



Universiteit
Leiden
The Netherlands

Exploring deep learning for data-efficient and generalizable visual recognition

He, K.

Citation

He, K. (2026, September 22). *Exploring deep learning for data-efficient and generalizable visual recognition*. Retrieved from <https://hdl.handle.net/1887/4310276>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4310276>

Note: To cite this publication please use the final published version (if applicable).

Chapter 7

Conclusions

In this thesis, we have devoted previous four research chapters to addressing four research questions regarding the three themes: metric learning, feature enrichment and knowledge transfer. In this chapter, we formulate the main findings from our approaches and results. In addition, we discuss the limitations of our methods and possible solutions to address them. Lastly, we point out several trends for future work.

7.1 Main Findings

In each research chapter, we have proposed a new approach to answer the corresponding research question. Next, we conclude these approaches and present the main findings inspired by empirical and theoretical analyses.

(1) We start the research part in Chapter 3 with proposing a new yet practical prototype similarity network (ProtoSimNet) for few-shot learning (FSL). A key difficulty in FSL is that each new class is represented by only a few labeled examples; as a result, the estimated class prototypes can be noisy and easily perturbed, so even small changes in the support set may noticeably shift prototypes in the embedding space and blur the decision boundary. This makes the choice of metric critical: a metric that is overly sensitive to such shifts can cause unstable predictions across tasks. To alleviate these issues, we propose a refined metric formulation, named cosine sharpening transformation, which focuses comparison on angular (directional) similarity and sharpens the similarity distribution to increase separation between classes in the decision space. Intuitively, this makes classification depend less on accidental variations in feature magnitude and more on consistent directional structure, which helps preserve clearer margins when prototypes are estimated from very limited data. On top of this, we introduce an elastic regularization term that constrains the deformation of the embedding space, preventing unstable shifts in feature geometry while still keeping sufficient flexibility for adapting to new categories.

(2) After investigating the metric learning theme, we now move our focus to address the feature enrichment theme in Chapter 4. In few-shot settings, the model must generalize from very limited supervision, so relying on a single feature level can be fragile. To more effectively capture the structural diversity required for robust few-shot recognition, we introduce an attention-based feature pyramid network (AFPN), which enriches the model’s representational capability by integrating multi-level features extracted from different layers of the backbone. AFPN uses an adaptive top-down attention mechanism to select and fuse informative spatial scales in a task-aware manner, allowing the model to combine local details with higher-level semantics. Experimental results across standard few-shot benchmarks demonstrate that the proposed AFPN significantly improves the discriminability and stability of features under limited supervision.

(3) In Chapter 5 we further explore the efficient strategies for enriching visual representations in FSL scenarios. Traditional meta-learning pipelines often aggregate features in a fixed way, but under few-shot supervision a fixed fusion strategy can be suboptimal because different tasks may require different spatial resolutions. Beyond traditional meta-learning approaches that often rely on a fixed feature aggregation strategy, we propose a novel adaptive multi-level feature fusion (AMFF) framework to strengthen representation diversity. AMFF dynamically integrates and recalibrates feature representations across multiple spatial resolutions, enriching the feature distribution while preserving the flexibility needed for adapting to new categories with only a few labeled samples. We believe that the AMFF can be extended to a wide range of representation-oriented tasks, such as fine-grained recognition, cross-domain FSL, and lightweight deep learning.

(4) After investigating the themes of metric learning and feature enrichment, we then turn to address the questions about the knowledge transfer theme for FSL. A central challenge in few-shot learning is that novel classes provide too little data to learn a strong representation from scratch; therefore, performance often depends on how effectively prior knowledge can be transferred. However, transferring everything indiscriminately can be harmful when the transferred information is not equally relevant to the new task. In Chapter 6, we introduce a dual selective knowledge transfer (DSKT) framework that transfers rich, efficient knowledge to guide the learning of novel categories. By emphasizing task-relevant knowledge while suppressing less useful ones, DSKT produces more effective distilled representations, yielding significant improvements over existing approaches and establishing new competitive baselines for FSL.

7.2 Limitations

Our methods in this thesis have addressed four research questions and achieved promising results in terms of three important research themes. However, they still

have some limitations which can be discussed from the following three perspectives.

7.2.1 Algorithmic perspective

In Chapter 3, ProtoSimNet integrates cosine sharpening transformation and elastic regularization to effectively strengthen inter-class prototype distinction. However, its effectiveness depends on the stable estimation of the prototype variance, which is inherently fragile in FSL settings. In addition, the fixed sharpen factor may suppress the model’s adaptability across diverse few-shot scenarios. A potential direction is to develop an adaptive sharpening mechanism that dynamically adjusts the transformation parameters based on episode difficulty.

