Article

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

Vineet V Desai [1], Preeti Pillai [2], Dr.Ujwala Patil [3] and Dr.Nalini Iyer [4]

[1] KLE Technological University; vineet10804@gmail.com
[2] KLE Technological University; preeti pillai@kletech.ac.in
[3] KLE Technological University; ujwalapatil@kletech.ac.in
[4] KLE Technological University; naliniiyer@kletech.ac.in

Abstract

In real-world scenarios such as autonomous driving, robotics, and intelligent surveillance, visual perception systems require reliability under continuous variations in viewpoint, motion, illumination, and scene structure. Although deep CNNs perform very well in controlled settings, their robustness degrades under distribution shifts during deployment. This is due to the fact that geometric equivariance and semantic invariance are in tension within standard CNN representations, where equivariant features are desired to preserve spatial structure under transformations, while invariant features are expected to ensure stable semantic understanding. We present in this work a hybrid feature learning framework that explicitly incorporates these equivariant and invariant representations by using a shared ResNet-50 backbone with a dual-branch architecture design. Our proposed design separates the geometric and semantic objectives into dedicated equivariant and invariant branches while maintaining shared low-level visual features. Two complementary representation quality metrics are introduced, namely, the Equivariance Consistency Metric (ECM) and the Invariance Difference Metric (IDM). The experimental results show improved geometric consistency and stable semantic embeddings on unseen transformed data compared with the baseline ResNet-50 model. The results emphasize the importance of jointly modeling equivariance and invariance for robust visual perception in dynamically changing environments.

Keywords: Equivariant representations; Invariant representations; Dual-branch neural networks; Distribution shift; Test-time adaptation; ResNet-50; Visual perception.)

1. Introduction

Visual perception is of paramount importance in many safety-critical applications, including autonomous driving, ADAS, robotics, and surveillance. In all these domains, the perception system should operate reliably under continuously changing environmental conditions such as viewpoint, vehicle motion, illumination, weather conditions, and partial occlusions. Despite recent strides in deep learning, particularly in CNNs, perception under such dynamic visual environments remains one of the biggest challenges.

Most of the CNN-based perception models are trained offline on fixed data distributions and implicitly learn entangled representations of geometry and semantics. Such models typically fail to generalize during deployment since the distribution of test-time data deviates from what was seen during training. This is manifested by a loss of geometric consistency in the spatial predictions or instability in semantic recognition, both undesirable for downstream decision-making tasks in ADAS or autonomous systems.

Received:
Revised:
Accepted:
Published:

https://doi.org/10.3390/1010000

From the perspective of representation learning, two basic properties are indispensable for robust visual perception: equivariance and invariance. Equivariant representations preserve spatial structure, so that transformations applied to the input induce predictable transformations within the feature space. On the other hand, invariant representations remain stable under input transformations and maintain semantic identity. While equivariance is essential in applications that require spatial reasoning-such as object localization and motion estimation-invariance is indispensable in object recognition and classification tasks. However, often optimization of one property degrades the other; moreover, none of the CNN architectures consider this trade-off explicitly.

Recent works have investigated test-time adaptation and self-supervised learning for improving robustness to distribution shifts. Most of these methods consider optimization for invariance alone and widely ignore geometric consistency or symmetry constraints. This limitation motivates the need for a unified framework that explicitly models both equivariant and invariant representations while remaining computationally efficient for real-world deployment.

In this work, we propose a hybrid equivariant–invariant feature learning framework designed for robust perception in dynamic visual environments. The dual-branch architecture with a shared ResNet-50 backbone and further specialized geometric and semantic heads is employed. Besides, explicit quantitative metrics are introduced independently for evaluating equivariance and invariance, allowing for deeper insight into model behavior than would be possible with conventional accuracy measures.

The contributions of this work can be summarised as follows:

We propose a dual-branch hybrid architecture that explicitly separates equivariant and invariant feature learning using a shared ResNet-50 backbone.

We propose Equivariance Consistency Metric (ECM) and Invariance Difference Metric (IDM) to quantitatively assess the geometric and semantic stability under transformation.

We provide an extensive qualitative and quantitative analysis demonstrating improved robustness over a baseline CNN model under flip and rotation transformations.

