vineet V Desai (01FE22BEC176)**. The project report has been approved as it satisfies the requirements with respect to the research experience for undergraduate work prescribed by the university curriculum for **B.E. (VII Semester)** in the School of Electronics and Communication Engineering of **KLE Technological University** for the academic year **2025–2026**.

Prof Preeti Pillai	**Suneeta V Budihal**	**B. S. Anami**
Dr.Ujwala Patil	**Head of Department**	**Registrar**
Dr. Nalini Iyer		
Guide		

External Viva:

Name of Examiners **Signature with date**

1.

2.

ACKNOWLEDGMENT

We express our heartfelt gratitude to the faculty and management for their professional guidance and support, which played a pivotal role in successfully completing our project work.

We sincerely thank Dr. Ashok Shettar, Pro Chancellor and Dr. B. S. Anami, Registrar of KLE Technological University, Hubballi, for their visionary leadership and encouragement throughout this journey. Our special thanks go to Dr. P.G. Tewari,Vice-Chancellor , for his unwavering support and commitment to fostering academic excellence. We are deeply grateful to Dr. Suneeta V. Budihal, Professor and Head of the Department of Electronics and Communication Engineering, for her valuable direction and for creating an environment conducive to skill enhancement and academic growth. We are immensely grateful to our guide, Prof. Preeti Pillai, for her constant guidance, insightful interactions, and constructive feedback during the reviews.

We also acknowledge the reviewers for their critical suggestions and valuable input, which significantly enriched our project. Finally, we thank the faculty in charge for their continuous support and encouragement, which were instrumental in the completion of our work

-The Project team

ABSTRACT

Visual perception systems operating in real-world environments are required to remain reliable under continuous variations in viewpoint, motion, illumination, and scene structure. While deep learning–based vision models achieve strong performance in controlled conditions, their reliability often degrades when exposed to distribution shifts during deployment. A key limitation arises from the inherent tension between geometric equivariance and semantic invariance in learned visual representations. Equivariance is essential for preserving spatial structure and ensuring consistent responses to geometric transformations, whereas invariance is required for maintaining stable semantic understanding across appearance changes. Existing approaches typically emphasize one of these properties while neglecting the other, limiting robustness in dynamic environments.

This project presents a hybrid feature learning framework that explicitly integrates equivariant and invariant representation learning within a unified architecture. The proposed approach separates geometric and semantic learning objectives into dedicated branches while sharing low-level visual features, enabling structured spatial representations and stable semantic embeddings to be learned simultaneously. The framework is trained using self-supervised objectives that encourage transformation-consistent feature behavior and appearance-invariant semantic representations.

Experimental evaluation under diverse transformations demonstrates stable learning behavior, improved geometric consistency, and robust semantic stability compared to conventional single-branch models. Qualitative analyses further confirm that the learned representations exhibit predictable spatial transformations while maintaining invariant semantic responses. The findings highlight the importance of jointly modeling equivariance and invariance for building robust visual perception systems capable of reliable operation in dynamic real-world settings.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Visual perception is a core component of intelligent systems such as autonomous driving, robotics, and surveillance, where reliable interpretation of visual data is required under continuously changing environmental conditions. Although convolutional neural networks (CNNs) have achieved remarkable success in controlled settings [1, 6], their performance often degrades when deployed in real-world scenarios involving viewpoint changes, illumination variations, and scene dynamics [10]. This degradation is largely attributed to the entangled treatment of geometric structure and semantic information in conventional CNN representations.

From a representation learning perspective, robust perception requires both *equivariance* and *invariance*. Equivariance ensures that geometric transformations applied to the input induce predictable transformations in the feature space, preserving spatial structure [8, 2]. In contrast, invariance aims to suppress nuisance variations and maintain semantic stability under transformations such as appearance changes [5]. While both properties are essential, enforcing one often compromises the other, and most existing approaches do not explicitly model their interaction.

Recent studies have highlighted that equivariance and invariance are complementary and should be jointly considered for improved generalization and robustness [4]. However, many methods either integrate them implicitly or focus on task-level performance without providing explicit representation-level analysis. Moreover, robustness techniques such as test-time adaptation primarily emphasize invariance and often neglect geometric consistency [9].

Motivated by these limitations, this project proposes a dual-branch learning framework that explicitly separates and jointly optimizes geometric equivariance and semantic invariance using a shared backbone architecture. By decoupling these objectives while preserving shared visual features, the framework aims to achieve stable and interpretable representations suitable for real-world visual perception under distribution shifts.

1.1 Motivation

Visual perception plays a critical role in modern intelligent systems such as autonomous driving, mobile robotics, and intelligent surveillance. These systems operate in highly dynamic and unconstrained environments where visual inputs are subject to continuous variations in viewpoint, object motion, illumination, weather conditions, and scene composition. For reliable decision-making, perception models must remain stable and consistent despite such variations.

Deep convolutional neural networks (CNNs) have demonstrated remarkable success in visual recognition tasks under controlled training conditions. However, their performance often degrades significantly when deployed in real-world environments due to distribution shifts between training and test data. These shifts frequently cause a loss of geometric consistency in spatial features or instability in semantic representations, leading to unreliable predictions in safety-critical scenarios.

