ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Self-Certified Deep Metric Learning with N-Tuple Losses

Oritsemisan Meggison, Sijia Zhou, Ata Kabán

University of Birmingham, School of Computer Science, B15 2TT

Abstract. Deep metric learning (DML) excels in retrieval and re-identification tasks, driven by tuple-wise loss functions that capture rich inter-sample relationships. Yet, the non-i.i.d. nature of such training complicates generalisation analysis, and the impact of tuple size remains unclear. While PAC-Bayes bounds have been applied to pairwise learning, their behaviour for higher-order tuples is unexplored. We extend this evaluation to general N-tuple settings using neural networks trained with a PAC-Bayes regularised surrogate loss. Experiments on CIFAR-10 show that sample complexity increases with tuple size, revealing trade-offs between tuple size, model capacity, and certificate tightness.

1 Introduction

Deep Metric Learning (DML) learns low-dimensional embeddings where semantic similarity aligns with a distance metric such as Euclidean or cosine distance [1]. It underpins tasks like person re-identification, face verification, and image retrieval, with its success often attributed to sophisticated tuple-wise loss functions. These operate on sample tuples (e.g., anchor, positive, and negatives), offering richer training signals than pairwise methods [2, 3].

However, tuple-wise training data are inherently *non-i.i.d.* since tuples share samples, violating assumptions of classical generalisation theory. Consequently, the interplay between tuple size and generalisation risk remains poorly understood. Recent PAC-Bayes advances [4, 5] provide computable generalisation bounds for i.i.d. data and have been extended to U-statistics [14], but have not been applied beyond pairwise learning. Building on this, we frame tuple-wise empirical risk as a U-statistic and compute risk certificates (high-confidence upper bounds on true risk) that account for tuple size and statistical dependencies to drive our learning algorithms.

We train neural networks with a generalised N-tuple loss combining a bounded surrogate loss and a PAC-Bayes regulariser, obtaining non-vacuous, high-confidence certificates. Experiments on CIFAR-10 reveal trade-offs, confirming that sample complexity grows with tuple size. Thus, the advantage of tuple-wise learning must be weighed against its higher data requirements.

2 Related Work

Significant advancements have recently been made in the field of deep metric learning, having evolved from early pairwise objectives such as contrastive loss [6], triplet loss [7], and losses that leverage richer relational information, e.g.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

the N-pair loss [2], multi-similarity loss [8], and supervised contrastive learning [9]. The latter contrasts an anchor against multiple positives and negatives simultaneously.

On the theoretical front, the PAC-Bayes framework has been successfully used to derive non-vacuous generalisation bounds for deep neural networks trained on i.i.d. data [11, 4]. While some research has explored generalisation for metric learning using algorithmic stability [12], and similar bounds has been applied to contrastive settings [13], a practical framework for computing risk certificates for metric learning models trained with N-tuple losses remains underexplored. Our work directly addresses this gap by integrating the practical machinery of self-certified learning with recent PAC-Bayes advancements for U-statistics [14], which provides a sound theoretical tool for handling the data dependencies inherent to tuple-wise objectives.

3 Self-Certified Tuple Wise Learning Framework

Consider a stochastic Convolutional Neural Network (CNN) where each weight and bias is a Gaussian distribution parameterized by a mean μ and a standard deviation $\sigma = \log(1 + \exp(\rho))$, with ρ being a learnable parameter. The inputs will be N-tuples, so the empirical risk for this model is defined by a generalised N-tuple loss, as follows. Let $S(\cdot, \cdot)$ be a similarity function (e.g. cosine similarity). For a tuple containing a anchor point x_a, one positive point x_p, and $N-2$ negative samples $\{x_{n_k}\}$, we use the loss defined by akin to a multi-class classification task [2]:

$$L_{N\text{-tuple}} = -\log \frac{\exp(\frac{1}{\tau} S(x_a, x_p))}{\exp(\frac{1}{\tau} S(x_a, x_p)) + \sum_{k=1}^{N-2} \exp(\frac{1}{\tau} S(x_a, x_{n_k}))} \tag{1}$$

where τ is a temperature parameter. The learner tries to predict which is the positive point.