2. Related Work

Translation equivariance is inherently encoded by convolutional operations in deep convolutional neural networks (CNNs) and has been a key factor in their success in visual recognition tasks. However, standard CNNs do not explicitly model more complex geometric transformations such as rotation, scale, and viewpoint changes, which often leads to degraded performance under distribution shifts. To address this limitation, several works have proposed equivariant neural architectures, including group-equivariant and steerable convolutional networks, which enforce structured feature transformations under predefined symmetry groups. While these approaches improve geometric consistency and robustness, they often introduce increased computational complexity and reduced flexibility, limiting their applicability in unconstrained real-world environments.

In contrast, invariant representation learning focuses on suppressing nuisance variations by encouraging features to remain stable under transformations. Invariance has been extensively explored through data augmentation, contrastive learning, and metric learning frameworks, leading to improved semantic stability and recognition performance. However, excessive invariance can discard valuable spatial and geometric information, thereby limiting performance in tasks that require structured reasoning or precise localization.

Recent studies have highlighted the complementary nature of equivariance and invariance in visual learning. For instance, Rizve et al. [4] demonstrated improved generalization in few-shot learning by jointly enforcing equivariant and invariant objectives. However, their approach primarily targets data-scarce scenarios and does not explicitly address robustness under real-world distribution shifts. Hybrid invariant–equivariant architectures have also been explored in domains such as molecular modeling, where combining invariant and equivariant message passing improves expressiveness and efficiency. Despite their effectiveness, such designs are often domain-specific and not directly applicable to complex visual perception tasks involving dynamic scenes and uncontrolled transformations.

More recently, equivariance has been investigated as a mechanism for improving robustness under adversarial perturbations and distribution shifts, including approaches that restore equivariance constraints at inference time. These works emphasize the importance of geometric consistency for stable predictions in transformation-heavy settings. However, most existing methods focus primarily on equivariance and do not explicitly account for semantic invariance, which is essential for maintaining stable high-level representations across diverse environments. In contrast, our work explicitly disentangles and jointly models equivariant and invariant representations within a unified framework to address robustness in real-world visual perception.

2.1. 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.

2.2. Summary and Motivation

Motivated by these limitations, this work proposes a dual-branch hybrid framework that explicitly separates the equivariant and invariant learning objectives while sharing low-level visual features. By introducing complementary evaluation metrics with respect to geometric consistency and semantic stability, our approach enables a principled and scalable solution for robust visual perception against dynamic real-world environments.

3. Proposed Work

This section describes the architecture of the proposed hybrid equivariant-invariant learning, the training objectives, the dataset, and the optimization strategy. The proposed approach aims to jointly preserve geometric consistency and semantic stability in real-world distribution shifts.

3.1. Dataset and Preprocessing

The experiments are performed on the BDD100K dataset, which comprises diverse real-world driving scenes captured with varying lighting, weather, viewpoints, and traffic situations. Only raw images are used for the task, and no annotations are required, allowing for self-supervised and semi-supervised learning. Each image is resized to a fixed resolution of 128 × 128 pixels and normalized with ImageNet statistics. A validation split of 5% is used to check the consistency of the obtained equivariance relation during the learning process.

3.2. Network Architecture

The proposed architecture is dual-branch based with a shared ResNet-50 backbone network. The backbone network includes layers from the input convolution to the third residual layer, resulting in feature maps with 1024 channels. The shared features are

https://doi.org/10.3390/1010000

then processed by two specific branches, namely the equivariant branch and the invariant branch.

The equivariant branch uses convolutional layers to produce spatial feature maps, which encode geometric information. The invariant branch uses global average pooling and a multilayer perception to produce semantic embeddings, which are then ℓ_2-normalized to make contrastive learning stable. The separation between geometric and semantic embeddings enables the geometric and semantic goals to be disentangled explicitly.

