All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction

Imanol G. Estepa
Universitat de Barcelona,
Barcelona, Spain
igonzaes42@alumnes.ub.edu

Ignacio Sarasúa
NVIDIA
isarasua@nvidia.com

Bhalaji Nagarajan
Universitat de Barcelona,
Barcelona, Spain
bhalaji.nagarajan@ub.edu

Petia Radeva
Universitat de Barcelona,
Barcelona, Spain
Computer Vision Center,
Cerdanyola (Barcelona), Spain
petia.ivanova@ub.edu

Abstract

Nearest neighbour-based methods have proved to be one of the most successful self-supervised learning (SSL) approaches due to their high generalization capabilities. However, their computational efficiency decreases when more than one neighbour is used. In this paper, we propose a novel contrastive SSL approach, which we call All4One, that reduces the distance between neighbour representations using "centroids" created through a self-attention mechanism. We use a Centroid Contrasting objective along with single Neighbour Contrasting and Feature Contrasting objectives. Centroids help in learning contextual information from multiple neighbours whereas the neighbour contrast enables learning representations directly from the neighbours and the feature contrast allows learning representations unique to the features. This combination enables All4One to outperform popular instance discrimination approaches by more than 1% on linear classification evaluation for popular benchmark datasets and obtains state-of-the-art (SoTA) results. Finally, we show that All4One is robust towards embedding dimensionalities and augmentations, surpassing NNCLR and Barlow Twins by more than 5% on low dimensionality and weak augmentation settings. Source code is available in **https://github.com/ImaGonEs/all4one**.

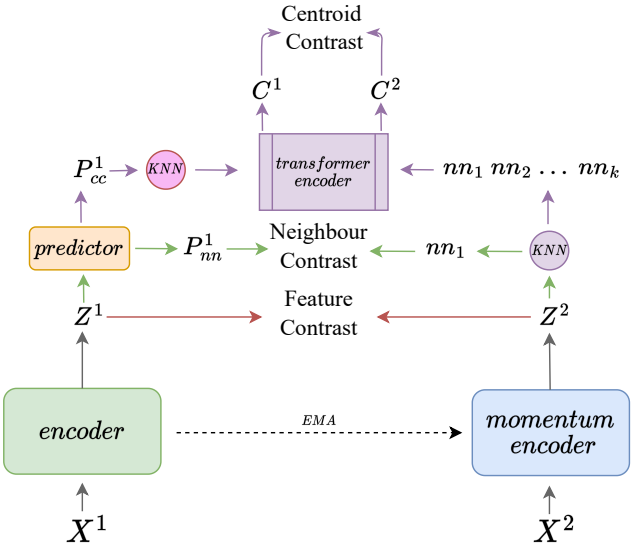

Figure 1: **Simplified architecture of All4One.** All4One uses three different objective functions that contrast different representations: Centroid objective contrasts the contextual information extracted from multiple neighbours while the Neighbour objective assures diversity [14]. Additionally, the Feature contrast objective measures the correlation of the generated features and increases their independence.

1. Introduction

Deep learning (DL) models strongly depend on the availability of large and high-quality training datasets whose construction is very expensive [25]. Self-supervised learning (SSL) claims to allow the training of DL models without the need for large annotated data, which serves as a milestone in speeding up the DL progression [25]. The

most popular SSL approaches rely on instance discrimination learning, a strategy that trains the model to be invariant to the distortions applied to a single image defined as positive samples [6, 18, 31]. As all the views belong to the same image, consequently, they belong to the same semantic class. Bringing them together in the same feature space encourages the model to create similar representations for similar images. The encouraging results of initial works such as SimCLR [6] and BYOL [15] boosted multiple improvements that address common problems of instance discrimination such as lack of diversity between samples and model collapse.

