Effectiveness of Proxy-Based Metric-Learning for Large-Scale Stock-Keeping Unit Recognition

Barbara Borkowska[1][0009−0009−0746−5613] and
Grzegorz Sarwas[1,2][0000−0003−4113−2387]

[1] Warsaw University of Technology, Faculty of Electrical Engineering,
Pl. Politechniki 1, 00-661 Warsaw, Poland
{barbara.borkowska.stud, grzegorz.sarwas}@pw.edu.pl
[2] Omnishelf Sp. z o.o. grzegorz@omnishelf.io

Abstract. Automatic recognition of retail products at the stock-keeping unit (SKU) level is a fundamental component of modern retail automation, enabling applications such as self-checkout, automated shelf monitoring, and inventory auditing. From a computer vision perspective, SKU recognition poses significant challenges due to extreme class cardinality, strong inter-class similarity, and substantial intra-class variability arising from real-world acquisition conditions. Deep metric learning offers a scalable alternative to conventional classification-based approaches; however, traditional pair- and triplet-based losses exhibit limited scalability and unstable optimisation in large-class settings. In this paper, we conduct a controlled empirical study of proxy-based metric learning methods for large-scale SKU recognition under realistic retail constraints. We compare a classical contrastive baseline with representative single- and multi-proxy objectives, including ProxyNCA, Proxy Anchor, SoftTriple, and a hierarchical proxy formulation. Experiments on the public RP2K benchmark and a real-world in-house retail dataset, covering both dense and sparse class distributions, demonstrate that proxy-based formulations consistently outperform contrastive learning across training regimes. The relative ranking of methods remains stable across initialisation strategies and dataset regimes, indicating robustness of the observed performance trends. In particular, multi-proxy approaches achieve the most favourable trade-off between retrieval accuracy, convergence speed, and computational efficiency, supporting their suitability for scalable industrial SKU recognition systems.

Keywords: Retail product recognition · SKU identification · Proxy-Based methods · Metric learning.

1 Introduction

Automatic recognition of retail products at the stock-keeping unit (SKU) level is a foundational component of modern retail automation. Applications such as self-checkout systems, automated shelf monitoring, inventory auditing, and planogram compliance require the reliable identification of thousands of visually similar products captured under unconstrained in-store conditions. From

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

a computer vision perspective, SKU recognition constitutes a large-scale fine-grained visual recognition problem characterised by extreme class cardinality, subtle inter-class differences, and substantial intra-class variability caused by viewpoint changes, illumination and white-balance variation, occlusions, and packaging reflections.

Conventional deep convolutional neural networks trained for closed-set classification have achieved remarkable success on large-scale benchmarks. In this setting, a model assigns each input to one of a fixed set of predefined classes using a parametric classifier (e.g., softmax). However, their direct application to SKU-level recognition presents several challenges. Retail catalogues evolve continuously, new products are frequently introduced, and maintaining balanced, fully annotated datasets at SKU granularity is costly and labour-intensive. Moreover, the number of product classes may reach several thousand, making conventional classification approaches increasingly inefficient and difficult to scale in both training and inference.

Deep metric learning offers a principled and scalable alternative by learning embedding spaces in which visually similar products are mapped close together, while dissimilar products are separated. In this formulation, SKU recognition is naturally treated as an image retrieval task, where a query image is matched against a reference catalogue by searching for nearest neighbours in the embedding space. Unlike classification-based approaches, retrieval-based formulations enable open-set recognition, support incremental catalogue updates, and avoid retraining large classification heads when new products are introduced.

However, traditional metric learning objectives based on contrastive or triplet losses suffer from limited scalability in large-class settings due to the combinatorial growth of sample pairs or triplets and the need for carefully designed mining strategies.

Proxy-based metric learning has emerged as a scalable solution to these limitations. By replacing instance-level comparisons with learnable class representatives (proxies), such methods reduce computational complexity and improve optimisation stability. Approaches such as ProxyNCA [8], ProxyNCA++ [17], Proxy Anchor [5], and SoftTriple [11] have demonstrated strong performance on fine-grained benchmarks. Nevertheless, their behaviour under realistic retail constraints—characterised by severe class imbalance, limited samples per SKU, long-tail distributions, and domain-specific acquisition noise—remains insufficiently explored.