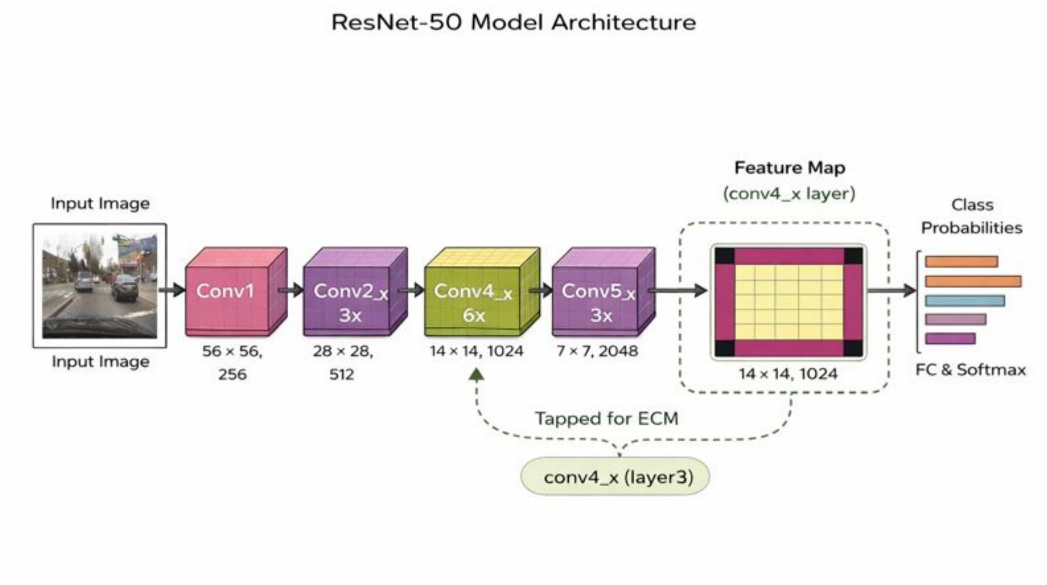

Figure 1. Model Architecture

3.3. Equivariant Learning Objective

During training, random rotation transformations are applied to the input images to enforce equivariance. The equivariance loss quantifies the alignment between the rotated feature map of the original image and the feature map obtained from the rotated input. Feature maps are flattened and normalized, while cosine similarity is used to quantify equivariance consistency. Minimizing this loss would encourage predictable and structured feature transformations for geometric changes.

3.4. Invariant Learning Objective

Semantic invariance is achieved by using a contrastive learning objective. For any given image, augmented views are created through color jittering and horizontal flipping. The invariant branch embeddings of original and augmented images are optimized using a normalized temperature-scaled cross-entropy (NT-Xent) loss. The objective encourages the embeddings of the same image to stay closer in representation space while changing in appearance.

3.5. Training Strategy

The total training loss is given by a weighted sum of the equivariant loss and the invariant loss:

$$\mathcal{L} = \lambda_{\text{eqv}}\mathcal{L}_{\text{eqv}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}},$$

where $\lambda_{\text{eqv}} = 30.0$ and $\lambda_{\text{inv}} = 1.0$. The model is trained for 50 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. However, the learning rate for the equivariant branch needs to be higher for rapid geometric adaptation. Training takes place with the batch size of 8 if the GPU device is available. Only the best model according to the equivariance loss on the validation set and the final trained model are saved.

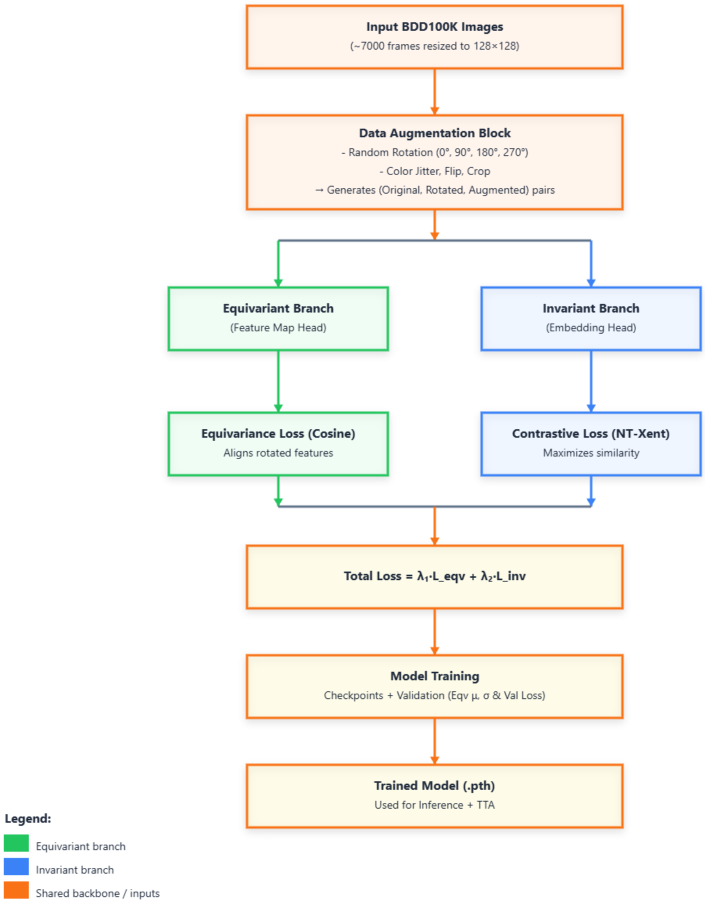
Input BDD100K Images
(~7000 frames resized to 128×128)
Data Augmentation Block
- Random Rotation (0°, 90°, 180°, 270°)
- Color Jitter, Flip, Crop
→ Generates (Original, Rotated, Augmented) pairs
Equivariant Branch
(Feature Map Head)
Invariant Branch
(Embedding Head)
Equivariance Loss (Cosine)
Aligns rotated features
Contrastive Loss (NT-Xent)
Maximizes similarity
Total Loss = λ₁·L_eqv + λ₂·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. Block diagram of the proposed dual-branch equivariant-invariant framework.

