Silhouette Distance Loss for Learning Few-Shot Contrastive Representations

Kosmas Pinitas[*] KOSMAS.PINITAS@UM.EDU.MT
Institute of Digital Games, University of Malta, Malta

Nemanja Rasajski[†] NEMANJA.RASAJSKI@UM.EDU.MT
Institute of Digital Games, University of Malta, Malta

Konstantinos Makantasis KONSTANTINOS.MAKANTASIS@UM.EDU.MT
University of Malta, Malta

Georgios N. Yannakakis GEORGIOS.YANNAKAKIS@UM.EDU.MT
Institute of Digital Games, University of Malta, Malta

Editors: Manuel Grana, Pawel Ksieniewicz, Leandro Minku, Pawel Zyblewski

Abstract

Conventional supervised contrastive learning methods excel in optimising encoders for discriminative tasks. In scenarios where only a few labelled samples are available, however, they struggle in eliminating the inductive bias when transferring from source to target classes. This is a byproduct (and inherent limitation) of their underlying optimisation process that involves training a representation to maximise class separation, without directly optimising for within-class cohesion. As a response to this limitation this paper introduces the *Silhouette Distance* (SD) loss, a new optimisation objective for supervised contrastive representation learning. SD aims to enhance the quality of learned embeddings by emphasising both the cohesion and separation of representation clusters for each class. We test SD extensively across several few-shot learning scenarios—where labelled data is limited—and we compare its performance against supervised contrastive loss and prototypical network loss for various text and image classification tasks. We also test SD in a cross-domain manner, by training a model on one dataset and testing it on another, within the same modality. Our results demonstrate the superior, at worst competitive, performance of the SD loss compared to its baselines. By leveraging pre-trained models and fine-tuning techniques, our study highlights how the SD loss can effectively improve representation learning across different modalities and domains. This initial study showcases the potential of the SD loss as a robust alternative within the few-shot learning setting.

Keywords: Limited label access

1. Introduction

Traditional supervised learning relies on large labelled datasets for training, typically with test data from similar statistical distributions. Obtaining labelled data, however, is far from being scalable, especially for new or rare concepts. *Few-Shot Learning* (FSL) (Wang et al., 2020) addresses this by allowing pre-trained models to generalise to new data categories with minimal labelled samples per class. By extending pre-trained models, such as those trained in a self-supervised manner, to new categories, the need for re-training from

[*] Corresponding author.

[†] Pinitas and Rasajski have equal contribution in this paper

scratch is eliminated, conserving computational resources. The core challenge of FSL lies in overcoming the inductive bias from source classes to better adapt to target classes. This challenge can be tackled by further refining the learned embedding space so that instances from distinct classes can be effectively differentiated into separate clusters (Tian et al., 2020).

To this end, we adopt a popular representation learning method, Supervised Contrastive Learning (SCL) (Khosla et al., 2020). Based on class labelling information, SCL creates pairs of similar and dissimilar samples. Those pairs are then used to cluster together, in the embedding space, data points that belong to the same class, and, at the same time, drive apart clusters of samples from different classes. However, the conventional supervised contrastive loss (L_{SC})—used as the training signal for SCL—primarily focuses on creating dissimilar class representations (Bukchin et al., 2021). While the resulting clusters are somewhat separated, there is significant potential for further improvement as we can observe in Fig. 1. One avenue for improvement is directly promoting both high intra-class similarity and minimal dispersion of representations within a class.

We rely on earlier work (Vapnik, 2013) suggesting that better class separation leads to better generalisation and allows for training with less labelled data. Thus, in this paper, we assume that forcing representations to maintain tight and separable clusters of data will be beneficial for the performance of an SCL process as a whole. With this aim we explore the efficacy of *Silhouette score* as an optimisation objective of representation learning for few-shot learning tasks. Inspired by recent work in this area (Minnehan and Savakis, 2018b,a), we introduce and test a novel loss for contrastive learning, namely *Silhouette Distance* (SD) loss (L_{SD}). Unlike the conventional supervised contrastive loss that prioritises learning discriminative representations per class, our loss focuses on creating representation clusters, one per class, by considering both clustering *cohesion* and *separation*. We thus test the hypothesis that through the optimisation of the SD loss, the trained encoders can learn compact context representations from a limited number of labelled examples. Figure 1 depicts the high level concept of our approach via an illustrative example on a synthetic dataset.

In this study, we focus on testing the proposed SD loss in an FSL setting. We thus use only a small number of labelled samples and compare the proposed loss (SD) against both the vanilla SC loss and the Prototypical Network (PN) loss (Snell et al., 2017)—an extensively studied method shown to be effective in FSL—on both image and text classification tasks across four datasets; two for image classification, and two for intent classification in text. Specifically, we extract representations of images and text using frozen pre-trained backbone models and we fine-tune a projection layer on top of those representations with the aforementioned training methods.