In this work, we provide a controlled empirical evaluation of proxy-based metric learning methods for stock-keeping unit recognition under realistic retail constraints. Unlike prior evaluations conducted primarily on generic fine-grained benchmarks, we explicitly analyse optimisation behaviour under long-tail distributions, sparse per-class sampling, and varying backbone initialisation regimes.

The main contributions of this paper are as follows:

– We provide a systematic, large-scale empirical comparison of contrastive and proxy-based metric learning objectives for stock-keeping unit recognition under realistic retail conditions, including both public and real-world datasets;

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

– We analyse the impact of class cardinality, data sparsity, and backbone initialisation on optimisation dynamics, convergence stability, and retrieval performance across different data regimes;
– We evaluate the scalability and computational characteristics of proxy-based methods, providing practical insights for real-world deployment.

2 Related Work

2.1 Deep Metric Learning

Deep metric learning (DML) aims to learn embedding spaces in which semantically similar samples are mapped close together and dissimilar samples are separated. Early approaches relied on Siamese architectures trained with contrastive loss [2] or triplet loss [13], which explicitly enforce pairwise or triplet-based distance constraints. Although effective in face recognition and re-identification tasks, these methods suffer from limited scalability due to the combinatorial growth of sample pairs and the need for carefully designed hard-negative mining strategies [4].

To mitigate sampling inefficiencies, several structured objectives were proposed. Lifted Structured Loss [16] leverages all positive and negative pairs within a batch to improve gradient utilisation, while the N-pair loss [14] extends triplet learning to multi-class settings by jointly comparing one positive and multiple negatives. Despite improved optimisation behaviour, such approaches remain sensitive to batch composition and scale poorly to problems involving thousands of classes.

More recently, contrastive formulations have been widely adopted in self-supervised representation learning. Methods such as SimCLR [1] and MoCo [3] maximise the agreement between the augmented views of the same instance while treating other samples as negatives. Although these approaches achieve strong representation quality on large-scale datasets, they typically require very large batch sizes or memory banks and are not explicitly tailored to fine-grained, large-class retrieval scenarios such as SKU recognition.

Comprehensive empirical analyses of DML training strategies have further highlighted the sensitivity of metric learning to sampling schemes, embedding dimensionality, and optimisation settings [12, 9]. These findings underscore the importance of scalable and stable objectives for industrial-scale applications.

2.2 Proxy-Based Metric Learning

To improve scalability, proxy-based metric learning methods replace instance-level comparisons with learnable class representatives, referred to as proxies. ProxyNCA [8] introduced the idea of optimising the distances between samples and class proxies rather than between individual instances, thus reducing computational complexity and improving convergence stability.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

Subsequent developments refined this formulation. Proxy Anchor [5] restructured the optimisation process around proxy-centric anchors with similarity-weighted gradients, improving the global embedding structure. SoftTriple [11] addressed the limitations of single-proxy representations by modelling each class with multiple proxies, allowing for a more expressive modelling of intra-class variability. Hierarchical proxy formulations further extend this paradigm by explicitly capturing multi-level semantic relationships between classes.

Proxy-based losses have demonstrated strong results on standard fine-grained benchmarks such as CUB and Stanford Online Products (SOP) [15]. However, their behaviour under realistic retail constraints characterised by severe class imbalance, long-tail distributions, sparse per-class sampling, and acquisition noise remains insufficiently explored.

2.3 Retail Product and SKU Recognition

Retail product recognition constitutes a challenging fine-grained visual recognition problem due to high inter-class similarity, subtle packaging variations, and continuously evolving product catalogues. The RP2K dataset [10] was introduced as a large-scale benchmark for SKU-level recognition under realistic in-store conditions, while earlier fine-grained retrieval benchmarks such as Stanford Online Products [15] have played a central role in evaluating large-class metric learning systems.

Recent research has explored multiple learning paradigms tailored to SKU recognition under different data regimes. Prototype-based one-shot learning approaches, such as the Variational Prototyping Encoder (VPE), demonstrate that learning from a single prototype image per class can yield competitive performance in retail environments [6]. However, such methods primarily target low-data scenarios and rely on generative prototype modelling.