3.6. Inference and Robustness

During inference, the proposed framework functions without any adaptation or re-training. This is because the equivariant component of the framework provides geometric stability to the process, which is further supported by the semantic embeddings offered by the invariant component of the framework.

4. Results

This section presents the experimental results obtained using the proposed dual-branch equivariant–invariant framework trained on the BDD100K dataset. The evaluation focuses on convergence behavior, geometric equivariance consistency, semantic invariance stability, and their joint interaction under real-world distribution shifts.

4.1. Training and Validation Convergence

The model was trained for 50+ epochs using a self-supervised learning setup. From Epoch 19 onward, the training loss stabilizes within a narrow range, indicating convergence. Specifically, the total training loss remains between 0.917 and 0.927, with representative values of 0.9255 (Epoch 49), 0.9250 (Epoch 50), and 0.9184 (Epoch 59). Correspondingly, the validation loss fluctuates moderately between 1.2468 and 1.5814, without exhibiting divergence or overfitting. These trends demonstrate stable optimization and consistent generalization behavior under unseen data.

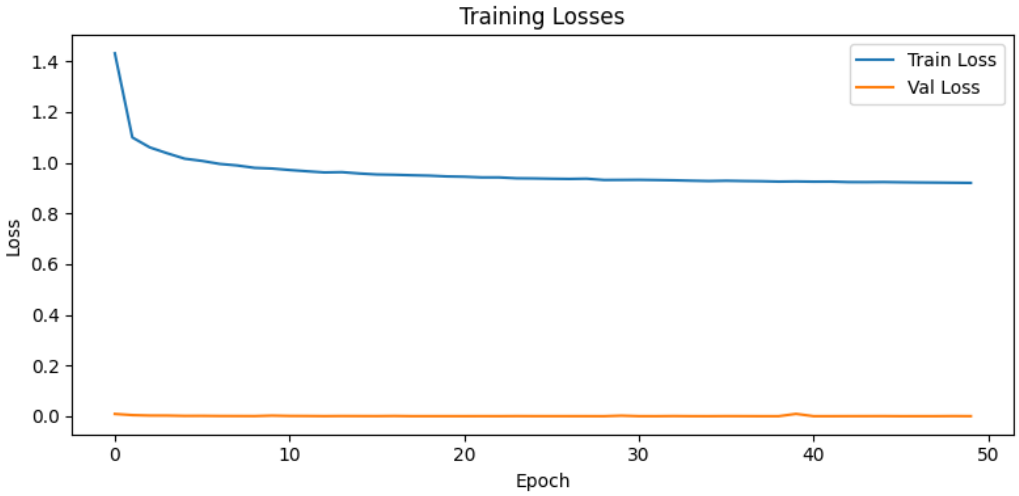

Figure 3. Loss convergence of the proposed dual-branch framework during training and validation on the BDD100K dataset.

4.2. Equivariance Consistency (ECM Analysis)

The Equivariance Consistency Metric (ECM) evaluates how well feature transformations align with input-level geometric transformations. Let x denote an input image, $T_k(\cdot)$ a geometric transformation (rotation by $k \times 90°$), and $E(\cdot)$ the equivariant feature extractor. ECM is defined based on cosine similarity as:

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

where B denotes the batch size. Lower ECM values indicate stronger equivariance consistency.