From a representation learning perspective, robustness in visual perception depends on two complementary properties: equivariance and invariance. Equivariance ensures that geometric transformations applied to the input result in predictable transformations in the feature space, preserving spatial structure. Invariance, on the other hand, ensures that semantic information remains stable under such transformations. Existing methods often emphasize one of these properties at the expense of the other, motivating the need for a unified approach that explicitly models both.

1.2 Objectives

The specific objectives are as follows:

- To design a dual-branch neural architecture that explicitly separates equivariant (geometric) and invariant (semantic) feature learning while sharing a common backbone.
- To study the behavior of learned representations under geometric and appearance transformations such as rotation, flipping, and color variations.
- To quantitatively evaluate geometric consistency and semantic stability using dedicated metrics.
- To compare the proposed hybrid framework against a baseline CNN model and demonstrate improved robustness under distribution shifts.

1.3 Literature Survey

The field of visual perception has been significantly advanced by convolutional neural networks, whose inherent translation equivariance has enabled remarkable success in image recognition and scene understanding tasks [1, 6]. However, real-world visual data is rarely limited to translations alone and often includes complex transformations such as rotations, scale variations, viewpoint changes, and perspective distortions. Standard CNNs do not explicitly model these transformations, leading to unstable feature representations and reduced robustness under distribution shifts [10, 8]. This limitation has motivated extensive research into equivariant representation learning, where the goal is to ensure that transformations applied to the input induce predictable and structured changes in the feature space. Group-equivariant and steerable convolutional networks exemplify this direction by enforcing symmetry constraints within the network architecture [2, 3, 17]. While these approaches improve geometric consistency, they often rely on predefined transformation groups and introduce increased computational complexity, limiting their applicability in unconstrained real-world scenarios.

In parallel, invariant representation learning has emerged as a dominant paradigm for improving semantic robustness. Invariant learning methods aim to suppress nuisance variations by encouraging features to remain stable under transformations that do not alter semantic identity. This idea has been widely explored through data augmentation, metric learning, and self-supervised contrastive learning frameworks [5, 12, 15]. Such approaches have demonstrated strong generalization and semantic stability, particularly in classification and recognition tasks. However, excessive invariance can be detrimental in tasks requiring spatial awareness or structured reasoning, as collapsing spatial information may discard valuable geometric cues necessary for localization and motion understanding [8, 10].

Recent research has increasingly recognized that equivariance and invariance are not competing objectives but complementary properties of effective visual representations. Equivariance preserves geometric structure, while invariance ensures semantic stability. Some hybrid approaches have attempted to integrate both aspects within a single learning framework, showing improved generalization in constrained settings such as few-shot learning [4]. Nevertheless, many existing methods lack a clear architectural separation between geometric and semantic objectives, resulting in entangled representations that are difficult to analyze and optimize. Moreover, quantitative evaluation of equivariance and invariance is often indirect, relying on downstream task performance rather than explicit representation-level metrics [8].

Another important research direction focuses on robustness under distribution shifts, adversarial perturbations, and test-time variations [9, 11, 14]. These methods typically employ test-time adaptation or consistency regularization strategies to stabilize model predictions. While effective to some extent, most of these approaches prioritize invariant behavior and largely overlook geometric consistency, leaving spatial instability unresolved in transformation-heavy environments such as autonomous driving.

Collectively, existing literature highlights a clear gap in designing practical, scalable frameworks that explicitly and jointly model equivariance and invariance while remaining suitable for real-world deployment [16, 4]. This project addresses this gap by proposing a dual-branch hybrid learning architecture that disentangles geometric and semantic learning objectives, supported by dedicated evaluation metrics to analyze representation behavior under transformations. By grounding the framework in real-world visual data [7] and focusing on representation-level robustness, the project contributes toward more reliable and interpretable visual perception systems.

1.3.1 Limitations of Existing Approaches

Despite significant advances, current approaches generally target either equivariance or invariance exclusively, or study their combination under highly restricted conditions or narrow application areas. Explicitly disentangling geometric and semantic goals under a single architecture remains uncommon; further, quantitative studies under realistic distribution shifts for both properties are also relatively scarce. Moreover, quantitative metrics that consider both equivariance consistency and invariance stability jointly are seldom explored.

1.4 Problem Statement

The primary objective of this work is to design and analyze a hybrid feature learning framework that jointly captures equivariant and invariant representations for robust visual perception in dynamic environments.

Therefore the problem statement: Hybrid Equivariant-Invariant Feature Learning for visual perception in Dynamic visual environments.

Chapter 2

System Design

System design represents a critical stage in the development of the proposed visual perception framework, as it bridges the gap between theoretical concepts and practical implementation. The objective of this chapter is to present a clear, modular, and scalable design that translates the core idea of joint equivariant–invariant representation learning into an operational system.