In parallel, contrastive and self-supervised learning approaches have been investigated for SKU recognition [7], demonstrating strong representation quality on both public benchmarks and real-world retail datasets. These studies also highlight challenges in dataset composition, including near-duplicate samples across splits in RP2K, which may affect evaluation protocols.

More broadly, prior work in retail recognition has explored both classification-based and retrieval-oriented formulations for applications such as automated checkout and shelf monitoring. However, despite these advances, systematic analyses of scalable metric learning strategies across varying levels of class cardinality, data sparsity, and backbone initialisation remain limited. In particular, comparative studies examining the optimisation behaviour and scalability trade-offs of modern proxy-based objectives in realistic industrial settings are scarce.

These gaps motivate the controlled empirical investigation presented in this paper.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

3 Methodology

This section describes the experimental methodology used to evaluate proxy-based metric-learning methods for large-scale SKU recognition in retail environments. The emphasis is on reproducibility, scalability, and relevance to real-world deployment scenarios.

3.1 Problem Definition

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes an image of a retail product and $y_i \in \{1, \ldots, C\}$ is its corresponding SKU label, the objective is to learn an embedding function

$$f_\theta : x \to \mathbf{z} \in \mathbb{R}^d, \tag{1}$$

parameterised by θ, such that images of the same SKUs are mapped close together in the embedding space, while the embeddings of different SKUs are well separated.

This can be formulated as an optimisation problem:

$$\min_\theta \ \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}} \left[\mathcal{L}(f_\theta(x_i), y_i)\right], \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes a metric learning loss function that enforces small distances between embeddings of samples sharing the same label and large distances otherwise. Formally, for an embedding space $(\mathcal{Z}, d)$ with distance function $d(\cdot, \cdot)$, the objective is to satisfy:

$$d(f_\theta(x_i), f_\theta(x_j)) \ll d(f_\theta(x_i), f_\theta(x_k)), \tag{3}$$

for $y_i = y_j$ and $y_i \neq y_k$.

We formulate SKU recognition as a retrieval problem rather than a closed-set classification task. In classification settings, a model assigns each input to one of a fixed set of predefined classes using a parametric classifier, which limits scalability and requires retraining when new classes are introduced. In contrast, the retrieval formulation operates in the learned embedding space, where a query image is matched against a reference catalogue by searching for its nearest neighbours according to a similarity measure (cosine similarity in our experiments).

During evaluation, query images are matched against a gallery constructed from the validation split, ensuring no overlap with training data in the class-disjoint setting. This formulation naturally supports open-set recognition, incremental catalogue updates, and deployment without retraining large classifiers, which are critical requirements in retail systems.

3.2 Datasets

Experiments were conducted on two datasets that represent complementary retail recognition scenarios.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

RP2K Dataset The RP2K dataset [10] is a large-scale public benchmark for retail product recognition, comprising more than 380,000 images across approximately 2,400 SKU classes. The images are captured under realistic in-store conditions, including cluttered backgrounds, varying illumination, occlusions, and multiple viewpoints. Each SKU is represented by a relatively large number of images, making RP2K suitable for evaluating the scalability and convergence behaviour of metric learning methods.

In-House Retail Dataset The in-house dataset was collected in operational retail environments and comprises over 3,300 SKU classes and approximately 32,000 images. In contrast to RP2K, this dataset is characterised by a low average number of images per class and pronounced class imbalance. As such, it reflects a realistic industrial scenario in which data acquisition and annotation are expensive, and newly introduced products are underrepresented.

Data Splits and Augmentation For RP2K, we follow the standard train–validation split provided by the dataset authors, in which the partition is predefined and commonly adopted in prior work. For the in-house dataset, a class-disjoint split is applied, ensuring that SKUs used for evaluation are not observed during training. This setting assesses the generalisation capability of learned embeddings to unseen products.

During training, images are resized and augmented using random rotations, colour jittering, and brightness and contrast variations. These augmentations are selected to mimic common sources of appearance variation in retail environments, such as changes in camera angle and non-uniform lighting.