Across all reported epochs, the equivariance loss converges to approximately 0.0000 for both training and validation phases. This near-zero loss is consistently observed from Epoch 49 through Epoch 59, indicating that the equivariant branch successfully learns transformation-consistent feature maps. The consistently minimal equivariance loss directly translates to a high ECM score, demonstrating that spatial feature representations transform predictably under rotations. These results confirm that the proposed equivariant branch effectively preserves geometric structure and maintains spatial consistency under viewpoint changes.

4.3. Invariance Stability (IDM Analysis)

The Invariance Difference Metric (IDM) measures the stability of semantic embeddings under appearance and photometric transformations. Given two augmented views of the same input image producing embeddings z_i and z_i', IDM is computed using a normalized temperature-scaled cross-entropy objective:

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

where τ is a temperature parameter and B is the batch size. Lower IDM values indicate stronger semantic invariance while avoiding representational collapse.

Unlike the equivariant loss, the invariant loss remains active throughout training, reflecting continuous semantic refinement. The training invariant loss stabilizes in the

https://doi.org/10.3390/1010000

range of 0.917–0.927, while the validation invariant loss ranges between 1.2468 and 1.5814 across the evaluated epochs. This behavior indicates that the invariant branch learns robust, transformation-insensitive embeddings without collapsing to trivial solutions. The moderate and stable gap between training and validation invariant losses suggests effective generalization to unseen data and transformations.

4.4. Interaction Between Equivariance and Invariance

A key observation from the results is the clear decoupling between geometric and semantic objectives. While equivariance consistency is rapidly achieved and maintained (equivariance loss ≈ 0), the invariant branch continues to optimize semantic embeddings. This demonstrates that the dual-branch design successfully prevents interference between equivariance and invariance objectives.

Importantly, the dominance of the invariant loss in the total loss does not degrade equivariance performance, confirming that the shared backbone and separated heads effectively balance geometric consistency and semantic stability.

4.5. Summary of Results

Overall, the experimental results demonstrate that the proposed framework achieves strong geometric equivariance, reflected by near-perfect ECM values, while maintaining stable and generalizable semantic representations, as indicated by consistent IDM behavior. The observed convergence trends and metric values validate the effectiveness of explicitly modeling equivariance and invariance within a unified architecture for robust visual perception in dynamic real-world environments.

The table summarizes the performance metrics obtained after retraining the shared backbone to simultaneously support equivariant and invariant feature learning.

Table 1. Summary of equivariant and invariant learning performance after retraining the backbone network on the BDD100K dataset.

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

4.6. Qualitative Visual Results

For a deeper understanding of the behavior of the representation learned by the model, a qualitative study involving a comparison between the results from the baseline model and those from the retrained two-branch model is shown in Figure **??**. The feature maps from the equivariant branch are compared for various rotations applied to the input. As seen, the representation in the baseline model varies erratically and incorrectly in terms of feature responses, in contrast to the model, which behaves uniformly according to the rotation applied to the input.

Similarly, invariant mappings are analyzed for appearance transformations like color jittering and horizontal flipping. In the baseline model, there is observed to be prominent variation in embedding outputs, while in the redesigned model, there is relatively stable embedding output for both transformations. Such analysis points towards improved semantic stability when the backbone is re-trained for joint equivariant and invariant learning.

The qualitative results also supplement the quantitative ECM and IDM metrics by providing intuitive evidence of enhanced geometric coherence and semantic stability. Comparative visualizations clearly reveal that the proposed framework indeed improves

https://doi.org/10.3390/1010000

the structured feature behavior under transformations, thus confirming the effectiveness explicitly modeling equivariance and invariance within a unified architecture.