Since the inputs are tuples where samples are shared across many tuples, tuple-based training violates the standard assumption of i.i.d data. To address this, we leverage the PAC-Bayes framework for U-statistics [14], which correctly models these dependencies. The effective number of independent samples is reduced from the total number of data points n to $\lfloor n/N \rfloor$, where N is the tuple size. The work [14] demonstrated the use of this framework in pairwise learning; our goal is to take advantage of it for general tuple-wise learning.

To this end, our training objective f_{obj} is a the upper bound on the true risk given in [14], which we directly minimise:

$$f_{\text{obj}} = \hat{\mathcal{R}}_{S,U}(q) + \sqrt{\frac{KL(q\|p) + \ln\left(\frac{\binom{n}{N}+1}{\delta}\right)}{2\lfloor n/N \rfloor}} \tag{2}$$

where $\hat{\mathcal{R}}_{S,U}(q)$ is the expected empirical risk of N-tuple in the posterior distribution q, p is a fixed prior, and $KL(q\|p)$ is the Kullback-Leibler divergence, which

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

penalises complexity. The denominator term $\lfloor n/N \rfloor$ correctly reflects the higher sample complexity of tuple-wise learning.

We train the model by directly minimising the objective function f_{obj} with respect to the posterior parameters (μ, ρ) using mini-batch stochastic gradient descent (SGD). The posterior parameters are initialised to match the prior, such that $(\mu, \rho) \leftarrow (\mu_0, \rho_0)$. Each training iteration involves estimating the gradient of the objective. To do this, we first sample a single network from the current posterior distribution. This is achieved efficiently via a reparameterization trick, where a random vector ϕ is drawn from a standard normal distribution $\mathcal{N}(0, I)$, and the network weights are calculated as $\mathbf{w} = \mu + \sigma \odot \phi$. This sampled network is then used to evaluate the objective on a mini-batch of tuples drawn from the training set. The resulting gradients with respect to μ and ρ are then used to update the posterior parameters via an SGD step. This process is repeated until convergence to optimise posterior distribution over the model weights.

After training, we compute a high-confidence risk certificate for the learned posterior q. As $\hat{\mathcal{R}}_{S,U}(q)$ (the expectation of empirical risk w.r.t. q) is analytically intractable, we compute an upper estimate of it as follows. First, we draw k weight samples from q and compute the average empirical risk (MC estimate), $\hat{\mathcal{R}}_{S,U}(\hat{q}_k)$. Using this, we compute a high-probability upper bound on $\hat{\mathcal{R}}_{S,U}(q)$, denoted $\overline{\mathcal{R}}_{S,U}(q)$, using binary KL-inversion to account for the variance of the MC estimate. Finally, we plug this into tight PAC-Bayes bound to yield the final certified risk, $\mathcal{R}_{\text{cert}}(q)$:

$$\mathcal{R}_{\text{cert}}(q) = f^{kl}\left(\overline{\mathcal{R}}_{S,U}(q), \frac{KL(q\|p) + \ln \frac{\binom{n}{N}+1}{\delta}}{\lfloor n/N \rfloor}\right) \tag{3}$$

where f^{kl} is the upper confidence bound on the true risk obtained by inverting the binary KL divergence.

At inference, predictions can be made deterministically using the mean of the posterior weights ($\mathbf{w} = \mu$), or stochastically by a model drawn from q, or by ensembling predictions from multiple models drawn from q. In our context, the test input is an N-tuple, in which predictor knows which is the anchor point, but does not know which is the positive point. It then computes cosine similarities between embeddings the anchor with embeddings of all other points in the tuple, and identifies highest of these. If the highest similarity point is the positive, the prediction is considered to be correct, otherwise incorrect.