The proposed system is designed to address the limitations of conventional convolutional neural networks, which often entangle geometric and semantic information within a single representation. Such entanglement leads to instability under real-world distribution shifts, including changes in viewpoint, orientation, and appearance. To overcome this limitation, the system is explicitly designed to separate geometric consistency and semantic stability into dedicated processing pathways while maintaining shared low-level feature extraction.

The system design emphasizes clear separation of functional responsibilities, efficient reuse of shared representations, robustness under geometric and photometric transformations, and extensibility for future enhancements.

2.1 Functional Block Diagram

The functional block diagram provides a high-level representation of the proposed system and illustrates the sequential flow of data through its major components. It serves as a conceptual blueprint that captures how raw visual data is transformed into robust equivariant and invariant representations.

The system begins with an input acquisition block, which receives raw RGB images from the dataset. These images are then passed to a preprocessing block, where resizing, normalization, and basic transformations are applied to standardize the input format and ensure stable numerical behavior during training.

The preprocessed images are fed into a shared feature extraction backbone, which is responsible for learning low-level and mid-level visual features such as edges, textures, and object parts. This backbone acts as a common foundation for both geometric and semantic reasoning, ensuring efficient parameter utilization.

From the shared backbone, the architecture branches into two parallel processing paths: an equivariant branch that preserves spatial structure and enforces predictable feature transformations under geometric changes, and an invariant branch that produces compact semantic embeddings that remain stable under appearance variations.

Each branch is optimized using a dedicated loss function, and the outputs are analyzed using representation-level metrics to evaluate geometric consistency and semantic stability. This modular block structure ensures clarity, interpretability, and robustness of the system.

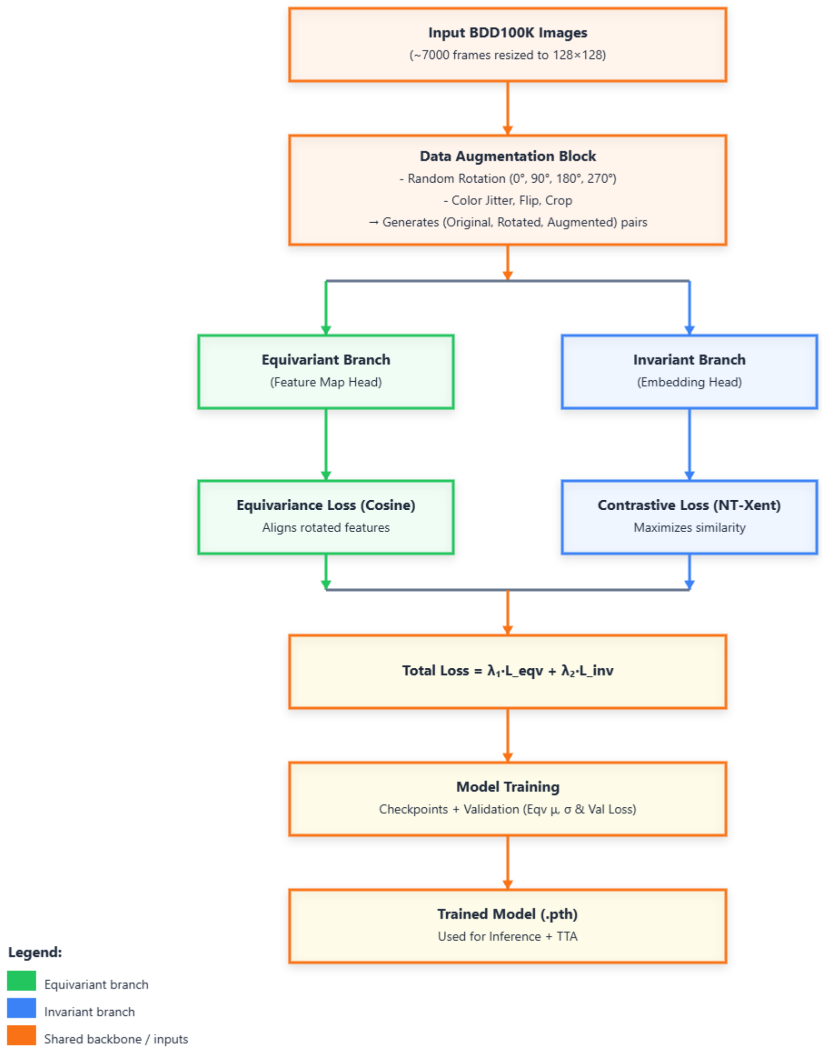
Input BDD100K Images
(~7000 frames resized to 128×128)
Data Augmentation Block
- Random Rotation (0°, 90°, 180°, 270°)
- Color Jitter, Flip, Crop
→ Generates (Original, Rotated, Augmented) pairs
Equivariant Branch
(Feature Map Head)
Invariant Branch
(Embedding Head)
Equivariance Loss (Cosine)
Aligns rotated features
Contrastive Loss (NT-Xent)
Maximizes similarity
Total Loss = $\lambda_1 \cdot L_{eqv} + \lambda_2 \cdot L_{inv}$
Model Training
Checkpoints + Validation (Eqv μ, σ & Val Loss)
Trained Model (.pth)
Used for Inference + TTA
Legend:
🟩 Equivariant branch
🟦 Invariant branch
🟧 Shared backbone / inputs

Figure 2.1: Block diagram of the proposed dual-branch equivariant-invariant framework.