3.3 Network Architectures

All experiments reported in this study are conducted using ResNet-50 as the convolutional backbone. The model follows a two-stage architecture consisting of a convolutional feature extractor and a metric embedding head.

The final classification layer of ResNet-50 is replaced with a fully connected projection layer producing d-dimensional embeddings, followed by ℓ_2 normalisation. This modification enables the network to operate in a deep metric-learning setting rather than in a standard classification regime.

ResNet-50 was selected due to its strong representational capacity, stable optimisation behaviour, and widespread use as a baseline architecture in metric learning research. In addition, it provides a favourable balance between accuracy and computational efficiency, making it suitable for large-scale retail product recognition.

3.4 Metric Learning Loss Functions

This study evaluates representative metric learning objectives using a fixed ResNet-50 backbone. The analysed losses span classical pair-based formulations

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

and modern proxy-based approaches, differing in similarity modelling, optimisation dynamics, and scalability to large SKU vocabularies. By keeping the backbone fixed, we isolate the impact of the loss formulation on embedding quality and retrieval performance.

Contrastive Loss Contrastive loss operates on labelled pairs and enforces attraction between positive pairs and separation between negative pairs. For a pair (x_i, x_j) with the label $y_{ij} \in \{0, 1\}$, the loss is defined as

$$\mathcal{L}_{\text{contrastive}} = (1 - y_{ij})\frac{1}{2}d^2(\mathbf{z}_i, \mathbf{z}_j) + y_{ij}\frac{1}{2}\max(0, m - d(\mathbf{z}_i, \mathbf{z}_j))^2, \tag{4}$$

where $\mathbf{z}_i = f_\theta(x_i)$ and $d(\cdot)$ denotes the Euclidean distance.

Although simple and effective in small-scale settings, its quadratic complexity $\mathcal{O}(N^2)$ and the reliance on informative negative sampling significantly limit scalability in large-class SKU recognition.

ProxyNCA ProxyNCA replaces instance-level comparisons with learnable class representatives (proxies). Each class c is associated with a proxy $\mathbf{p}_c$, and for a sample x_i:

$$\mathcal{L}_{\text{ProxyNCA}}(x_i) = -\log \frac{\exp\left(-d(\mathbf{z}_i, \mathbf{p}_{y_i})\right)}{\sum_{c=1}^{C} \exp\left(-d(\mathbf{z}_i, \mathbf{p}_c)\right)}. \tag{5}$$

This formulation reduces complexity to $\mathcal{O}(NC)$ and improves optimisation stability. However, single-proxy representations may be insufficient when intra-class variability is strong.

Proxy Anchor Loss Proxy Anchor reformulates optimisation around proxy-centric anchors. Given proxies $\mathcal{P}$ and embeddings $\mathcal{Z}$, the loss is

$$\begin{aligned}\mathcal{L}_{\text{PA}} = &\frac{1}{|\mathcal{P}^+|}\sum_{p\in\mathcal{P}^+}\log\left(1 + \sum_{\mathbf{z}\in\mathcal{Z}_p^+} e^{-\alpha(s(\mathbf{z},p)-\delta)}\right) \\ &+ \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\log\left(1 + \sum_{\mathbf{z}\in\mathcal{Z}_p^-} e^{\alpha(s(\mathbf{z},p)+\delta)}\right),\end{aligned} \tag{6}$$

where $s(\cdot)$ denotes cosine similarity.

By weighting gradients based on similarity, Proxy Anchor improves convergence stability while maintaining linear complexity with respect to the number of classes.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

SoftTriple Loss SoftTriple extends proxy-based learning by assigning multiple proxies per class to model multi-modal intra-class distributions. For class c with proxies $\{\mathbf{w}_c^k\}_{k=1}^K$, the soft similarity is computed as

$$S'_{i,c} = \sum_{k=1}^{K} \frac{\exp\left(\frac{1}{\gamma}\mathbf{z}_i^\top \mathbf{w}_c^k\right)}{\sum_{l=1}^{K} \exp\left(\frac{1}{\gamma}\mathbf{z}_i^\top \mathbf{w}_c^l\right)} \mathbf{z}_i^\top \mathbf{w}_c^k, \tag{7}$$

and the final loss is

$$\mathcal{L}_{\text{SoftTriple}}(x_i) = -\log \frac{\exp\left(\lambda(S'_{i,y_i} - \delta)\right)}{\exp\left(\lambda(S'_{i,y_i} - \delta)\right) + \sum_{c \neq y_i} \exp\left(\lambda S'_{i,c}\right)}. \tag{8}$$

SoftTriple provides increased representational flexibility with moderate computational overhead, making it well-suited for fine-grained retail recognition.