The paper presents several notable contributions. First, we introduce a novel approach by utilising a loss function derived from a clustering quality metric as the optimisation objective for supervised contrastive representation learning, marking a pioneering endeavour in this domain. Second, we perform an initial investigation into the properties of SD using synthetic data. Finally, we extend our investigation to encompass two distinct modalities, namely images and text. Employing various pre-trained backbones, we conduct experiments across 6 different scenarios, in terms of the number of samples and classes considered. Additionally, we conduct cross-domain experiments by training models on one dataset and

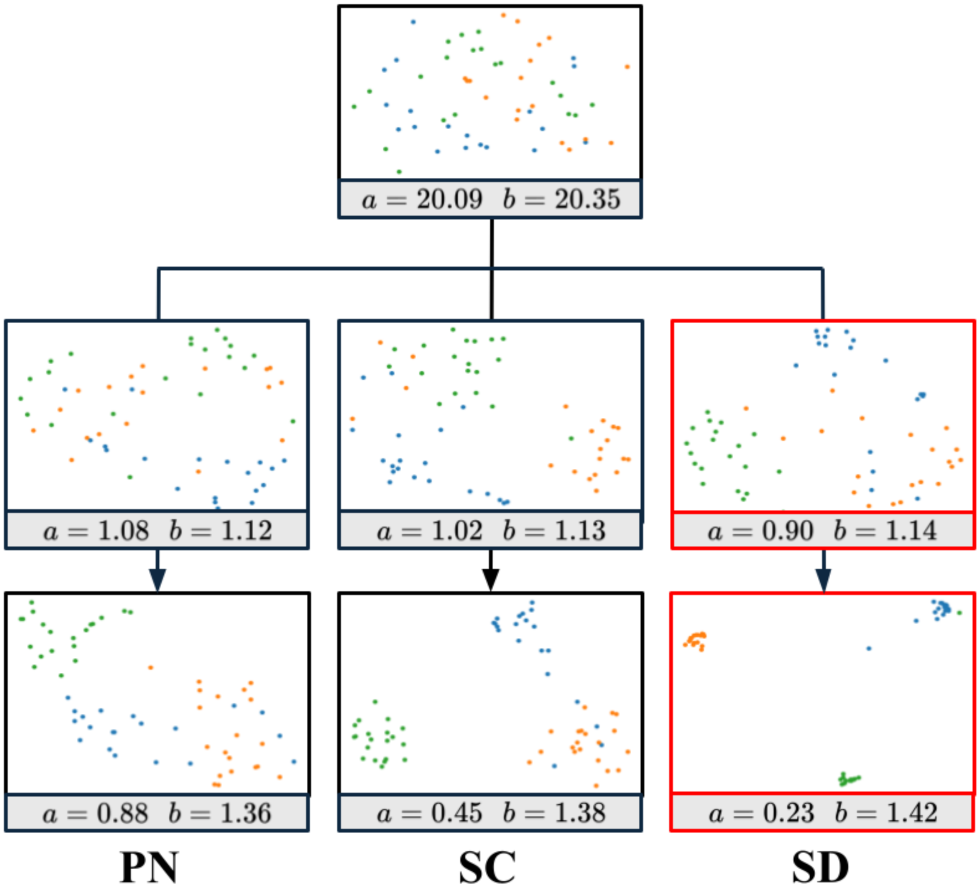

Figure 1: **Learning to Contrast Representations via the Silhouette Distance Loss—Core Concept**: t-SNE plots of the embedding spaces generated via optimising three different loss functions: the PN (Prototypical Network), SC (Supervised Contrastive), and the introduced SD (Silhouette Distance) loss. Their shared starting point (top figure) is a synthetic dataset comprising and 20 input dimensions and three classes (coloured differently). The middle figures display a snapshot of the optimisation process of a single-layer perceptron at epoch 50, while convergence is observed after 100 epochs (bottom figures). Parameters a and b correspond to intra-class distance (lower is better) and inter-class distance (higher is better) from the nearest class, respectively. Our proposed SD loss, highlighted in red, demonstrates superior efficacy in separating observations from different classes into distinct clusters (see Section 3.4).

evaluating them on another dataset, within the same modality. The results consistently demonstrate that our introduced SD loss method achieves performance levels comparable to, and often surpassing, those of SC and Prototypical Networks. Importantly, the SD loss shows significant performance advantages in harder classification tasks involving more classes and less samples.

2. Related Work

2.1. Few-Shot Classification

Few-shot classification refers to the learning paradigm that aims to develop computational models capable of categorizing unseen samples from a few annotated examples (Wang et al., 2020). Most few-shot learning algorithms can be grouped into three distinct categories based on their learning strategy: *optimization-based* algorithms, *metric learning* approaches, and *hybrid* methods.

Optimisation-based algorithms enable the rapid adaptation of models to new tasks with minimal updates. Indicatively, Finn et al. (Finn et al., 2017) introduced MAML, a model-agnostic optimisation approach focused on identifying a set of initial parameters capable of improving the performance of the model on novel tasks using limited data and minimal updates. This is achieved by iteratively fine-tuning or re-initialising the parameters of the model across each task. Nichol et al. (Nichol and Schulman, 2018) developed Reptile, a modified version of MAML. Their method drops the task-specific re-initialisation proposed of MAML resulting in reduced computational load. Finally, Li et al. (Li et al., 2017) introduced Meta-SGD that aims to find both optimal initial parameters and learning rate for each task.

Metric-learning approaches, instead, aim to project the input into a common latent space and learn a metric to discern between samples from new classes. The main benefit of such approaches is their simplicity since they do not require fine-tuning for specific tasks. Notably, Snell et al. (Snell et al., 2017) developed the *Prototypical Networks* that learn a prototype for each class in a known metric space, where samples from the same class are closer to their prototype than to prototypes of other classes. Sung et al. (Sung et al., 2018) introduced the Relation Networks that attempt to capture the relationship between data points by learning a deep distance metric that compares a small number of samples per iteration. Vinyals et al. (Vinyals et al., 2016) utilised an attention mechanism to produce a weighted nearest neighbour classifier given the support set of an episode. Finally, hybrid methods combine the benefits of both optimisation-based and metric-learning approaches. An indicative study under this category is Proto-MAML (Triantafillou et al., 2019), a method that learns a prototype for each class in a meta-learning framework, allowing for rapid adaptation to new tasks with limited data.