2.2 Design Objectives and Constraints

The system design is guided by several key objectives and practical constraints that influence architectural decisions and optimization strategies.

The primary design objectives include explicit separation of equivariant and invariant feature learning, preservation of spatial structure under geometric transformations, semantic stability under appearance variations, computational efficiency suitable for real-world deployment, and extensibility for future enhancements.

The design constraints include limited computational resources during training, the need for self-supervised or weakly supervised learning, stable convergence under multi-objective optimization, and compatibility with standard deep learning frameworks.

2.3 Design Alternatives

Several architectural alternatives were explored before arriving at the final system design. One alternative involved a single-branch architecture with a combined loss function enforcing both equivariance and invariance. While simpler in structure, this approach resulted in entangled representations and limited interpretability.

Another alternative considered fully equivariant architectures using specialized convolutional layers designed for predefined symmetry groups. Although such architectures provide strong geometric guarantees, they introduce significant computational overhead and lack flexibility when applied to unconstrained real-world data.

A third alternative focused on purely invariant contrastive learning frameworks, which provide strong semantic stability but discard spatial information essential for tasks requiring geometric reasoning.

Through comparative analysis, these alternatives were found to be either computationally expensive, overly restrictive, or insufficiently robust, motivating the adoption of a dual-branch architecture with shared low-level features.

2.4 Final System Architecture

The final system architecture adopts a dual-branch hybrid design built upon a shared convolutional backbone. The shared backbone extracts common visual features from the input images, reducing redundancy and ensuring consistency across branches.

2.4.1 Equivariant Branch

The equivariant branch is designed to preserve spatial and geometric information. It processes backbone feature maps using convolutional layers that maintain spatial resolution. The branch is optimized to ensure that geometric transformations applied to the input induce predictable transformations in the feature space.

This explicit enforcement of equivariance ensures geometric consistency, which is essential for robust perception under viewpoint and orientation changes.

2.4.2 Invariant Branch

The invariant branch focuses on learning compact semantic representations. It employs global pooling operations followed by fully connected layers to remove spatial dependencies and produce transformation-insensitive embeddings.

Normalization is applied to stabilize contrastive learning and prevent representational collapse. This branch ensures semantic stability under appearance variations such as illumination changes and color distortions.

2.5 Summary of System Design

This chapter presented a comprehensive system design for a hybrid equivariant–invariant visual perception framework. By explicitly separating geometric and semantic learning objectives within a dual-branch architecture while sharing low-level features, the system achieves robustness, interpretability, and scalability. This design forms a strong foundation for the implementation and evaluation discussed in subsequent chapters.

Chapter 3

Implementation Details

This chapter presents the practical realization of the proposed dual-branch equivariant–invariant learning framework. It describes the complete implementation pipeline, including data preparation, feature extraction, branch-wise processing, loss formulation, and optimization strategy. While the previous chapter focused on architectural design principles, this chapter explains how those principles are translated into an operational system.

The implementation is designed to jointly learn geometric equivariance and semantic invariance in a stable and controlled manner, while maintaining modularity and extensibility for future research and deployment.

3.1 Data Preparation and Input Processing

The implementation begins with raw visual inputs obtained from a large-scale real-world driving dataset. All images are resized to a fixed spatial resolution to ensure uniformity across training batches and to maintain stable computational requirements. Pixel values are normalized using standard channel-wise statistics to align the input distribution with that expected by deep convolutional backbones.

The framework operates in a self-supervised manner and does not require semantic annotations during training. This enables the model to learn robust representations directly from visual data, making the approach suitable for large-scale and continuously evolving datasets.

3.2 Implementation Framework and Data Pipeline

The implementation follows an end-to-end self-supervised learning paradigm designed to operate directly on raw visual data without relying on explicit semantic annotations. Input images are first subjected to a preprocessing stage to ensure consistency and numerical stability. Each image is resized to a fixed spatial resolution and normalized using standard channel-wise statistics. This preprocessing step ensures that the input distribution aligns with the expectations of deep convolutional feature extractors and facilitates stable gradient-based optimization.

Once preprocessed, images are passed through a shared deep convolutional backbone responsible for extracting hierarchical visual features. The backbone serves as the common representation space for both equivariant and invariant learning objectives. Instead of freezing the backbone parameters, the implementation retrains the backbone jointly with both learning branches. This allows the feature extractor itself to adapt to geometric transformations and appearance variations, leading to representations that are simultaneously spatially structured and semantically meaningful.

During training, two distinct categories of transformations are applied to each input image. Geometric transformations, such as controlled discrete rotations, are applied to enforce equivariance by altering the spatial configuration of visual content. Appearance-based transformations, including color jittering and horizontal flipping, are applied to enforce invariance by modifying visual appearance while preserving semantic identity. These transformations are applied in a coordinated manner so that the relationship between original and transformed samples is explicitly known and exploited during optimization.