Hierarchical Proxy Loss Hierarchical proxy loss models a multi-level semantic structure by introducing main and sub-proxies per class [18–20]:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_1 \mathcal{L}_{\text{sub}} + \lambda_2 \mathcal{L}_{\text{rel}}. \tag{9}$$

Although expressive, this formulation increases memory and computational cost, making it less attractive for large-scale industrial deployment.

Comparative Perspective Table 1 summarises the evaluated losses in terms of computational complexity, memory requirements, convergence stability, and practical suitability.

Table 1: Comparison of metric learning loss functions in terms of computational complexity, memory requirements, convergence stability, and practical suitability for large-scale SKU recognition. Here, N denotes the number of training samples, C the number of classes, and M the number of proxies per class.

Loss Function	**Computational Complexity**	**Memory Overhead**	**Convergence Stability**	**Practical Suitability**
Contrastive	$\mathcal{O}(N^2)$	Low	Low	Limited
ProxyNCA	$\mathcal{O}(NC)$	Low	Medium	Good
Proxy Anchor	$\mathcal{O}(NC)$	Low – Medium	High	Very Good
SoftTriple	$\mathcal{O}(NCM)$	Medium	High	Excellent
Hierarchical Proxy	$\mathcal{O}(NCM)$	High	Medium	Conditional

Qualitative assessments are based on empirical observations across both datasets. Convergence stability was assessed based on loss smoothness and sensitivity to hyperparameter variation.

Contrastive loss suffers from limited scalability due to its quadratic complexity. Single-proxy methods (ProxyNCA, Proxy Anchor) offer favourable efficiency with linear complexity in C, with Proxy Anchor typically demonstrating

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

more stable optimisation. Multi-proxy approaches increase overhead but provide stronger modelling capacity; among them, SoftTriple achieves the most balanced trade-off between accuracy, stability, and computational cost in our experiments.

4 Experimental Setup

4.1 Training Protocol and Experimental Configuration

The experimental design ensured a controlled and methodologically consistent comparison of loss functions within a fixed backbone architecture (ResNet-50).

All models were trained using the AdamW optimiser with decoupled weight decay. A step-based learning rate schedule was employed, reducing the learning rate at predefined epochs to stabilise convergence. Training was performed with fixed-size mini-batches, and only complete batches were considered to ensure consistent proxy updates. Mini-batches were sampled uniformly with shuffling at each epoch.

To evaluate robustness with respect to initial feature representations, two initialisation strategies were considered:

1. training from scratch using He initialisation;
2. fine-tuning from ImageNet – pretrained weights.

This setup reflects practical deployment scenarios in which pretrained backbones may or may not be available for a specific retail domain.

Hyperparameters were selected within predefined ranges to ensure stable and comparable convergence behaviour without favouring any specific method. The explored configurations included batch sizes between 100 and 300, embedding dimensions of 256, 512, and 1024, and training durations between 50 and 150 epochs. The initial learning rate ranged from 2×10^{-4} to 1×10^{-3} and was reduced using a step-based schedule.

The final configuration used for the comparative evaluation is summarised in Table 2. The relative ranking of the methods remained consistent across alternative hyperparameter configurations, indicating that the observed trends are not attributable to specific parameter choices.

All experiments were conducted with a fixed random seed; additional runs with alternative seeds confirmed consistent ranking behaviour across loss functions.

4.2 Evaluation Metrics

Performance is evaluated using Recall@K, a standard retrieval metric in deep metric learning. For each query image, the embedding is compared against the reference catalogue, and the top-K nearest neighbours are retrieved based on cosine similarity. A query is considered correctly retrieved if at least one image of the same SKU appears among the top-K results.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

Table 2: Final hyperparameter configuration used for comparative evaluation.

