Few-shot learning in industrial applications

MOLEK Vojtech and ALIJANI Zahra

Institute for Research and Applications of Fuzzy Modeling
University of Ostrava, Dvorakova 7., 701 03, Ostrava
Czech Republic
E-mail: {vojtech.molek, zahra.alijani}@osu.cz

Abstract

This paper reports on the empirical performance of few-shot learning (FSL) for visual defect classification using confidential industrial datasets. We evaluate 16 combinations of four backbone models (Perception Encoder, DINOv2, DINOv3, ConvNeXt-v2) and four FSL classifiers (Prototypical Networks, Neighborhood Component Analysis, Relation Networks, Linear Adapter). The evaluation covers three conditions: a baseline comparison, deterministic support set augmentation, and a learnable attention preprocessor. Results demonstrate that support set augmentation is a highly effective strategy, improving performance in nearly all configurations. Furthermore, the DINOv2 and ConvNeXt-V2-T backbones emerged as top performers, achieving the most competitive and highest-accuracy results, respectively. These findings suggest that for industrial FSL applications, combining a strong, pre-trained backbone with a simple augmentation strategy is a practical approach for building data-efficient classification systems.

1 Introduction

The deployment of deep learning for industrial visual quality control is often challenged by data scarcity, making it difficult to collect the large labeled datasets required for traditional supervised learning. Few-shot learning (FSL) addresses this by enabling models to generalize from very few examples. This work serves as a practical report on the application of FSL to this domain. We conduct a systematic empirical evaluation of 16 combinations of modern backbone architectures and specialized FSL classifiers on three confidential, real-world datasets for binary classification. We compare four classifiers (Prototypical Networks [5], Neighborhood Component Analysis [6], Relation Networks [7], and Linear Adapter) paired with four pre-trained vision models (Perception Encoder (PE) [1], DINOv2 [2], DINOv3 [3], ConvNeXt-v2 [4]). Our goal is to provide practical comparisons of these FSL pipelines for industrial environments.

2 Methodology

Our methodology is designed as a systematic, comparative study. We evaluate four backbone models paired with four FSL classifiers across three confidential datasets, under various experimental conditions.

Acknowledgement The contribution has been funded from the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583", which is co-financed by the European Union. This article has been produced with the financial support of the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition.

https://www.doi.org/10.15452/978-80-7599-515-5.2026.14

2.1 Datasets

We utilize three confidential industrial datasets for binary visual defect classification (defected/non-defected) of varying resolution, size, and difficulty. Each experiment was conducted with a balanced training set of 20 samples per class, totaling 40 samples, and a whole unbalanced test set.

2.2 Backbone Architectures

We selected four pre-trained models to serve as feature extractors, covering both Vision Transformer (ViT) and Convolutional Neural Network (CNN) architectures. All models are used with frozen weights to simulate a realistic FSL scenario where the backbone is not fine-tuned.
PE$_{CORE}$T. A Vision Transformer from the Perception Encoder family.
DINOv2 (ViT-S/14). A ViT model pre-trained with the DINOv2 self-supervised learning method.
DINOv3 (ViT-S+/16). A newer generation ViT model from the DINOv3 family.
ConvNeXt-V2-T. A modern CNN architecture, serving as a convolutional baseline.

2.3 Few-Shot Learning Classifiers

We evaluate four distinct classifiers that attach to the frozen backbones.
Prototypical Networks. A metric-learning method that computes a single prototype vector for each class as the mean of its support embeddings. Classification is performed by finding the nearest class prototype.
Neighborhood Component Analysis (NCA). An instance-based metric-learning algorithm that learns a distance metric to maximize the probability of correct classification by soft k-NN. It retains all support samples for prediction.
Relation Networks. A method that learns a deep, non-linear similarity metric. A relation module is trained to output a similarity score between query and support examples.
Linear Adapter Classifier. A fine-tuning approach where a simple linear layer, preceded by a small trainable adapter, is trained on top of the frozen backbone. This serves as a simple baseline against the other metric-learning approaches.

3 Experiments

We conducted a series of experiments to evaluate the 16 model-classifier pairs. All evaluations are performed in a 40-shot setting.

3.1 Experimental Setups

1. Baseline Evaluation. A direct comparison of all 16 model-classifier pairs to establish baseline performance on the three datasets.
2. Support Set Augmentation. We investigate the effect of expanding the support set using deterministic data augmentation. Each of the 40 support samples is augmented to create 9 additional versions, resulting in a 10x larger support set for training the classifier.
3. Attention-based Preprocessing. To address challenges with variable aspect ratios, we evaluate a learnable ‘AttentionPreprocessor‘ module. This module is inserted before the backbone and learns to generate a soft attention map, focusing on relevant image regions while avoiding aspect ratio distortion.