The processed images then flow through a unified pipeline consisting of feature extraction, branch-specific processing, loss computation, and parameter updates. This integrated pipeline ensures that both geometric consistency and semantic stability are optimized jointly at every training iteration.

3.3 Equivariant and Invariant Branch Implementation with Learning Objectives

The equivariant branch is designed to preserve spatial structure and enforce predictable feature transformations under geometric changes. Feature maps obtained from the shared backbone are forwarded to this branch, which produces spatially structured representations. To enforce equivariance, the system compares the feature map obtained from a geometrically transformed input with the correspondingly transformed feature map of the original input [8, 2].

Let x denote an input image, $T_k(\cdot)$ represent a geometric transformation such as a rotation by $k \times 90^\circ$, and $E(\cdot)$ denote the equivariant feature extractor. The equivariance objective is defined using a cosine similarity-based Equivariance Consistency Metric (ECM) [8]:

$$\mathcal{L}_{eqv} = 1 - \frac{1}{B}\sum_{i=1}^{B} \frac{\langle T_k(E(x_i)), E(T_k(x_i))\rangle}{\|T_k(E(x_i))\|_2 \, \|E(T_k(x_i))\|_2}, \tag{3.1}$$

where B denotes the batch size. Minimizing this loss encourages the feature representations to transform in a structured and predictable manner that mirrors input-level geometric transformations, thereby preserving spatial coherence.

In parallel, the invariant branch focuses on learning transformation-insensitive semantic embeddings. Feature maps from the shared backbone are spatially aggregated using global average pooling to remove positional dependencies [6]. The pooled features are then passed through a projection head to obtain compact embedding vectors. These embeddings are normalized to improve numerical stability and prevent degenerate solutions.

Semantic invariance is enforced using a contrastive learning objective that compares embeddings derived from different appearance-based augmentations of the same image [5]. Let z_i and z_i' denote embeddings obtained from two augmented views of the same input image. The Invariance Difference Metric (IDM) is defined as:

$$\mathcal{L}_{inv} = -\frac{1}{2B}\sum_{i=1}^{B} \log \frac{\exp\left(\frac{\langle z_i, z_i'\rangle}{\tau}\right)}{\sum_{j=1}^{2B} 1_{[j\neq i]} \exp\left(\frac{\langle z_i, z_j\rangle}{\tau}\right)}, \tag{3.2}$$

where τ is a temperature parameter controlling the sharpness of similarity comparisons. This objective ensures that embeddings corresponding to the same semantic content remain close despite appearance changes, while maintaining discrimination between different samples.

3.4 Training Dynamics

The final training objective combines the equivariant and invariant losses into a single multi-objective optimization problem. The total loss is expressed as a weighted sum:

$$\mathcal{L}_{total} = \lambda_{eqv}\mathcal{L}_{eqv} + \lambda_{inv}\mathcal{L}_{inv}, \tag{3.3}$$

where λ_{eqv} and λ_{inv} control the relative contribution of geometric and semantic learning objectives. These weights are carefully selected to ensure that strong geometric consistency does not suppress semantic learning and vice versa [4].

During backpropagation, gradients from both loss components are propagated through their respective branches and the shared backbone. This coordinated gradient flow enables the backbone to learn representations that simultaneously support equivariance and invariance without mutual interference. An adaptive gradient-based optimizer is employed to ensure stable convergence under this multi-objective setting [18].

Throughout training, validation monitoring is performed to ensure stable learning behavior and to prevent overfitting. The best-performing model is retained based on validation consistency, ensuring robustness of the learned representations. During inference, the trained framework operates without additional adaptation, providing spatially structured equivariant feature maps and stable invariant embeddings suitable for downstream perception tasks.

In summary, the implementation integrates controlled data transformations, shared feature extraction, branch-specific learning objectives, and joint optimization into a cohesive and extensible system. This practical realization successfully translates the theoretical motivation of equivariant and invariant learning into an effective framework for robust real-world visual perception.

3.5 Summary

This chapter detailed the implementation of the proposed hybrid equivariant–invariant learning framework. By integrating controlled data transformations, branch-specific learning objectives, and joint optimization through a shared backbone, the system effectively learns robust geometric and semantic representations. The modular and extensible design makes the implementation suitable for both academic research and real-world deployment.