Parameter	**Value**
Embedding dimension	512
Batch size	128
Number of proxies per class	12
Initial learning rate	2×10^{-4}
Weight decay	1×10^{-4}
Learning rate step	30 epochs
Scaling factor α	24
Margin	0.5
Number of epochs	75

Recall@K is computed as the average proportion of correctly retrieved queries across the evaluation set. We report Recall@1, Recall@10, and Recall@100 to reflect both strict identification performance (top-1 accuracy) and more relaxed retrieval requirements relevant to retail applications, such as assisted checkout or shelf analysis, where multiple candidate matches can be inspected.

5 Results

5.1 Quantitative Results

Tables 3 and 4 summarise the retrieval performance of all evaluated methods under both random initialisation and ImageNet pretraining. The comparison highlights two complementary regimes: a sparse in-house dataset with limited samples per SKU, and the larger-scale RP2K benchmark with denser per-class sampling.

On the in-house dataset (Table 3), proxy-based methods substantially outperform the contrastive baseline when trained from scratch, with improvements exceeding 10 percentage points in Recall@1. This indicates that proxy-based optimisation provides stronger structural guidance in low-data settings. However, when pretrained weights are used, the performance gap narrows considerably, and all methods achieve comparably high recall values. This suggests that in sparse regimes, representation quality may dominate optimisation choice, reducing the relative advantage of more structured proxy-based objectives.

In contrast, the results on RP2K (Table 4) demonstrate a consistent advantage of proxy-based formulations across both initialisation strategies. All proxy variants substantially outperform the contrastive baseline at every retrieval depth. Differences among proxy-based methods are more nuanced and depend on the retrieval depth K: hierarchical and multi-proxy formulations tend to achieve the strongest Recall@1 performance, while single-proxy methods such as ProxyNCA remain highly competitive, particularly at higher K values. Importantly, the overall performance hierarchy remains stable across initialisation strategies, indicating the robustness of proxy-based optimisation to the choice of feature initialisation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

Table 3: Retrieval performance (Recall@K, %) on the in-house retail dataset under random initialisation and ImageNet pretraining. For each initialisation regime, the best result at a given K is shown in bold, and the worst is underlined.

	R@1	**R@10**	**R@100**	**R@1**	**R@10**	**R@100**
	Random initialisation			*ImageNet pretraining*		
Contrastive	<u>74.10</u>	<u>84.84</u>	<u>93.96</u>	92.78	97.43	99.27
Hierarchical	**87.67**	**94.00**	**97.99**	92.53	97.52	99.39
Proxy Anchor	87.64	93.87	97.79	92.92	97.82	**99.49**
ProxyNCA	86.25	93.06	97.40	<u>91.31</u>	<u>96.95</u>	<u>99.21</u>
SoftTriple	86.99	93.49	97.52	**93.73**	**98.00**	99.45

Table 4: Retrieval performance (Recall@K, %) on the RP2K dataset under random initialisation and ImageNet pretraining. Within each initialisation block and for each retrieval depth K, the best value is shown in bold, and the worst value is underlined.

	R@1	**R@10**	**R@100**	**R@1**	**R@10**	**R@100**
	Random initialisation			*ImageNet pretraining*		
Contrastive	<u>83.08</u>	<u>91.29</u>	<u>95.16</u>	<u>86.75</u>	<u>93.37</u>	<u>96.07</u>
Hierarchical	**94.62**	96.38	97.57	**95.39**	96.89	97.85
Proxy Anchor	92.94	96.25	97.52	93.87	96.82	97.70
ProxyNCA	92.67	96.45	**98.18**	94.37	**97.34**	**98.45**
SoftTriple	94.12	**96.62**	97.81	95.33	97.26	98.15

In general, these findings confirm that the benefits of proxy-based metric learning become more pronounced as class cardinality and dataset scale increase, supporting their suitability for large-scale retail SKU recognition.

5.2 Convergence Analysis

Figure 1 illustrates the evolution of Recall@1 during training on the in-house dataset under He initialisation. ProxyNCA initially lags behind the contrastive baseline but surpasses it within the first few epochs, achieving substantially higher final retrieval performance.

