Exploiting Graph-Based Structural Priors for Visual Recognition

Depanshu Sani

IIIT-Delhi, India
depanshus@iiitd.ac.in

Abstract

Nature is inherently structured! The entities in the real world are naturally organized in rich relationships. For example, dolphins and sharks, despite their striking visual resemblance in body shape and fins, are actually from entirely different branches of the animal hierarchy, i.e., mammals and fishes, respectively. This remarkable similarity is a prime example of 'convergent evolution', where unrelated species develop similar features because they face similar environmental challenges. This illustrates how nature's underlying organization often transcends superficial visual resemblances. Although humans intuitively grasp and utilize these profound natural constraints, they are typically underutilized in most AI systems. As a result, trained AI models tend to align with statistical patterns in the data, such as sampling biases or class imbalance, rather than adhering to the underlying relational consistency. This thesis argues that AI systems must evolve beyond learning *"flat"* feature representations, which are domain-agnostic and derived purely from data correlations, to *"explicitly model the domain-specific structural relationships"*. A key benefit of encoding relational priors in the learning process is that it can inject domain knowledge as an inductive bias, leading to more robust and reliable models. My research investigates incorporating domain knowledge by leveraging *"graph-based structural priors"* that explicitly model relational constraints in various visual recognition tasks. This work spans three distinct dimensions of visual recognition, progressing from coarse-level (image-level) to fine-grained (scene-level) understanding.

My research highlights a crucial limitation in existing AI models: they often fail to incorporate real-world constraints, leading to significant errors. I show that even powerful, pretrained neural networks can make severe mistakes due to a lack of domain knowledge. I argue that standard metrics like top-1 accuracy, precision and recall are insufficient for evaluating model robustness, and propose a new metric based on rank order of the predictions as a better indicator of reliability. The benchmark on various large-scale datasets confirms that existing solutions do not sufficiently capture the domain knowledge, which is often available as a taxonomy tree, motivating our design of better learning frameworks. I also examine complex visual re-identification (Re-ID) tasks, such as monitoring animals in the wild. I find that existing foundational models struggle with new species and environments. This challenge is compounded by the high cost of manual annotation for adapting these systems to new settings. While existing unsupervised learning methods can help reduce the need for extensive labeling, they often suffer from under- and over-segmentation errors, which led me to develop more effective active learning strategies. Finally, I address the limitations of the classic Kalman filter, a widely used tool for dynamic systems. I point out that this filter makes a flawed assumption that the movement of each individual object is independent of its dynamic surroundings. In the real world, this is rarely the case. I demonstrate the need for a new filtering mechanism that not only considers an object's past movements but also its spatial relationship with other dynamic entities in its environment.

In my analysis, I observed the vision foundation models for all recognition tasks, i.e., classification, detection and segmentation, lack the domain knowledge. I believe that our learning framework, which was designed specifically for classification, can be adapted for other recognition tasks. I speculate that a unified learning framework can be designed that can be leveraged for making vision foundation models aware of the available taxonomy.

Related Works & Thesis Contributions

Image-Level Recognition (Leveraging Hierarchical Relations): Most of the image classification algorithms operate under the assumption that all negative classes are equally unrelated to the target class, effectively treating each class independently. This oversimplification can lead to feature representations where errors are arbitrary and unstructured. Semantic knowledge is often available but remains underutilized. By ignoring these inter-class relationships, despite showing impressive performance, classification algorithms can drastically fail when they make mistakes, which makes them especially risky in safety-critical applications. Incorporating hierarchy-aware feature representations in such a scenario offers a significant advantage by structuring the feature space based on semantic relationships. This property has been extensively studied and applied to mitigate failures in safety-critical systems. Such representations ensure that classes with greater semantic similarity are closer in the feature space, enabling more meaningful predictions even in the event of errors. I make the following contributions:

1. **Hierarchically Ordered Preference Score:** I highlight crucial shortcomings in the hierarchical evaluation metrics , and introduce a novel, ranking-based metric for

evaluating the performance of hierarchical classifiers (Sani and Anand 2025).

2. **Learning Hierarchy-Aware Features with Auxiliary Classifiers:** I learn additional auxiliary classifiers for classifying images at coarser levels (Garg, Sani, and Anand 2022), but with improved hierarchical consistency and therefore, performance. I show that auxiliary classifiers also allow learning a unified feature representation for all hierarchical levels, thereby resulting in a hierarchy-aware feature representation.
3. **Learning to Transform Flat Features into a Hierarchy-Aware Feature Space:** I also propose a novel framework that learns to map deep feature embeddings into a vector space defined using fixed orthogonal bases, that is, by design, consistent with the structure of a given taxonomy tree (Sani and Anand 2025). This helps reduce the overhead introduced because of learning auxiliary classifiers and the complex loss function.