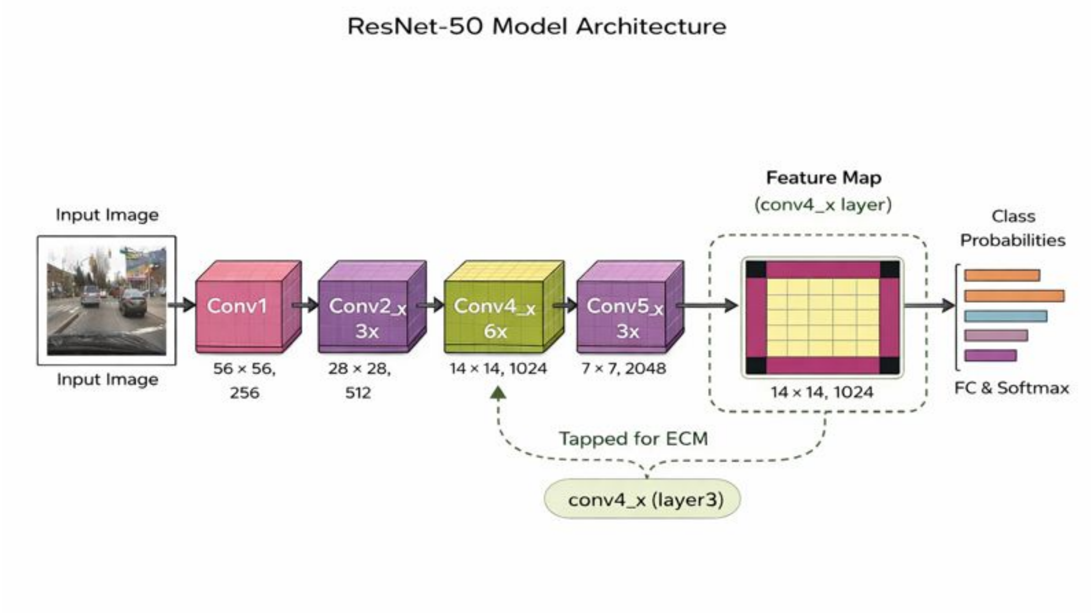

Figure 3.1: Model Architecture

Chapter 4

Result and Discussion

This chapter presents a comprehensive evaluation of the proposed dual-branch equivariant–invariant learning framework trained on the BDD100K dataset. The analysis examines training behavior, geometric equivariance, semantic invariance, and the interaction between these objectives under real-world distribution shifts. Both quantitative metrics and qualitative visual evidence are used to validate the effectiveness of the proposed approach.

4.1 Training and Validation Convergence

The model is trained for more than 50 epochs using a self-supervised learning strategy. From approximately Epoch 19 onward, the total training loss stabilizes within a narrow range, indicating convergence of the optimization process. Specifically, the training loss remains between 0.917 and 0.927, with representative values of 0.9255 at Epoch 49, 0.9250 at Epoch 50, and 0.9184 at Epoch 59.

The validation loss follows a stable trend, fluctuating moderately between 1.2468 and 1.5814 without exhibiting divergence or overfitting. The consistent gap between training and validation losses throughout the training process indicates robust generalization to unseen data. These convergence characteristics demonstrate that the proposed multi-objective learning formulation enables stable and reliable optimization.

4.2 Geometric Equivariance Consistency

Geometric equivariance is quantitatively evaluated using the Equivariance Consistency Metric (ECM), which measures how accurately feature representations transform in response to input-level geometric transformations [8]. Let x denote an input image, $T_k(\cdot)$ a rotation by $k \times 90°$, and $E(\cdot)$ the equivariant feature extractor. The ECM loss is defined as:

$$\mathcal{L}_{ECM} = 1 - \frac{1}{B} \sum_{i=1}^{B} \frac{\langle T_k(E(x_i)), E(T_k(x_i)) \rangle}{\|T_k(E(x_i))\|_2 \, \|E(T_k(x_i))\|_2}, \tag{4.1}$$

where B denotes the batch size.

Across all evaluated epochs, the equivariance loss converges to values close to zero for both training and validation sets. This behavior is consistently observed from Epoch 49 through Epoch 59, demonstrating that the equivariant branch successfully learns transformation-consistent feature mappings. The near-zero ECM loss confirms that spatial feature representations rotate predictably with the input, preserving geometric structure under viewpoint changes.

4.3 Semantic Invariance Stability

Semantic invariance is assessed using the Invariance Difference Metric (IDM), which evaluates the stability of learned embeddings under appearance-based transformations. Given two augmented views of the same image producing embeddings z_i and z_i', the IDM loss is defined using a normalized temperature-scaled cross-entropy formulation [5]:

$$\mathcal{L}_{IDM} = -\frac{1}{2B} \sum_{i=1}^{B} \log \frac{\exp\left(\frac{\langle z_i, z_i' \rangle}{\tau}\right)}{\sum_{j=1}^{2B} 1_{[j \neq i]} \exp\left(\frac{\langle z_i, z_j \rangle}{\tau}\right)}, \tag{4.2}$$

where τ is the temperature parameter.

Unlike the equivariant loss, the invariant loss remains active throughout training, reflecting continuous refinement of semantic representations. The training invariant loss stabilizes within the range of 0.917–0.927, while the validation invariant loss remains between 1.2468 and 1.5814. These trends indicate that the invariant branch successfully learns transformation-insensitive embeddings without collapsing to trivial solutions, while maintaining effective generalization across unseen appearance variations.

4.4 Quantitative Performance Summary

Table 4.1 summarizes the quantitative performance of the proposed framework after retraining the shared backbone to support joint equivariant and invariant feature learning. The results demonstrate strong geometric consistency alongside stable semantic representations across both training and validation sets.