In Chapter 4, AFPN demonstrates its improvements in recognition performance by integrating multi-level features with attention-guided fusion. Yet the top-down attention propagation increases computational cost as the pyramid depth and feature resolution grow. One possible optimization is to implement hierarchical attention mechanisms that progressively refine the features fused with reduced computational cost.

In Chapter 5, AMFF improves feature expressiveness by enriching representations through parallel convolutional branches. However, its multi-branch convolutions inevitably increase the parameter budget. Potential improvements include investigating dynamic neural architectures that adaptively activate different branches based on input complexity, or developing knowledge distillation techniques to compress the multi-branch knowledge into more compact representations.

In Chapter 6, DSKT facilitates knowledge transfer by selectively extracting efficient feature representations while simultaneously filtering out task-specific noise that may influence generalization. However, this selectivity requires the student to compare its representations with those of the teacher at multiple points, which increases the overall computational cost. A potential solution for future research is to develop in-place selective distillation, enabling the student to obtain relevant information without reconstructing every component of the teacher model.

7.2.2 Practical perspective

In Chapter 3, ProtoSimNet demonstrates its effectiveness in tightening class boundaries, yet it is sensitive to highly distorted feature distributions —as commonly occurs in long-tailed or domain-shifted datasets. A promising direction is to incorporate uncertainty-aware sharpening, where cosine sharpening transformation is modulated by confidence estimates or Bayesian feature variance to avoid excessive margin tightening.

In Chapter 4, AFPN enhances feature representation through multi-scale feature pyramids, but storing all scale outputs introduces a high demand for memory space. This makes it difficult to deploy AFPN on real-time applications where computational and memory resources are limited. A potential solution is exploring a more lightweight and efficient pyramid designs to maintain multi-scale benefits while fitting to practical constraints.

In Chapter 5, AMFF enriches feature diversity through multi-branch convolutions, but the additional branches introduce noticeable extra computational cost when applied to deeper backbones. One practical alternative is to reduce redundant branches or employing dynamic fusion mechanisms.

In Chapter 6, DSKT transfers selective teacher knowledge to improve generalization, but comparing student and teacher representations still requires excessive computational cost. A potential option is to distill compressed teacher features—such as low-rank projections or compact codes—allowing the student to rely on lightweight embedding descriptors rather than full feature matching.

7.3 Future Research Directions

In the previous four chapters, we have introduced several methods addressing the key research questions within the three major themes. These explorations also provide a range of promising directions for further advancement. In this section, we briefly discuss future research directions regarding each theme.

7.3.1 Metric Learning

Metric learning is a practical solution for FSL and has obtained noticeable performance in past years. However, most existing metrics are fixed once trained, making it difficult for the model to adapt to new tasks across episodes. New works, such as adaptive prototype networks [327] and prototype-rectification strategies [328], show that episode-dependent or distribution-aware metrics can significantly improve robustness. Therefore, a practical direction for future research is to design adaptive metric formulations that adjust distance functions based on support-set structure, or cross-domain statistics, rather than using a single global metric.

7.3.2 Feature Enrichment

Recent work has shown that feature enrichment approaches can significantly improve FSL performance by mitigating the lack of feature diversity. However, most existing enrichment methods require increased memory usage. A realistic challenge is to design enrichment modules that can enhance representation quality without requiring additional backbone stages or heavy multi-branch operations. A potential solution is to explore implicit feature enrichment, where fine-grained representations

are strengthened without explicitly expanding network width or depth. Recent studies demonstrate that lightweight modulation layers can effectively reshape feature distributions for few-shot generalization [329], cross-layer aggregation enhances hierarchical consistency across backbone stages [294], and self-augmenting features generated directly in latent space provide a parameter-efficient way to boost discriminative capability [330]. These insights indicate that implicit enrichment can obtain significant improvements in representation quality while keeping lower computational costs and memory usage.

7.3.3 Knowledge Transfer

Knowledge transfer for FSL has recently evolved from simple pretraining and fine-tuning to more sophisticated distillation-based frameworks, including dual selective distillation and multi-teacher knowledge transfer [331]. There is also growing interest in combining FSL with knowledge distillation in NLP and multi-modal settings, for instance, distilling few-shot capabilities from large language models into smaller learners [332]. Yet, a critical challenge arises in resource-constrained environments, where storing the full teacher model is impractical. Thus, it will be promising to explore compressed teacher representations or episodic subspace transfer, enabling stable knowledge transfer without excessive memory or computation.