3.2 Evaluation Metrics

To provide a quantitative comparison, we use a set of metrics designed to evaluate both absolute performance and relative competitiveness across the three datasets. For each of the 16 model-classifier combinations, we calculate the following:

Average Raw Accuracy. The mean of the balanced test accuracy scores across the three datasets. This measures the expected absolute performance.

Average Normalized Gap. This metric measures how consistently competitive a model is. For each dataset, we first calculate a normalized gap: Normalized Gap $= \frac{\text{max\_acc} - \text{model\_acc}}{\text{max\_acc} - \text{min\_acc}}$ where 'max_acc' and 'min_acc' are the maximum and minimum accuracies achieved by any combination on that dataset. This score, which ranges from 0 (best) to 1 (worst), is then averaged across the three datasets.

Table 1: Aggregated summary of results. For each combination, 'Accuracy' and 'Norm. Gap' show the baseline value followed by the best value achieved. 'Method' indicates the experiment that yielded the best result. Select performers are highlighted.

Model-Classifier Combination	Accuracy (%)	Method	Norm. Gap
PE_{CORE}T - Prototypical Net.	81.15 / 86.77	Augmented	0.50 / 0.60
PE_{CORE}T - NCA	86.01 / 86.01	Baseline	0.32 / 0.32
PE_{CORE}T - Relation Net.	72.36 / 85.75	Augmented	0.75 / 0.75
PE_{CORE}T - Linear Adapter	80.98 / 83.74	Augmented	0.62 / 0.86
DINOv2 - Prototypical Net.	90.03 / 91.62	Augmented	0.09 / 0.22
DINOv2 - NCA	87.59 / 91.29	Augmented	0.20 / 0.18
DINOv2 - Relation Net.	77.88 / 92.30	Augmented	0.60 / 0.17
DINOv2 - Linear Adapter	87.98 / 89.54	Augmented	0.15 / 0.46
DINOv3 - Prototypical Net.	84.26 / 86.40	Augmented	0.33 / 0.47
DINOv3 - NCA	85.72 / 87.72	Augmented	0.25 / 0.43
DINOv3 - Relation Net.	85.19 / 86.89	Augmented	0.31 / 0.60
DINOv3 - Linear Adapter	85.51 / 88.76	Augmented	0.35 / 0.42
ConvNeXt-V2-T - Prototypical Net.	84.38 / 91.18	Augmented	0.34 / 0.51
ConvNeXt-V2-T - NCA	86.09 / 90.50	Augmented	0.27 / 0.49
ConvNeXt-V2-T - Relation Net.	87.97 / 92.92	Augmented	0.20 / 0.34
ConvNeXt-V2-T - Linear Adapter	78.51 / 89.84	Augmented	0.60 / 0.59

4 Discussion

An analysis of the results in Table 1 reveals two clear findings. First, support set augmentation is highly effective, improving performance in 15 of 16 configurations. Second, the choice of backbone is critical, with **DINOv2** and **ConvNeXt-V2-T** emerging as the top performers. The 'ConvNeXt-V2-T - Relation Net.' pairing achieved the highest accuracy (92.92%), while 'DINOv2' models delivered the most consistently competitive (lowest) normalized gaps. The experiment involving adaptive preprocessing did not yield significant improvements in this context.

5 Conclusion

This empirical study found that the most effective strategy for the evaluated industrial few-shot learning tasks was combining a modern backbone with support set augmentation. The results indicate that **DINOv2** and **ConvNeXt-V2-T** are strong feature extractors, and that augmentation is a critical step for improving performance. For practitioners, these findings suggest that focusing on these elements is a promising approach for data-efficient quality control.

References

[1] Bolya, Daniel, et al. *Perception encoder: The best visual embeddings are not at the output of the network.* arXiv preprint (2025) 2504.13181.

[2] Oquab, Maxime, et al. *Dinov2: Learning robust visual features without supervision.* arXiv preprint (2023) 2304.07193 .

[3] Darcet, Timothée, et al. *Vision transformers need registers.* arXiv preprint (2023) 2309.16588.

[4] Woo, Sanghyun, et al. *Convnext v2: Co-designing and scaling convnets with masked autoencoders.* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023).

[5] Snell, Jake, Kevin Swersky, and Richard Zemel. *Prototypical networks for few-shot learning.* Advances in neural information processing systems **30** (2017).

[6] Goldberger, Jacob, et al. *Neighbourhood components analysis.* Advances in neural information processing systems **17** (2004).

[7] Sung, Flood, et al. *Learning to compare: relation network for few-shot learning.* Proceedings of the IEEE conference on computer vision and pattern recognition (2018).