In contrast, the contrastive objective converges more slowly and plateaus earlier, indicating reduced optimisation effectiveness in sparse data regimes.

5.3 Computational Efficiency

Beyond recognition accuracy, practical suitability depends on computational cost. Although hierarchical proxy loss achieves strong retrieval performance, it incurs substantially longer training time and higher memory consumption. In contrast, SoftTriple provides competitive performance with moderate computational overhead and stable optimisation behaviour.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

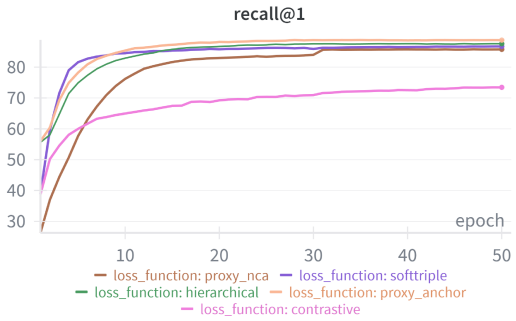
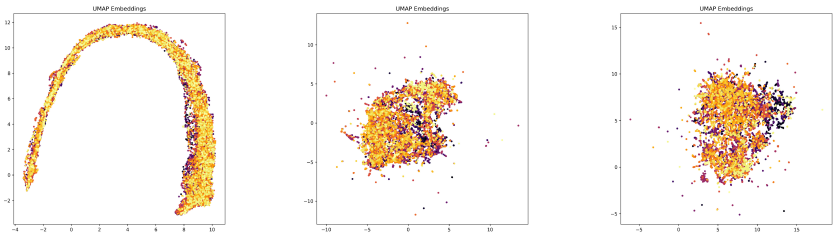


Fig. 1: Evolution of Recall@1 during training on the in-house dataset. proxy-based losses converge faster and achieve higher performance than the contrastive baseline.

5.4 Embedding Space Analysis

To qualitatively analyse the learning dynamics, we visualised the evolution of the embedding space using UMAP projections (Fig. 2). Early in training, embeddings are scattered with significant overlap between classes. As training progresses, clear clustering emerges, with progressively more compact and well-separated class regions at convergence.

These observations indicate that proxy-based optimisation rapidly establishes global structure in the embedding space, followed by refinement of intra-class compactness, consistent with the faster quantitative convergence.



(a) Epoch 1 (b) Epoch 10 (c) Epoch 75

Fig. 2: Evolution of the embedding space during training using ProxyNCA.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

Multi-proxy approaches capture intra-class variability by modelling sub-cluster structure, yielding more expressive embeddings than single-proxy representations.

5.5 Discussion

From an applied perspective, the results indicate that proxy-based metric learning is well-suited for large-scale SKU recognition systems. In particular, methods employing multiple proxies per class offer robustness to intra-class variability while remaining computationally tractable.

Nevertheless, no single loss function is universally optimal. In scenarios with limited data per SKU, the choice of backbone architecture and pretraining strategy may be as important as the choice of metric learning objective. Conversely, in large-scale settings with thousands of classes, proxy-based losses provide clear advantages in scalability and training stability. These findings indicate that the primary benefit of proxy-based learning lies not only in improved accuracy but also in predictable optimisation dynamics across heterogeneous retail conditions.

6 Limitations and Practical Considerations

Although proxy-based metric learning demonstrates strong scalability and retrieval performance, several practical considerations remain relevant for real-world deployment. Furthermore, real-world retail systems may impose additional constraints related to latency, hardware limitations, and privacy regulations, which were beyond the scope of this study. First, all evaluated methods rely on high-quality SKU annotations; label noise caused by packaging variations or manual errors may distort proxy representations, particularly in multi-proxy settings. Second, extremely long-tail distributions and limited per-class samples may reduce the stability of learned class anchors, making performance dependent on pretrained feature representations.

From a computational perspective, hierarchical and multi-proxy formulations introduce additional memory and training overhead, whereas simpler objectives such as ProxyNCA or Proxy Anchor offer better efficiency at the cost of reduced representational flexibility. At inference time, large-scale deployment requires approximate nearest-neighbour indexing to ensure real-time retrieval.