Instance-Level Recognition (Leveraging Edge Relations in Graphs): Adapting pre-trained Re-ID systems to unseen environments remains challenging due to the high cost of manual annotation. Since all images of an individual must be manually separated out from a huge collection of images, collecting labeled data at scale is also labor-intensive and error-prone. A natural way to approach this is to frame supervision in terms of pairwise annotations, where annotators simply indicate whether two images are of the same individual. Unfortunately, this labeling process is expensive and time-consuming as it requires identifying matching pairs within a large dataset, where the number of comparisons grows quickly with the dataset size. To mitigate the cost of exhaustive annotation, unsupervised Re-ID approaches rely on pseudo-labeling via clustering, where the model is iteratively refined using automatically generated labels from the structure of the feature space. However, the effectiveness of pseudo-labeling hinges critically on cluster purity. Errors in clustering can introduce harmful supervision. Broadly, clustering errors can be categorized into two types: under-segmentation, where samples from different identities are erroneously grouped into the same cluster, and over-segmentation, where samples from the same identity are incorrectly split into multiple clusters. Both scenarios are detrimental—under-segmentation leads to identity confusion and suppresses inter-class separability, while over-segmentation fragments identity representations and weakens intra-class consistency. These challenges highlight the fragility of unsupervised pseudo-labeling and underscore the importance of guiding the learning process using reliable supervisory signals. Active learning (AL) is thus employed to identify and label the most informative examples under a constrained annotation budget. However, because correctly identifying the clustering algorithm and its hyperparameters for unlabeled data is non-trivial, existing AL Re-ID methods are biased towards the inductive bias of the chosen clustering algorithm. To this end, I make the following contributions:

1. **Non-Parametric, Plug-and-Play Constrained Clustering:** I introduce NP3, which is a post-hoc constrained clustering algorithm and is agnostic to the underlying clustering technique used to obtain the labels (Sani, Khurana, and Anand 2025). Specifically, it enforces *must-link* and *cannot-link* constraints between pairs of data points.
2. **AL for Visual Re-ID with Ambiguity-Aware Sampling:** I propose an Ambiguity-Aware Sampling (AAS) strategy that leverages disagreements between two complementary clustering algorithms to identify and sample the most informative, uncertain and diverse pairs of images for annotation, reducing both under- and over- segmentation errors (Sani, Khurana, and Anand 2025).

Scene-Level Recognition (Leveraging Spatio-Temporal Relations): Many Kalman filter-based Multi-Object Tracking (MOT) approaches assume the independence of object trajectories, overlooking potential inter-object relationships. For instance, at a busy urban intersection, an autonomous vehicle can utilize the trajectory of an oncoming vehicle that is decelerating to anticipate the presence and movement of pedestrians who are temporarily occluded by a passing truck. This allows the autonomous vehicle to make safer and more informed navigation decisions, even when direct visibility is compromised. While some efforts have been made to incorporate these relationships, they often concentrate on learning feature representations to facilitate better association. Moreover, the existing filter-based method for estimating graphs from noisy data is unsuitable for online MOT applications. To alleviate these problems, I introduce **Sensor-Agnostic Graph-Aware Kalman Filter** (Sani et al. 2025), which is the first online state estimation technique designed to fuse multi-modal graphs derived from noisy multi-sensor data. Specifically, I propose a novel dynamical model that captures the inter-object relationships in the form of a time-varying and topology-aware state-transition function on graph nodes. This dynamical model helps better estimate the state of the objects in the presence of noisy data.

My Contributions

In all the works, all the student co-authors helped in setting up the baselines and running our experiments. Thesis supervisor was involved in all the brainstorming sessions. I was involved in everything else, like, problem formulation, finding literature, writing code and designing experiments.

References

Garg, A.; Sani, D.; and Anand, S. 2022. Learning Hierarchy Aware Features for Reducing Mistake Severity. In *Computer Vision – ECCV 2022*, 252–267. Cham: Springer Nature Switzerland. ISBN 978-3-031-20053-3.

Sani, D.; and Anand, S. 2025. Learning and Evaluating Hierarchical Feature Representations. arXiv:2503.07853.

Sani, D.; Iyer, A.; Rai, P.; Anand, S.; Srivastava, A.; and Kalyanaraman, K. 2025. Sensor-Agnostic Graph-Aware Kalman Filter for Multi-Modal Multi-Object Tracking. In *Pattern Recognition*, 380–398. Cham: Springer Nature Switzerland. ISBN 978-3-031-78444-6.

Sani, D.; Khurana, M.; and Anand, S. 2025. Active Learning for Animal Re-Identification with Ambiguity-Aware Sampling. In *The 40th Annual AAAI Conference on Artificial Intelligence*.