Unlike the methodologies mentioned above, this work focuses on learning representations in a few-shot manner and introduces a new contrastive learning loss inspired by the silhouette cluster validity index that optimises for both inter-class separation and intra-class cohesion. Hence, we aim to finetune self-supervised trained models to construct few-shot embedding models for both image and text classification tasks.

2.2. Contrastive Learning

Contrastive Learning (CL) methods have gained significant attention within the study of representation learning across several domains. Indicatively, Nakamasa and Keita (Inoue and Goto, 2020) introduced a semi-supervised contrastive learning framework for text-independent speaker verification. Qin et al. (Qin et al., 2024) proposed a novel distribution-aware contrastive learning algorithm to address the inconsistencies in medical image segmentation. Pinitas et al. (Pinitas et al., 2022) employed supervised contrastive learning for predicting human affect by relying on multiple modalities of user input such as facial expression and physiology.

Although research at the intersection of few-shot learning (FSL) and contrastive learning (CL) has been active in recent years, the literature remains relatively sparse. Liu et al. (Liu et al., 2021) used CL with noise contrastive estimation to develop a few-shot embedding model for image classification. Chen et al. (Chen et al., 2022) proposed ContrastNet, a contrastive learning framework aimed at addressing overfitting in few-shot text classification. Zhen et al. (Zheng et al., 2022) introduced mixed-supervised hierarchical contrastive learning to differentiate videos at various levels and employed weak supervision to align discriminative temporal clips or spatial patches. Lastly, Jian et al. (Jian et al., 2022) combined supervised CL with standard masked language modelling loss in prompt-based few-shot learners across 15 diverse language tasks. In this paper, we introduce a new contrastive learning loss based on a clustering validity metric. We showcase the robustness of the proposed method across dissimilar datasets and few-shot learning settings.

2.3. Learning via Silhouette Score

While the Silhouette score is a widely used clustering validity index, research on Silhouette-based optimisation objectives is limited. Minehan et al. (Minnehan and Savakis, 2018b) introduced a simplified Silhouette score (Sil) as a regularisation term on cross-entropy loss for improved image classification. The authors also (Minnehan and Savakis, 2018a) developed a framework combining Sil loss with a retraction onto the Grassmann manifold, enhancing performance in smaller network architectures. Vardakas et al. (Vardakas et al., 2024) introduced a deep learning clustering method incorporating a probabilistic Silhouette score to improve clustering accuracy.

Unlike the works mentioned above, our proposed SD loss is differentiable in the embedding space. It also considers the actual Silhouette score and calculates distances across all samples. We show how the typical silhouette score is converted into a smooth loss suitable for few-shot representation learning without the need for additional optimisation objectives.

3. Method

3.1. Problem Setting

Few-shot classification aims to teach models to adapt to new, unseen categories with minimal labeled examples. We define D^{train}, D^{val}, and D^{test} as the non-overlapping training, validation, and test sets, respectively. In each iteration, termed episode or task, data is drawn from D^{train}, D^{val}, or D^{test} for training, validation, or test. Each episode involves N classes (referred to as N-way) and K samples per class (referred to as K-shot). Within an episode, a support set (S) with labeled samples for training and a query set (Q) with labeled samples for evaluation are defined. The model is then trained to classify samples in the query set based on insights gained from the support set. Formally, we denote the i-th sample in the support and query sets as (x_i^s, y_i^s) and (x_i^q, y_i^q), respectively. It's worth noting that both N (classes) and K (samples/shots) are hyperparameters affecting the few-

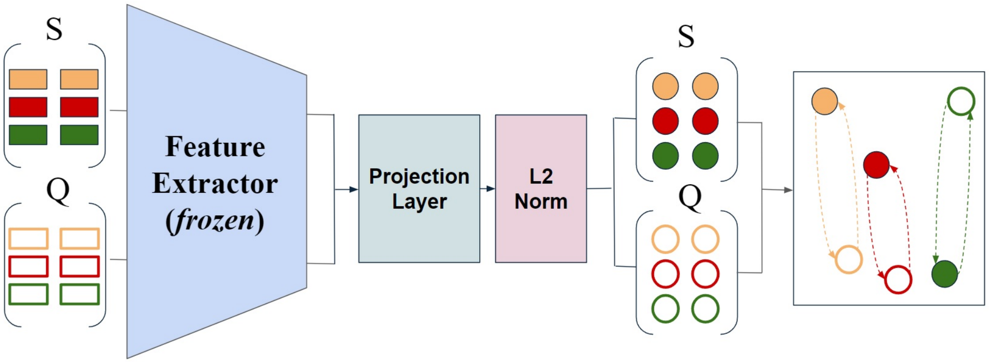

Figure 2: The overall setting of our evaluation process. S and Q represent the support and query sets, respectively. We first extract embeddings using a pre-trained frozen feature extractor. Following this step we pass the extracted embeddings through a trainable projection layer, and perform $L2$ normalisation. Finally we optimise on the tested loss functions using the resulting S and Q sets.

shot setting's difficulty, increasing the number of comparisons and reducing support set information per episode, respectively.

3.2. Representation Components

The overall methodology employed in this paper is illustrated in Figure 2. In the context of representation learning an encoder model refers to a neural network architecture capable of projecting high-dimensional data into a lower-dimensional latent space. Therefore, post-training, it serves as an efficient coding function, that generates low-dimensional high-level representations of the input space. In this work, we test the predictive power of the learnt representations within the domain of few-shot learning across two modalities (image and text). We employ 4 different backbones (2 for each modality) which are then finetuned to tackle few-shot learning problems. The first encoder, employed in both the text and image classification tasks, is the base version of CLIP (Radford et al., 2021), with a patch size of 16 and an embedding size of 512 in both scenarios. When it comes to the image classification tasks the second tested encoder is the small version of DINOv2 (Oquab et al., 2023), with a patch size of 14 and an embedding size of 384. In the case of text encoders, we also test the base uncased version of BERT (Devlin et al., 2018) with an embedding size of 768. During the finetuning phase, a ReLU-activated layer is added on top of the frozen backbones which is optimised by the learning objectives described in the next Section (see Section 3.3). The trainable layer is of the same dimension as the backbones' output while $L2$-norm is applied to project the embeddings on the unit sphere.