Finally, continuously evolving retail catalogues require careful proxy initialisation and controlled retraining strategies to prevent degradation of existing class representations. Despite these limitations, proxy-based metric learning provides a practical and scalable foundation for industrial SKU recognition systems.

7 Conclusions

This paper investigated the effectiveness of proxy-based metric learning for large-scale SKU recognition in retail environments. Across both dense and sparse data

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

regimes, proxy-based losses consistently outperformed contrastive objectives, offering improved convergence behaviour and better scalability with respect to class cardinality.

The relative ranking of methods remained stable across initialisation strategies, indicating that performance differences are primarily driven by the loss formulation rather than feature initialisation. This empirical stability suggests that proxy-based objectives provide predictable optimisation dynamics, which is particularly valuable in industrial deployment scenarios.

Among the evaluated approaches, multi-proxy methods, particularly SoftTriple, provided the most favourable trade-off between representational flexibility and computational efficiency.

These findings position proxy-based metric learning as a principled and practically scalable alternative to pair-based metric objectives and classification-centric training regimes in large-scale retail recognition settings. Importantly, the observed performance trends remain consistent across datasets with markedly different data regimes, reinforcing the generalisability of proxy-based optimisation principles.

Future work will explore hybrid training pipelines that integrate self-supervised pretraining with proxy-based fine-tuning, as well as adaptive proxy allocation mechanisms for dynamically evolving retail catalogues.

Acknowledgments. The research was carried out on devices co-funded by WUT within the Excellence Initiative: Research University (IDUB) programme.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. ICML'20, JMLR.org (2020)
2. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 1735–1742 (2006). https://doi.org/10.1109/CVPR.2006.100
3. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
4. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Feragen, A., Pelillo, M., Loog, M. (eds.) Similarity-Based Pattern Recognition. pp. 84–92. Springer International Publishing, Cham (2015)
5. Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3238–3247 (2020). https://doi.org/10.1109/CVPR42600.2020.00330
6. Kowalczyk, A., Sarwas, G.: One-shot learning from prototype stock keeping unit images. Information **15**(9) (2024). https://doi.org/10.3390/info15090526, https://www.mdpi.com/2078-2489/15/9/526

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33

7. Kępiński, W., Sarwas, G.: Contrastive learning in stock keeping unit image recognition. Applied Sciences **16**(6) (2026). https://doi.org/10.3390/app16062810, https://www.mdpi.com/2076-3417/16/6/2810
8. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE international conference on computer vision. pp. 360–368 (2017). https://doi.org/10.1109/ICCV.2017.47
9. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) European Conference of Computer Vision. pp. 681–699. Springer International Publishing, Cham (2020)
10. Peng, J., Xiao, C., Wei, X., Li, Y.: RP2K: A large-scale retail product dataset for fine-grained image classification (2020), https://api.semanticscholar.org/CorpusID:219980255
11. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: SoftTriple loss: Deep metric learning without triplet sampling. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6450–6458 (2019). https://doi.org/10.1109/ICCV.2019.00655
12. Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., Cohen, J.P.: Revisiting training strategies and generalization performance in deep metric learning. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 8242–8252 (2020), https://proceedings.mlr.press/v119/roth20a.html
13. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
14. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016)
15. Song, H.O., Jegelka, S., Rathod, V., Murphy, K.: Deep metric learning via facility location. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2206–2214 (2017). https://doi.org/10.1109/CVPR.2017.237
16. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4004–4012 (2016). https://doi.org/10.1109/CVPR.2016.434
17. Teh, E.W., DeVries, T., Taylor, G.W.: ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis. In: European Conference on Computer Vision. pp. 448–464 (2020). https://doi.org/10.1007/978-3-030-58586-0_27
18. Wang, J., Li, X., Song, W., Zhang, Z., Guo, W.: Multi-hierarchy proxy structure for deep metric learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1645–1649 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747268
19. Wang, M., Yang, C., Xu, Y.: Hierarchical multiple proxy loss for deep metric learning. Digital Signal Processing **133**, 103826 (2023). https://doi.org/10.1016/j.dsp.2022.103826
20. Zhao, L., Zhang, X.L.: A hierarchical multi-proxy loss with dynamic main-proxy for deep metric learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2695–2699 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447991

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_33