4 Experiments and Results

We validate our framework on the **CIFAR-10 dataset**, adapted for a metric learning task. Code will be made publicly available upon publication.. The training set is used to construct tuples, each consisting of an anchor, a positive, and $N - 2$ negative samples. While not factored into the bounds, we observed that constructing batches by using hard mining to retrieve images with the largest disparity is beneficial, and have adopted this strategy in our experiments.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Our models are probabilistic **CNNs** of varying depths (4, 9 and 13 layers) to analyse the effect of model capacity. The models are pretrained on a 20% subset of the data, to fix the prior p. Our 'NTuple' objective, eq. (2) is compared against baselines, including a quadratic PAC-Bayes bound ('fquad')[4], which assumes i.i.d. data, and a simplified ntuple bound ('nested_tuple') that replaces the effective sample size $\lfloor n/m \rfloor$ with $|N_{\text{tuples}}|/m$, yielding an objective that is easier to optimise but no longer fully consistent with the U statistic derivation. A comprehensive ablation study of over 400 independent experiments was conducted by varying the N-tuple size ($N \in \{3, 4, 5, 6\}$), network architecture, and key model hyperparameters. Figure 1 displays a summary of obtained results, each marker representing an experiment. Regression lines fitted for each tuple size N consistently display a negative correlation between the empirically measured accuracy and the high-probability error risk certificate from our bounds. This means that the risk certificates are indeed predictive of the hold-out errors.

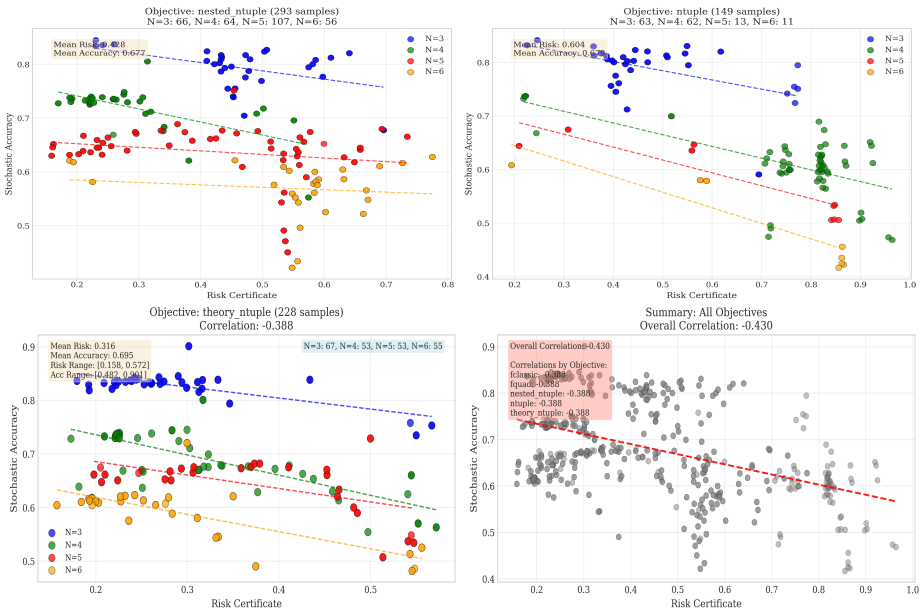

Fig. 1: Risk vs Accuracy Relationship by Tuple Size.

Moreover, all risk certificates are lower than 1. This means that our experiments successfully demonstrate the ability to train models with non-vacuous, theoretically sound generalisation guarantees. The best model, a 4-layer CNN trained with the 'NTuple' objective and a tuple size of $N = 3$, achieved a stochastic test accuracy of **82.8%** and a tight risk certificate of **0.21**, guaranteeing with 97.5% probability that its true error is no more than 21%.

We further give quantified results in Table 1, demonstrating that our theo-

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 1: A comparison of PAC-Bayes objectives across N-tuple sizes, showing model performance and certified risk.

Objective	**Metric**	N=3	N=4	N=5	N=6
`fquad`	Stoch. Acc.	0.8228	0.7094	0.6513	0.6000
	Ens. Acc.	0.8249	0.7122	0.6562	0.6074
	Stoch. Risk	0.2557	0.2583	0.2438	0.2644
	Ens. Risk	0.3291	0.3147	0.3190	0.3274
`Nested NTuple`	Stoch. Acc.	0.7939	0.7191	0.6284	0.5469
	Ens. Acc.	0.7962	0.7195	0.6288	0.5473
	Stoch. Risk	0.4172	0.2993	0.3881	0.5221
	Ens. Risk	0.4613	0.2915	0.4364	0.5693
`NTuple`	**Stoch. Acc.**	0.8277	0.6888	0.6425	0.5911
	Ens. Acc.	0.8278	0.7011	0.6447	0.5921
	Stoch. Risk	0.2108	0.2504	0.2999	0.2322
	Ens. Risk	0.2876	0.3251	0.3726	0.2880

retically grounded objective provides a reasonable balance of high accuracy and a tight, trustworthy risk certificate. While the baseline 'fquad' objective appears competitive, its certificate is overly optimistic as it fails to account for the data dependencies introduced by tuple based data.