3.3. Learning Objectives

The primary contribution of this paper lies in introducing a supervised contrastive learning (SCL) loss derived from the silhouette score for constructing representations capable of learning few-shot classification tasks. In particular, we first employ self-supervised trained models to extract representations for each sample. The extracted embeddings are then fine-tuned via the Silhouette Distance (SD) Loss within a few-shot classification setting. The proposed loss is compared against two additional optimisation objectives namely Prototypical Network (PN) Loss and Supervised Contrastive (SC) Loss.

3.3.1. Prototypical Network Loss

The Prototypical Network loss, which serves as the first baseline, was first introduced by Snell et al. (Snell et al., 2017). Using that loss, the authors aim to learn a metric space in which classification can be performed by computing the distances between the query samples and the prototypes derived for each of the N classes within the support set. The Prototypical Network loss is defined as

$$L_{PN} = -\frac{1}{|Q|} \sum_{(x_i^q, y_i^q) \in Q} \sum_{n=1}^{N_q} (y_i^q = n) \log(p_\theta(y_i^q = n | x_i^q)), \tag{1}$$

where $p_\theta(y_i^q = n|x_i^q) = softmax(-d(f_\theta(x_i^q), c_n^s)$ is the probability of a query sample x_i^q to fall into the class n and $c_n^s = \frac{1}{|S_n|} \sum_{(x_i^s, y_i^s) \in S} (y_i^s = n) f_\theta(x_i^s)$ is the prototype of class n. It should be noted that $d(\cdot)$ corresponds to the Euclidean distance, $f_\theta(\cdot)$ represents the learnable embedding functions and $|S_n|$ is the number of samples of class n in the support set. Finally, $|Q|$ is the cardinality of the query set.

3.3.2. Supervised Contrastive Loss

The optimisation objective of SC (Khosla et al., 2020) is to learn representations that make positive pairs (i.e. samples with the same label) more similar and negative pairs (i.e. samples with different labels) more dissimilar by, respectively, minimising the distance between representations of positive samples and maximising the distance between negative samples. The minimisation of this loss yields distinct and separable representations for each class. Inspired by previous work in few-shot representation learning for classification (Liu et al., 2021)(Section 2.2) we formulate SC as follows:

$$L_{SC} = \frac{1}{|S|} \sum_{s \in S} \frac{-1}{|P_s^q|} \sum_{p \in P_s^q} \log \frac{\exp(r_s \cdot r_p / \tau)}{\sum_{q \in Q} \exp(r_s \cdot r_q / \tau)}, \tag{2}$$

where S is a set that includes all samples in the support set and P_s^q is the set that includes only the query set samples that are assigned to the same class as s while $q \in Q$ denotes any element in the query set. With r_s, r_p and r_q we denote the latent representations produced from a function f_θ that samples x_s, x_p and x_q, respectively. Finally, τ stands for a non-negative temperature hyperparameter that transforms the representation similarity distribution. Finally, s, p and q correspond to the index of the current support set sample, a query set sample positive to the current support sample and a sample in the query set, respectively. As it can be discerned from Eq. 2, SC primarily focuses on creating dissimilar class representations.

3.3.3. Silhouette Distance Loss

In a supervised context, the Silhouette score is adapted to encourage representations that not only consider the separability of classes but also the cohesion within each class. The *Silhouette Distance* (SD) loss as introduced here evaluates the quality of clustering patterns based on labels by measuring the average Euclidean distance within the same class (intra-class cohesion) and the average nearest-class Euclidean distance (inter-class separation). The term "Silhouette Distance" was chosen because the loss function draws inspiration from the silhouette score, providing an intuitive understanding of its operation. While it does not meet the criteria to be classified as a metric distance, it can be considered a non-metric distance, similar to cosine distance. The minimisation of this loss yields representations that consider both the separability of classes and the compactness within each class. Formally, the SD loss is defined as:

$$L_{SD} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1 - Sil(q, S)}{2}, \tag{3}$$

where

$$Sil(q, S) = \frac{b(q, S) - a(q, S)}{max_{\delta}\{a(q, S), b(q, S)\}} \tag{4}$$

denotes the silhouette score. Given a query sample (x_q, y_q), $a(q, S)$ corresponds to the average distance between the representation of x_q and the representations of all the support set samples x_s that have label $y_s = y_q$. Furthermore, $b(q, S)$ is the average Euclidean distance between the representation of x_q and the nearest support set class. $max_{\delta}\{a(\cdot), b(\cdot)\} = max\{a(\cdot), \delta\}$ when $b(\cdot) \leq a(\cdot)$ and $max_{\delta}\{a(\cdot), b(\cdot)\} = b(\cdot)$ when $b(\cdot) > a(\cdot)$ ensuring that $S(\cdot)$ is differentiable for all samples. δ is a small scalar value guaranteeing numerical stability when $a(\cdot) = b(\cdot) = 0$ and is empirically set to 0.001. It is important to note that we apply L2 normalisation to all embeddings (see Fig 2). This normalisation ensures that the resulting Euclidean distances are proportional to cosine distances, making them approximately equivalent. (Qian et al., 2004)

3.4. Analysing the behaviour of L_{SD}

This section aims to provide a deeper insight into the behaviour of the Silhouette Distance Loss during optimisation. Hence we primarily focus on the two main components that constitute the SD loss, namely inter-class and intra-class distance. In particular, we generate two synthetic classification scenarios that vary in terms of class separation. The first scenario focuses on a setting where there is a clear separation between classes while the second scenario features slight to no separation across classes. Moreover, we train a simple ReLu layer with $L2$-normalised activations on both scenarios by focusing on either one or both components of the SD loss. It should be noted that the generation process creates clusters of points normally distributed ($\sigma = 1$) on vertices of a 20-dimensional hypercube. The class separation parameter h is a multiplicative factor that controls the size of the hypercube. In this experiment, h is set to 0.9 and 2 for the *No Separation* and the *Clear Separation* scenario, respectively.