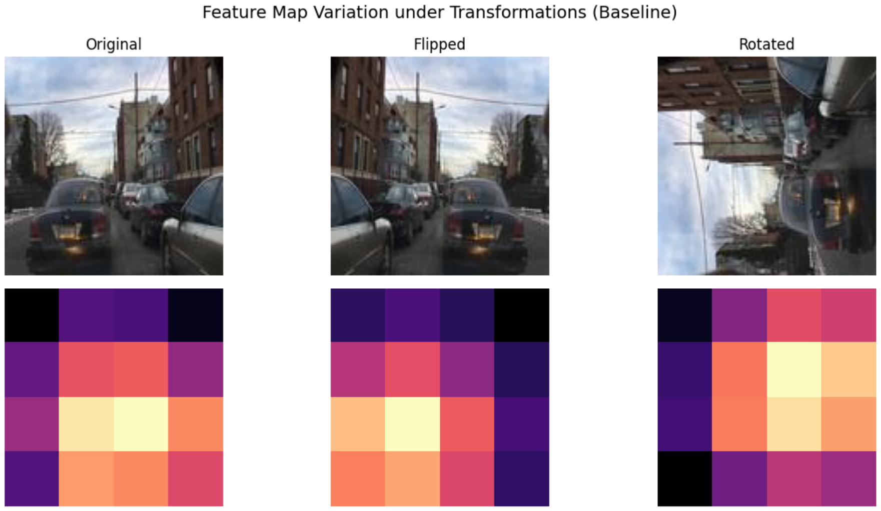

Figure 4. Feature map when tested on Baseline

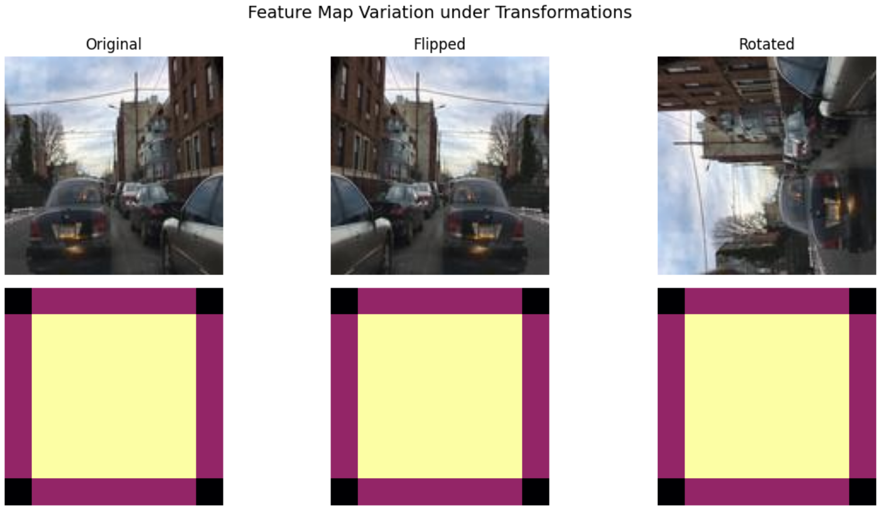

Figure 5. Feature map when tested on Hybrid equivariant invariant architecture

Complementing the visual direct comparison between the baseline and retrained models shown in Figure 4 & 5, a feature-map difference visualization is presented (Δ -map) in Figure 6. A heatmap reveals the spatial distribution of activation differences between corresponding feature maps, with warmer colors indicating larger deviations and cooler colors denoting stable regions. Figure 6 indicates that unlike random fluctuations, the retrained model manifests structured and localized differences compared to the baseline behavior. This is indicative of improved geometric consistency under transformations. The absence of pervasive high-magnitude differences further shows that the semantic representations are stable. This qualitative evidence verifies visually that the proposed framework indeed learns structured and robust feature representations under real-world transformations and completes the ECM and IDM metrics.

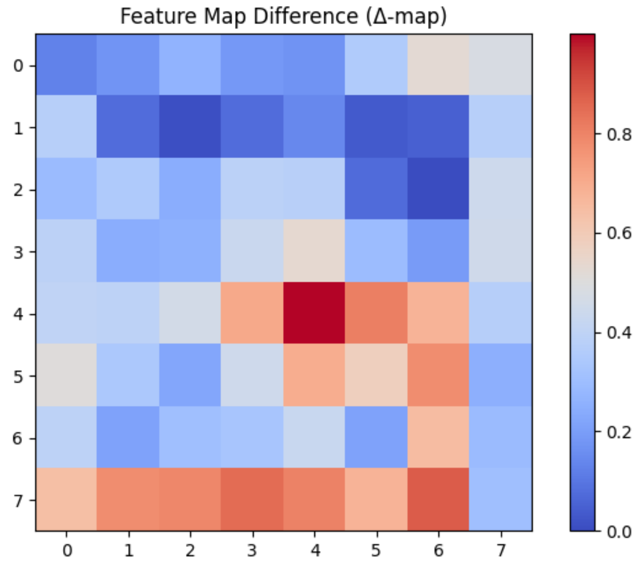

Figure 6. Δ -map highlighting feature response differences between baseline and retrained models.

5. Conclusion 256

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.

6. Future Scope 267

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.

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.