These results confirm the initial prediction that sample complexity increases with tuple size. As shown in Figure 1, the certified risk consistently loosens (worsens) as N increases from 3 to 6, while empirical accuracy shows diminishing returns. This is a direct consequence of the effective sample size $\lfloor n/N \rfloor$ in our bound, which penalises larger tuples, showing increased performanace for certifiable metric learning models with smaller tuple sizes at the sample sizes tested. More data would be required to afford larger tuple sizes.

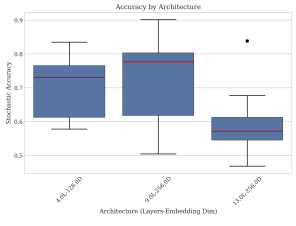

Fig. 2: Model Comparison of Predictor Types

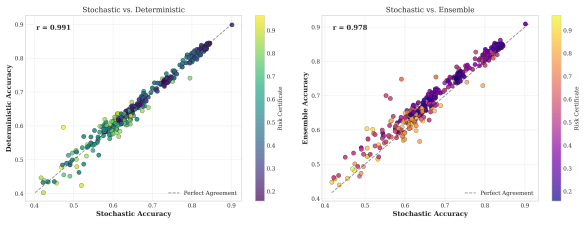

Fig. 3: Correlation of Predictor Types with Risk Certificate

Figures 2-3 compare the three different ways the PAC-Bayes learner can be used at test time, discussed at the end of Sec. 3. we see that the Ensemble predictor consistently achieved the highest accuracy by averaging embeddings from multiple weight samples, effectively reducing prediction variance. we also see from Fig. 3 that there is strong positive correlation between predictor types.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

5 Conclusion and Future Work

This study bridges the gap between the empirical success of deep metric learning and the theoretical understanding of its generalisation in the PAC-Bayes paradigm. By leveraging a PAC-Bayesian framework adapted for U-statistics, we developed and validated a method for training metric learning models that produce non-vacuous risk certificates. We find that smaller tuple sizes and simpler architectures produce more informative bounds. This research provides a novel training objective and a clear methodology for producing trustworthy similarity-learning models with high accuracy. Scaling this framework to large-scale re-identification datasets such as Market-1501, and integrating the bounds with larger visual networks, are subject to further research. We will also seek to better understand the observed beneficial effect of tuple mining strategies, such as batch-hard mining [10].

References

[1] K. Musgrave, S. Belongie, and S.-N. Lim, A Metric Learning Reality Check, *arXiv preprint arXiv:2003.08505*, 2020.

[2] K. Sohn, Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[3] B. Yu and D. Tao, Deep Metric Learning With Tuplet Margin Loss. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[4] M. Pérez-Ortiz, O. Rivasplata, E. Parrado-Hern'andez, B. Guedj, and J. Shawe-Taylor, Progress in Self-Certified Neural Networks, *arXiv preprint arXiv:2111.07737*, 2021.

[5] O. Rivasplata, et al., PAC-Bayes and Domain Adaptation. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[6] R. Hadsell, S. Chopra, and Y. LeCun, Dimensionality Reduction by Learning an Invariant Mapping. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1735-1742, 2006.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 815-823, 2015.

[8] X. Wang, et al., Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] P. Khosla, et al., Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[10] A. Hermans, L. Beyer, and B. Leibe, In Defense of the Triplet Loss for Person Re-Identification, *arXiv preprint arXiv:1703.07737*, 2017.

[11] G. K. Dziugaite and D. M. Roy, Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks via PAC-Bayes, *arXiv preprint arXiv:1703.11008*, 2017.

[12] Y. Jiang, B. Kulis, and O. Dikmen, Generalization Bounds for Deep Metric Learning, *arXiv preprint*, 2020.

[13] X. Wei, S. Kakade, and T. Ma, PAC-Bayesian Generalization Bounds for Contrastive Learning, *Preprint*, 2021.

[14] S. Zhou, Y. Lei, and A. Kabán, Self-certified Tuple-wise Deep Learning. In *Proc. of Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, Springer, 2024.