Figure 3 showcases the influence of each component on the properties of the latent space. Specifically in the *clear separation* scenario where the data generation process yields well-separated classes, it is evident, both visually and numerically (silhouette scores), that the separation is retained across both components of the SD loss. On the other hand, in the *no separation* scenario, the role of each component is clearer. The intra-class distance yields more tight yet less separated clusters and the inter-class distance results in better separated yet less cohesive representation clusters per class. The limitations of these individual components are also more apparent. While there is some enhancement of clustering compared to the initial space, as highlighted by silhouette scores, it is not as remarkable as observed in the prior situation. Overall, the SD loss optimisation leads to higher quality clusters than the initial state, as indicated by silhouette scores in both scenarios. However, the cluster shape and separability are closely tied to the input space's geometrical properties.

An additional analysis involves investigating the differences between the representations learned by optimising across all tested losses on synthetic data. Once again, we follow the same data generation process, however, in this case we set the class separation parameter to $h = 1.1$. It is evident from Figure 1—presented earlier in the introduction of the paper—that all losses facilitate learning discriminating representations in a supervised setting. We argue, however, that L_{PN} focuses on finding a boundary between the query sample and its corresponding prototype, thus acting as a cross entropy minimisation objective. Furthermore, L_{SC} places a primary emphasis on creating separable clusters for each class whereas L_{SD} considers the overall clustering quality resulting in cohesive and separable representation clusters. This is also illustrated in Figure 1 where SD marks the lowest intra-class and the highest inter-class distance.

4. Experiments

4.1. Datasets

We conduct experiments on four datasets in total: two for text classification (Banking77 and Clinic150) and two for image classification (FC100 and mini-Imagenet). Specifically, **Banking77** (Casanueva et al., 2020) is a fine-grained dataset specific to single-domain banking for intent classification. It consists of 13083 customer service queries labelled with 77 classes, in which some categories are similar and may overlap with others. **Clinic150** (Larson et al., 2019) contains 150 classes and $23,700$ examples across 10 domains. It features $22,500$ user utterances evenly distributed in every class and $1,200$ out-of-scope queries. We discard out-of-scope examples and only use $22,500$ data points. Shifting to image classification datasets, **FC100** (Oreshkin et al., 2018)—also known as Fewshot-CIFAR100—consists of $60,000$ samples labelled with 100 classes (600 samples per class). A common split consists of 60 categories for training, 20 categories for validation and 20 categories for testing. The RGB images of FC100 are resized to 224x224 pixels for the experiments presented here. Finally, **mini-Imagenet** (Vinyals et al., 2016) is a subset of Imagenet consisting of 100 categories; each category contains 600 samples. These categories are split into train, validation and test sets with 64, 16 and 20 classes, respectively, following the same partition as (Liu et al., 2021). Once again the RGB images are resized to 224x224 pixels. The detailed statistics of all four datasets employed in this paper are shown in Table 1.

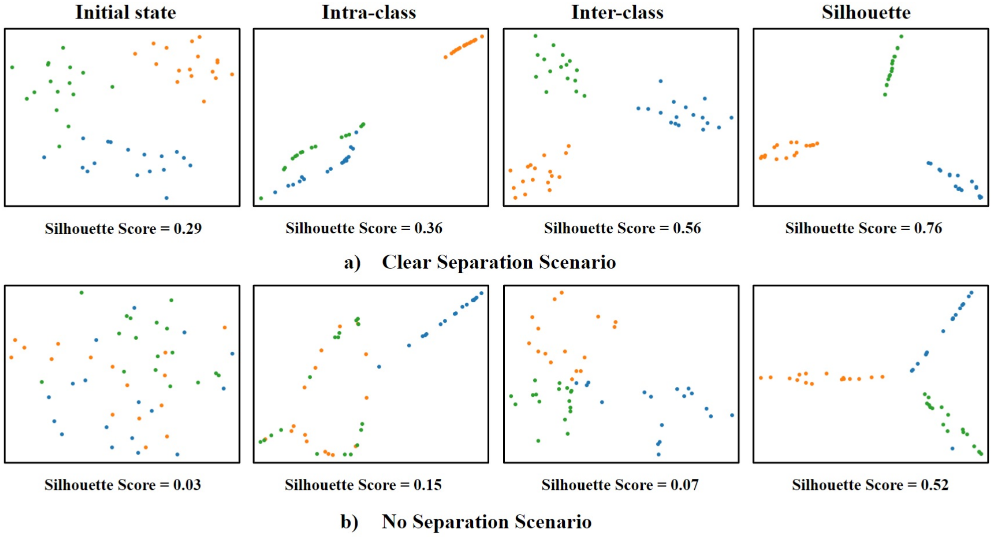

Figure 3: Analysis of the SD loss components, visualised with t-SNE plots, along with their respective silhouette scores in two scenarios. *Initial state* corresponds to the dataset input space. *Intra-class* and *inter-class*, respectively, refer to the projection of data optimised for intra-class distance minimisation and inter-class distance maximisation. *Silhouette* refers to a latent space that minimises the silhouette distance. It can be observed, both visually and through silhouette scores, that using individual components as loss functions slightly improves the clustering of different classes within the embedding space. However, when combined into the SD loss they exhibit notable enhancements (rightmost t-SNE plots).