Neighbour contrastive learning hinges on the fact that data augmentations do not provide enough diversity in selecting the positive samples, as all of them are extracted from the same initial image [14]. To solve it, Nearest neighbour Contrastive Learning (NNCLR) [14] proposes the use of nearest neighbours (NN) to increase the diversity among the positive samples which in turn boosts the generalization of the model. Instead of bringing together two distortions created from the same image, they increase the proximity between a distorted sample and the NN of another distorted sample. However, relying entirely on the first neighbour holds back the real potential of the approach. MSF [21] proposes the use of k neighbours to increase the generalization capability of the model. However, MSF suffers from high computation as the objective function needs to be computed for each neighbour (k times). Apart from the low diversity of positive samples, instance discrimination approaches suffer from model collapse, a scenario where the model learns a constant trivial solution [15]. Barlow Twins [38] proposes a redundancy reduction-based approach that naturally avoids the collapse by measuring the correlation among the features on the generated image representations. However, this collapse avoidance suffers from the requirement of projecting embeddings in high dimensions.

In our work, we contrast information from multiple neighbours in a more efficient way by avoiding multiple computations of the objective function. This way, we are able to increase the generalization from neighbour contrastive approaches while avoiding their flaws. For that, we propose the use of a new embedding constructed by a self-attention mechanism, such as a transformer encoder, that combines the extracted neighbour representations in a single representation containing contextual information about all of them. Hence, we are able to contrast all the neighbours' information on a single objective computation. We make use of a Support Set that actively stores the representations computed during the training [14] so that we can extract the required neighbours. In addition, we integrate our approach with a redundancy reduction approach [38]. Making the computed cross-correlation matrix close to the identity reduces the features redundancy of the same image representation while also making them invariant to their distortions. This idea contrasts the representations in a completely different way than the rest of instance discrimination approaches [28, 30, 14]. For this reason, we increase the richness of the representations learnt by the model by combining the neighbour contrast approach with the redundancy reduction objective that directly contrasts the features generated by the encoder and aims to increase their independence. In addition, the need for high-dimensional embeddings of redundancy reduction feature contrast approaches [38] is alleviated thanks to our SSL objective combination.

As a summary, in this paper, we introduce a new symbiotic SSL approach, which we call All4One, that leverages the idea of neighbour contrastive learning while combining it with a feature contrast approach (see Figure 1). All4One integrates three different objectives that prove to benefit each other and provide better representation learning. Our contributions are as follows: (i) We define a novel objective function, centroid contrast, that is based on a projection of sample neighbours in a new latent space through self-attention mechanisms. (ii) Our proposal, All4One, is based on a combination of centroid contrast, neighbour contrast and feature contrast objectives, going beyond the single neighbour contrast while avoiding multiple computations of the objective function; (iii) We demonstrate how contrasting different representations (neighbours and distorted samples) using InfoNCE [28] based and feature contrast objectives benefit the overall performance and alleviate individual flaws such as the reliance of high-dimensional embeddings on feature contrast approaches; (iv) We show that All4One, by contrasting the contextual information of the neighbours and multiple representations, outperforms single nearest neighbour SSL on low-augmentation settings and low-data regimes, proving much less reliance on augmentations and increased generalization capability; and (v) We prove that All4One outperforms single-neighbour contrastive approaches (among others) by more than 1% in different public datasets and using different backbones.

2. Related Works

Self-supervised Learning. In SSL methods, a model is trained to learn intermediate representations of the data in a completely unsupervised way to be transferred later to multiple tasks [25]. Since the introduction of contrastive loss [10], several works such as SimCLR [6] and MoCo [18] proved their usefulness in multiple downstream tasks by proposing variations of InfoNCE [28], a contrastive objective function inspired by Noise Contrastive Estimation (NCE) [26]. In image representation learning, this objective function works with the assumption of positive pairs (z^a, z^+), formed by two representations that share the same semantic class, and negative pairs (z^a, z^-), which are formed by image representations that do not belong to the