4.5 Qualitative Visual Analysis

To complement the quantitative evaluation, qualitative visual comparisons between the baseline model and the proposed retrained framework are presented. Feature maps from the equivariant branch are visualized under multiple input rotations. The baseline model exhibits irregular and inconsistent feature responses, whereas the proposed framework produces structured and rotation-consistent feature transformations.

To further highlight differences between the baseline and retrained models, a feature-map difference visualization (Δ-map) is presented in Figure 4.4. The heatmap illustrates spatial activation differences, where warmer regions indicate larger deviations and cooler regions denote stability.

The localized and structured differences observed in the retrained model indicate improved geometric consistency, while the absence of widespread high-magnitude deviations suggests preserved semantic stability. These visual observations provide intuitive confirmation of the quantitative ECM and IDM results.

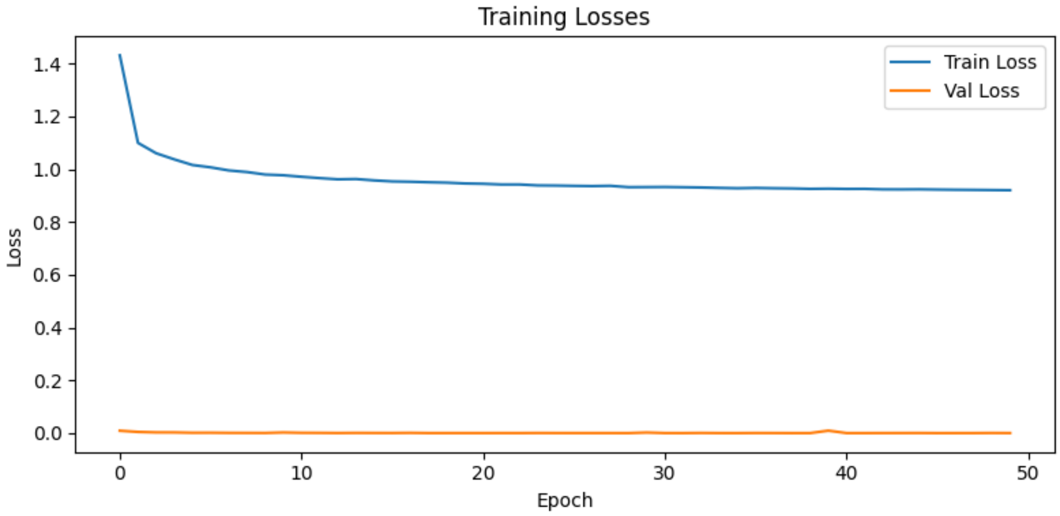

Figure 4.1: Training and validation loss convergence of the proposed dual-branch equivariant–invariant framework on the BDD100K dataset.

Table 4.1: Performance summary of equivariant and invariant learning.

Metric	**Training**	**Validation**
Total Loss Range	0.917–0.927	1.2468–1.5814
Equivariance Consistency (ECM)	≈ 0.0000	≈ 0.0000
Invariance Difference (IDM)	0.917–0.927	1.2468–1.5814

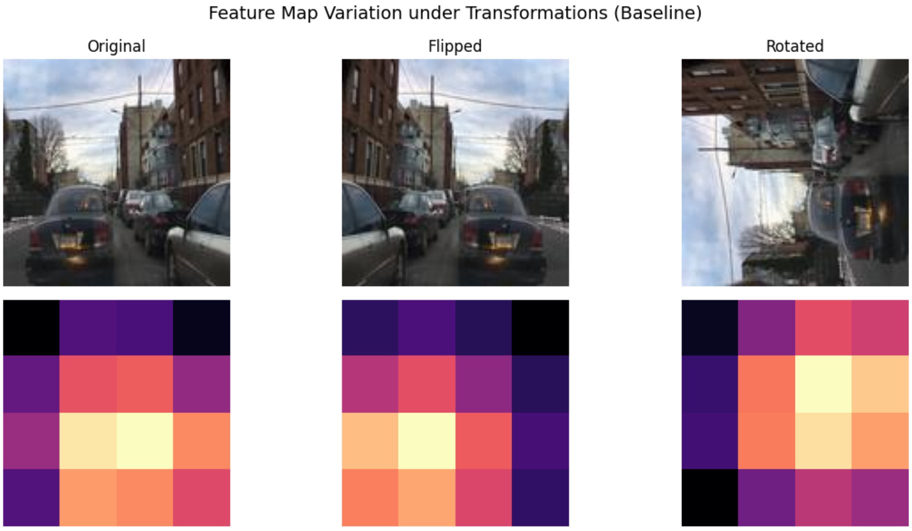

Figure 4.2: Feature map responses obtained from the baseline model under geometric transformations.