Table 1: High level statistics of the four datasets used in this paper. For the two text datasets (top two rows), the size column refers to the average sentence length. For the two image datasets (bottom two rows), instead, it refers to the number of colour channels, width and height of images.

dataset	#samples	size	#train / valid / test
FC100 (Oreshkin et al., 2018)	60, 000	3x224x224	60 / 20 / 20
mini-Imagenet (Vinyals et al., 2016)	60, 000	3x224x224	64 / 16 / 20
Banking77 (Casanueva et al., 2020)	13, 083	12	25 / 25 / 27
Clinic150 (Larson et al., 2019)	22, 500	9	50 / 50 / 50

4.2. Experiment Protocol

In this initial study, we compare models optimised by the losses mentioned in Section 3.3, using the N-way K-shot FSL evaluation setting. For each episode, we randomly sample N classes and K samples per class for both training and validation. Following best practices

Table 2: **Image classification**: The 5-way, 10-way and 20-way 1-shot and 5-shot image classification mean accuracy on the FC100 and mini-Imagenet datasets obtained by Prototypical Network Loss (PN), Supervised Contrastive Loss (SC) and Silhouette Distance Loss (SD). The bold values indicate the highest absolute accuracy while the underlined values demote which methods perform on par with SD based on 95% CI.

		FC100						mini-Imagenet					
Method	Backbone	5-way		10-way		20-way		5-way		10-way		20-way	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PN	CLIP	42.56	65.76	38.26	55.53	35.80	48.99	80.34	93.67	73.07	**91.13**	71.48	88.86
SC		48.34	**69.83**	40.19	58.83	37.23	51.67	84.46	94.43	**77.82**	90.19	71.45	89.11
SD (ours)		**53.14**	62.14	**43.79**	**59.89**	**39.05**	**53.83**	**84.52**	**94.54**	69.97	91.07	**76.83**	**90.5**
PN	DINOV2	54.80	74.70	43.77	**65.94**	41.07	60.11	86.22	94.73	79.17	92.86	74.31	91.71
SC		**54.84**	**75.12**	**44.45**	64.19	40.73	59.56	**88.16**	**96.10**	**81.63**	**94.59**	**79.05**	92.78
SD (ours)		44.26	74.6	42.98	65.38	**42.74**	**60.68**	84.58	96.02	81.59	94.44	78.90	**93.21**

Table 3: **Text classification**: The 5-way, 10-way and 20-way 1-shot and 5-shot text classification mean accuracy on the Banking77 and Clinic150 datasets.

		Banking77						Clinics150					
Method	Backbone	5-way		10-way		20-way		5-way		10-way		20-way	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PN	BERT	60.21	82.63	50.91	75.46	47.20	70.52	78.76	95.15	70.46	92.35	64.05	89.50
SC		**66.94**	87.40	61.62	79.80	55.59	70.68	85.22	**96.55**	77.35	92.55	69.37	90.89
SD (ours)		60.94	**87.52**	**64.71**	**83.45**	**61.22**	**79.24**	**87.74**	96.51	**83.34**	**94.72**	**77.47**	**92.92**
PN	CLIP	72.94	89.77	65.41	84.41	61.72	79.95	78.14	96.48	71.76	94.27	73.33	92.96
SC		78.41	**91.78**	71.25	86.67	66.46	80.26	78.20	96.46	80.06	95.37	72.74	92.91
SD (ours)		**82.32**	91.46	**74.43**	**87.46**	**69.62**	**84.32**	**90.08**	**97.47**	**82.23**	**95.74**	**77.82**	**93.8**

from earlier studies (Chen et al., 2022; Liu et al., 2021), we report results on 5-way using both 1-shot and 5-shot settings, and we extend this evaluation protocol for the more challenging 10-way and 20-way settings.

The models are trained using a protocol that stops training after 10 epochs (or 60-300 episodes, depending on the setting) if there is no improvement in validation accuracy, returning the best-performing model.SGD is used for optimisation, with a learning rate scheduler that halves the learning rate every 5 epochs. The optimal learning rate varies by model and setting, with values between $\alpha \in (10^{-3}, 10^{-1})$ benefiting SD, while in other configurations, the optimal range is $\alpha \in (10^{-6}, 10^{-3})$.

All hyperparameters are selected via a greedy search on the validation set. The performance of the models is evaluated in terms of accuracy score following the evaluation protocol of prototypical networks to promote a fair comparison across the presented methodologies. We repeat the model training experiments 5 times and we sample $1,000$ test episodes per run. All reported significance tests are measured at 95% confidence level (CI) with $p < 0.05$.

4.3. Few-Shot Learning

The proposed methodology is tested extensively across four datasets and 6 few-shot settings. Table 2 showcases the average accuracy of the models for the two image datasets and backbone architectures employed. SD performs on par with the top-performing models in 3 out of 12 settings while achieving the highest accuracy in the majority of settings (i.e. 7 out of 12 settings; significantly outperforming both baselines in 5 out of 7 settings) in the **FC100** dataset. Furthermore, SC marks significantly higher accuracy than SD only in 2 settings while PN performs on par with SD only in 1 setting. Observing the performances obtained in the **mini-Imagenet** dataset, once again, SD achieves the same levels of accuracy with the best models in 5 out of 12 settings while achieving the highest accuracy in those settings (significantly outperforming both baselines in 3 out of 5 settings). SC performs on par with SD in half (6 out of 12) of the settings while it significantly outperforms the latter in 2 settings. PN performs on par with SD only in a few cases (2 out of 12 settings) while it does not manage to outperform it in any setting.