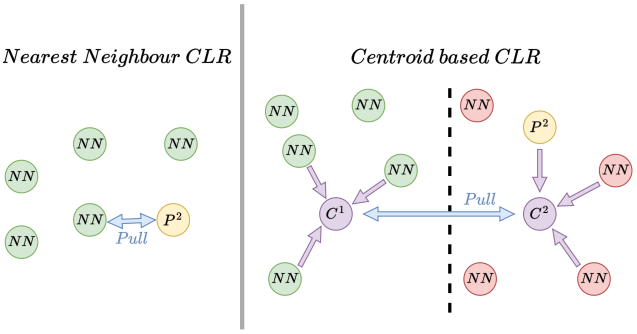

Figure 2: **Neighbour contrast comparison.** While the common neighbour contrastive approaches only contrast the first neighbour, we create representations that contain contextual information from the k NNs and contrast it in a single objective computation.

same semantic class. Thus, the positive pairs are pulled together in the same feature space while the negative pairs are repelled to avoid model collapse. Later, BYOL [15] proved that it is possible to achieve the same effect without using negative samples by avoiding the collapse with the introduction of architectural changes such as a predictor. Additionally, Barlow Twins [38] introduced a novel objective function based on the redundancy reduction principle instead of InfoNCE that naturally avoided the collapse.

Overall, discriminative frameworks have been obtaining exceptional results [5, 7, 8, 6] due to the improvements done by introducing new architectures [18, 5, 9], applying new objective functions individually (e. g. redundancy reduction) [27, 4, 2], alternative augmentation settings [32] or even proposing novel strategies such as NN based approaches [14, 21], providing an increased generalization by contrasting distortions with NNs.

During training, NN-based approaches store the representations in a queue and extract them by applying a k-NN algorithm that uses one of the representations in the positive pair as a query. This way, one of the pair representations is swapped by its neighbour for the loss computation. Nevertheless, using a single neighbour per sample holds back the potential of the approach meanwhile using k neighbours reduces the efficiency of the NN approaches as the objective function needs to be computed multiple times.

Self-attention. Since the introduction of the Transformer [33] architecture, self-attention-based models have proved to be one of the most successful approaches in Computer Vision (CV). In fact, multiple works analyse the behaviour of the self-attention mechanism and combine it with other well-known tools such as k-NN algorithms [35, 16, 22, 34]. In SSL, Self-attention has been widely used on generative frameworks [17, 3, 40], where they train the transformer backbone to reconstruct the given masked image. However, these reconstruction objectives used in works such as iBOT [40], BEiT [3] and MAE [17] are computationally expensive and rely on vision transformers exclusively. In our research, different to previous MIM works, we still maintain a contrastive objective. We take advantage of the capacity of the transformer encoder to mix the neighbour representations into a single one that contains information from the neighbours and contrast it using a variation of InfoNCE [28].

3. The All4One Symbiosis

In order to increase the performance and efficiency of previous neighbour contrastive methods, we propose a symbiotic SSL framework that combines three different approaches into one. We show our proposed All4One pipeline in Figure 3, where we show the three different objectives: the first objective is a neighbour contrast objective (green path); the second objective is a centroid contrast objective (purple path), which is carried out by the application of self-attention mechanisms [33] and the final objective is a redundancy reduction-based feature contrast objective (red path). During training, initial representations are transformed depending on the followed path to adapt them to the objective of the path.

The pipeline is composed of a pair of encoders or neural networks, f, and a pair of projectors, g. The projector consists of a basic MLP that transforms the output of the encoders [6]. In this case, we apply a momentum encoder, which is a smoothed version of the online encoder, similar to BYOL [15]. The pipeline is iterated by a batch of images, X. For each image, two batches of augmented/distorted images, X^1 and X^2 are generated using a data augmentation pipeline. Then, both distorted images are fed to the encoder and, next, to the projector. This sequence of encoder-projector is defined as momentum or online branch depending on the encoder used. We call the momentum branch the encoder-projector sequence that contains a momentum projector and vice-versa [15]. The output of momentum and online branches can be defined as $Z^1 = g^\xi(f^\xi(X^1))$ and $Z^2 = g^\theta(f^\theta(X^2))$ respectively, being g and f the projectors and encoders of each branch. Next, we brief each of the objectives used in the proposal.