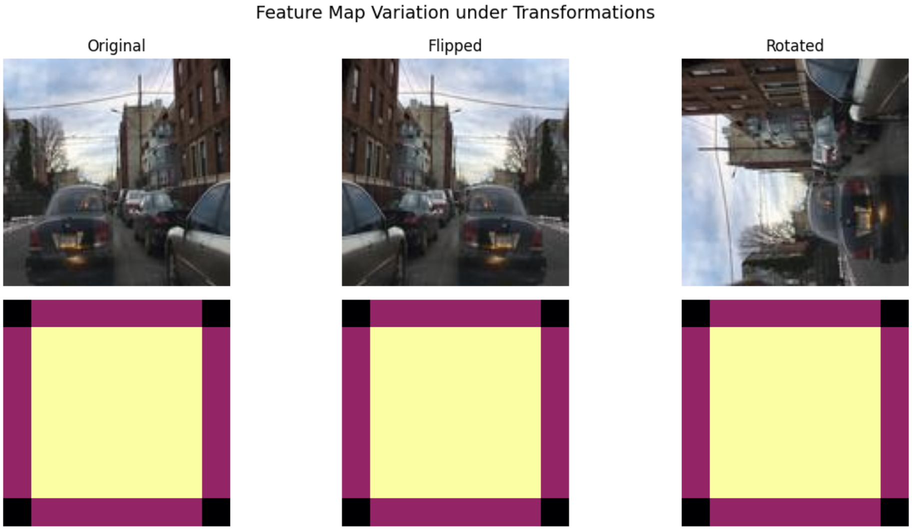

Figure 4.3: Feature map responses obtained from the proposed equivariant–invariant framework under geometric transformations.

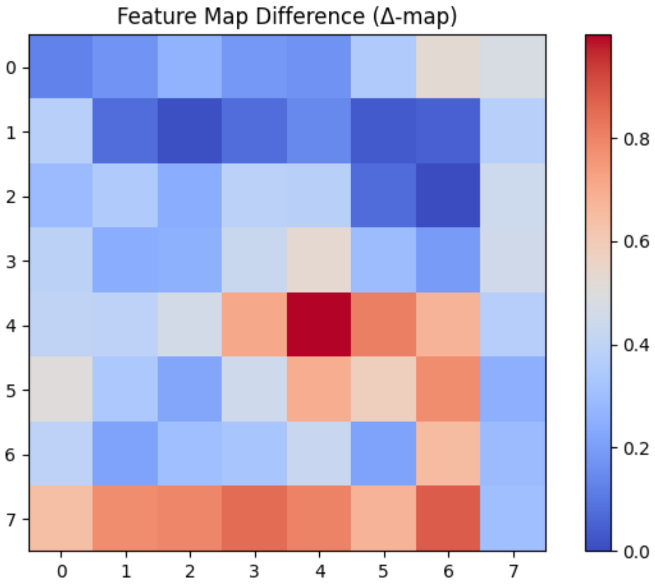

Figure 4.4: Δ-map highlighting feature response differences between baseline and re-trained models.

Chapter 5

Conclusion and future scope

This chapter summarizes the overall outcomes of the proposed work and reflects on the key insights gained through system design, implementation, and experimental evaluation. It consolidates the findings derived from both quantitative metrics and qualitative analyses to assess the effectiveness of the proposed dual-branch equivariant–invariant framework. In addition, the chapter outlines potential directions for future research and system enhancement, highlighting opportunities for extending the framework to more complex transformations, broader application domains, and advanced real-world perception tasks.

5.1 Conclusion

This work presented a dual-branch hybrid learning framework that explicitly integrates equivariant and invariant representations to enhance robustness in real-world visual perception systems. By retraining a shared backbone and separating geometric and semantic learning objectives into dedicated branches, the proposed approach effectively addresses the limitations of conventional convolutional neural networks under distribution shifts. Experimental results on the BDD100K dataset demonstrate stable convergence, strong geometric consistency, and robust semantic stability when compared to a baseline model. Quantitative ECM and IDM metrics, along with qualitative visual analyses, confirm that the proposed framework learns structured feature transformations while preserving invariant semantic representations.

5.2 Future Scope

Future research will explore extending the framework to additional transformation groups such as scale, perspective, and non-rigid deformations, as well as incorporating temporal equivariance for video-based perception tasks. Investigating adaptive or test-time loss balancing strategies and evaluating the learned representations on downstream tasks such as object detection and tracking remain promising directions for further improvement.

Bibliography

[1] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**, 2278–2324.

[2] Cohen, T.; Welling, M. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 2990–2999.

[3] Weiler, M.; Cesa, G. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Rizve, M.N.; Khan, S.H.; Hayat, M.; Khan, F.S. Invariant and equivariant representations for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[6] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Yu, F.; Chen, H.; Wang, X.; et al. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Lenc, K.; Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 991–999.

[9] Goodfellow, I.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[10] Azulay, A.; Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 2019, **20**, 1–25.

[11] Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. Improved baselines with momentum contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[14] Weng, T.; Zhang, H.; Chen, P.Y.; et al. Towards fast computation of certified robustness for deep neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

[15] Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[16] Bronstein, M.M.; Bruna, J.; Cohen, T.; Velickovic, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[17] Esteves, C.; Allen-Blanchette, C.; Zhou, X.; Daniilidis, K. Learning SO(3) equivariant representations with spherical CNNs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.