Table 3 illustrates the average accuracy of the models obtained in text classification tasks. When it comes to the **Banking77** dataset, SD performs on par with the best-performing models only in 1 case while it marks the highest accuracy in the vast majority of settings (i.e., 10 out of 12). SC performs on par with SD twice while it outperforms it significantly only once. PN once again performs on par with SD in 2 settings and does not manage to outperform it. Similarly, in the **Clinics150** dataset, SD achieves the highest accuracy in the vast majority of settings (i.e., 11 out of 12) performing significantly higher than the rest of the models in this dataset. SC performs on par with SD only in 1 setting while PN does not outperform or perform on par with SD in any setting.

Importantly, SD loss tends to outperform other optimisation objectives as the number of classes (N) increases. In tasks with a higher number of classes, such as 20-way classification, distinguishing between many categories becomes more difficult. Unlike PN, that aims to find decision boundaries where the query samples closer to their prototypes, and SC, which focuses on creating discriminative representations per class, SD addresses both within-class cohesion and between-class separation. This dual focus helps the model better define boundaries between many classes, leading to higher classification accuracy.

4.4. Testing Across Datasets

To further assess the quality of the representations learned we also perform experiments across datasets using the models trained on dataset A to predict samples from dataset B ($A \rightarrow B$). In particular, we focus on the 5-way and 20-way settings since they are the most common and most difficult scenarios, respectively. Table 4 showcases the average accuracy of the models across image datasets and backbone architectures. When we evaluate the capacity of the **FC100** trained models to test **mini-Imagenet** dataset (see **FC100→mini-Imagenet**) we observe that SD performs on par with the best-performing models in 1 out of 8 settings marking the highest accuracy in half (4 out of 8) of the settings. SC performs on par with SD in 3 settings (significantly higher in 2) while PN performs on par with SD in 3 settings. When it comes to **mini-Imagenet→FC100** SD performs on par with the best models in 2 settings and achieves the highest accuracy in 3 settings. SC performs on par with SD in 2 settings while it significantly outperforms the latter in 3. PN performs on par with SD in half of the settings (4 out of 8).

In a similar vein, Table 5 showcases the average accuracy of the models across text datasets and backbone architectures. Particularly, in the case of **Banking77→Clinics150** SD marks the highest accuracy in most settings (5 out of 8) yielding significantly higher

Table 4: **Image Classification Across Datasets**: The 5-way and 20-way (1-shot and 5-shot) image classification mean accuracy on experiments across datasets obtained by Prototypical Network Loss (PN), Supervised Contrastive Loss (SC) and Silhouette Distance Loss (SD). The right arrow (→) points at the dataset used for testing. The bold values indicate the highest absolute accuracy while the underlined values demote which methods perform on par with SD based on 95% CI.

Method	Backbone	FC100→mini-Imagenet				mini-Imagenet→FC100			
		5-way		20-way		5-way		20-way	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PN	CLIP	77.66	93.85	71.09	88.60	46.96	66.64	34.24	49.24
SC		75.44	**94.18**	71.35	88.99	**49.90**	**67.07**	32.45	48.06
SD (ours)		**82.67**	89.70	**71.90**	**89.65**	46.38	63.09	**35.86**	**51.44**
PN	DINOV2	82.10	94.75	71.60	91.82	54.12	74.43	41.34	59.59
SC		**82.44**	**94.90**	72.64	**92.26**	**57.90**	**74.51**	**41.99**	59.35
SD (ours)		74.12	94.86	**72.89**	91.51	52.14	74.46	41.76	**60.47**

Table 5: **Text Classification Across Datasets**: The 5-way and 20-way (1-shot and 5-shot) text classification mean accuracy.

Method	Backbone	Banking77→Clinics150				Clinics150→Banking77			
		5-way		20-way		5-way		20-way	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PN	BERT	78.78	95.57	62.88	89.84	59.76	80.01	46.1	67.86
SC		**82.4**	**96.21**	64.0	90.48	60.34	80.68	46.88	69.32
SD (ours)		80.32	96.16	**66.53**	**91.63**	**61.92**	**81.77**	**50.88**	**71.73**
PN	CLIP	78.62	97.1	69.61	93.14	69.86	87.57	59.02	77.38
SC		**86.54**	97.44	72.55	93.69	**72.64**	88.46	59.45	78.5
SD (ours)		83.92	**97.64**	**73.97**	**93.83**	72.36	**88.6**	**60.52**	**79.28**

performances than the other models in 3 settings. SC performs on par with SD in 3 settings while significantly outperforming it only once while PN performs on par with SD in 2 settings. When it comes to **Clinics150→Banking77** once again SD marks- statistically the same accuracy with the best model in 1 setting and marks the highest accuracy in the vast majority of settings. SC and PN perform on par with SD in 6 and 4 settings, respectively, but they do not outperform it significantly.

Once again, the capacity of the proposed SD method is more evident in the more challenging 20-way scenario. SD performs on par or even outperforms the other two losses in 15 out of 16 settings with SC performing on par or better than the former just in 7 out of 16 cases. PN marks the worst performance in the 20-way scenarios since it performs on par with SD in 5 settings and it does not manage to outperform any of the methods in any of the 20-way settings.

5. Discussion

This work introduced a novel contrastive loss based on silhouette score and investigated the potential of this loss in handling few-shot classification tasks. In particular, we compare SD with SC and PN losses across different modalities and datasets. The core findings indicate that the SD encoders yield more robust few-shot learners. A worthwhile discussion is our choice of not applying any data augmentation in this paper. We particularly decided to omit this step since our backbone models are pretrained in a self-supervised manner and thus are robust across an array of augmentations. However, it is worth exploring whether data augmentations based on feature manipulation such as the addition of noise and feature gating can be beneficial for the models.

Another aspect worth addressing involves the selection of silhouette score over other clustering validity indexes such as Calinki-Harabasz (Caliński and Harabasz, 1974) and Davies-Buldin (Davies and Bouldin, 1979) for representation learning. Silhouette is interpretable which, in turn, makes it advantageous for evaluating the training process (loss value inspection). It can also be easily converted to a smooth loss function suitable for gradient-based optimization. Lastly, according to previous research (Arbelaitz et al., 2013) silhouette can be regarded as the most robust and resilient clustering validity index.

Although, we can argue with relative confidence that our approach achieves SOTA performance in mini-Imagenet 5-way settings and FC100 5-way 5-shot among the metric and contrastive learning approaches and performs on par with the SOTA in the 5 way-5 shot setting of Clinics 150 (Hu et al., 2023; Liu et al., 2021), we particularly refrain from making such claims. This decision stems from the challenges we encountered in reproducing the results of other methods. Factors such as computational constraints and code availability posed significant obstacles in our attempts to validate or compare against existing benchmarks such as CAML and BAVARDAGE (Fifty et al., 2023; Hu et al., 2023). Additionally, most of the representation learning papers that focus on few-shot learning tasks usually treat the representation learning objective as an auxiliary loss and not as the main optimisation objective (Chen et al., 2022; Liu et al., 2021). Consequently, our focus remains on the insights gained from our approach, and comparison with methodologies that optimise for similar objectives.

We foresee several important directions for future research based on the results obtained in this paper. First, it is essential to test the efficacy of the proposed loss in predicting other modalities such as sound and graph-based structures. Another crucial area involves exploring the potential of the proposed loss as an auxiliary objective in other few-shot learning algorithms to improve adaptability and performance across various domains. We also plan to assess our method's effectiveness in producing representations for tasks like classification, segmentation, and preference learning. Additionally, a comprehensive analysis is needed to evaluate the behavior of the proposed loss with data imbalances and noise. Despite these open research directions, we highlight that the introduced loss is versatile and applicable to any representation learning domain where the downstream task can be framed as a classification problem.

6. Conclusion

This paper introduced the *Silhouette Distance Loss*, a loss function that encourages representations to form tight clusters for samples that belong to the same class (e.g., same object) and large separations between representation clusters from different classes. By leveraging pre-trained backbone models, we were able to learn high-level representations that accurately classify unseen classes across 6 different few-shot image and text classification settings and 4 different datasets. Comparing the performance of the proposed loss against the supervised contrastive loss—the standard optimisation objective for supervised contrastive learning—we observe that the former yields models that are superior, especially when the number of classes increases, by constructing tight and separable representation clusters. These results hold even in cases where the train and test data come from different datasets. While our initial findings already suggest that the method is generic and applicable to any few-shot classification task, its promise and full potential remains to be tested across more diverse datasets and downstream tasks.

Acknowledgments

Makantasis was supported by Project ERICA (GA: REP-2023-36) financed by the Malta Council for Science & Technology (MCST), for and on behalf of the Foundation for Science and Technology, through the FUSION: R&I Research Excellence Programme. Pinitas, Rasajski, and Yannakakis were supported by Project OPtiMaL funded by MCST through the SINO-MALTA Fund 2022.

References

Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern recognition*, 46(1):243–256, 2013.

Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Fine-grained angular contrastive learning with coarse labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8730–8740, 2021.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10492–10500, 2022.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Christopher Fifty, Dennis Duan, Ronald G Junkins, Ehsan Amid, Jure Leskovec, Christopher Ré, and Sebastian Thrun. Context-aware meta-learning. *arXiv preprint arXiv:2310.10971*, 2023.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Yuqing Hu, Stéphane Pateux, and Vincent Gripon. Adaptive dimension reduction and variational inference for transductive few-shot classification. In *International Conference on Artificial Intelligence and Statistics*, pages 5899–5917. PMLR, 2023.

Nakamasa Inoue and Keita Goto. Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1641–1646, 2020.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Contrastive learning for prompt-based few-shot language learners. *arXiv preprint arXiv:2205.01308*, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8635–8643, 2021.

Breton Minnehan and Andreas Savakis. Defrag: Deep euclidean feature representations through adaptation on the grassmann manifold. *arXiv preprint arXiv:1806.07688*, 2018a.

Breton Minnehan and Andreas Savakis. Learning robust feature representations in deep networks for image classification. In *2018 25th International Conference" Mixed Design of Integrated Circuits and System"(MIXDES)*, pages 29–33. IEEE, 2018b.

Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Supervised contrastive learning for affect modelling. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 531–539, 2022.

Gang Qian, Shamik Sural, Yuelong Gu, and Sakti Pramanik. Similarity between euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 1232–1237, 2004.

Zheyun Qin, Xiaoming Xi, and Yilong Yin. Distribution-aware contrastive learning for robust medical image segmentation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1991–1995, 2024. doi: 10.1109/ICASSP48485.2024.10446000.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

Vladimir Vapnik. *The nature of statistical learning theory.* Springer science & business media, 2013.

Georgios Vardakas, Ioannis Papakostas, and Aristidis Likas. Deep clustering using the soft silhouette score: Towards compact and well-separated clusters. *arXiv preprint arXiv:2402.00608*, 2024.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *European Conference on Computer Vision*, pages 297–313. Springer, 2022.