3.1. Neighbour Contrast

NNCLR [14] is the most popular neighbour contrastive approach. Instead of contrasting two distortions of the same image, it uses the simple KNN operator to extract from a queue or Support Set [14] the NN of the first distortion and contrast it against the second distortion using a variant of InfoNCE [28]. For each i-th pair in the batch, the neighbour contrast loss is defined as:

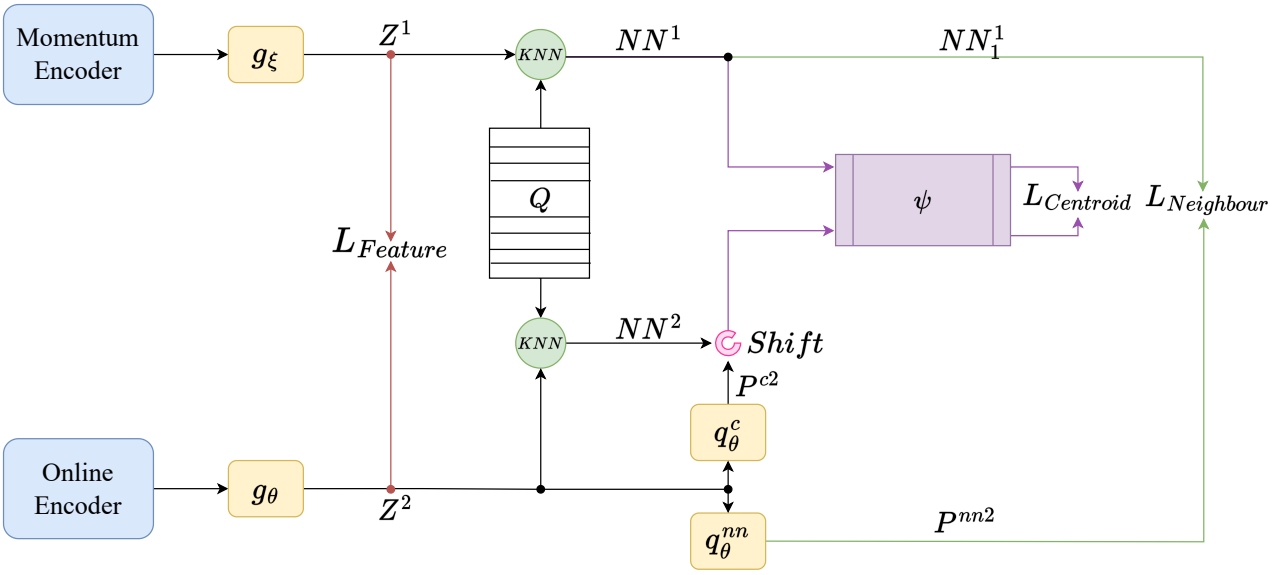

Figure 3: **Complete architecture of All4One framework.** Feature, Centroid and Neighbour contrast objective functions are indicated by red, purple, and green respectively.

$$L_i^{NNCLR} = -log(\frac{exp(nn_i^1 \cdot p_i^2/\tau)}{\sum_{k=1}^{N} exp(nn_i^1 \cdot p_k^2/\tau)}) \quad (1)$$

where nn_i is the i-th NN of the representation z, p_i is the second distortion, τ refers to the temperature constant and N is the number of samples in the batch. This way, it increases the generalization of the model through the use of more diverse samples. Note that, in our case, the second predictor q_θ^{nn} is used for the Neighbour Contrast objective.

3.2. Centroid Contrast

According to MSF [21], using a single NN could be holding back the potential of the approach. In fact, they showed that contrasting the second distorted image with multiple NN could provide a better SSL framework that also obtains higher accuracy regarding the selection of the neighbours. Nevertheless, contrasting multiple neighbours hurts the computational efficiency of the model, as they need to compute the objective function k times, where k is the number of extracted neighbours. For this reason, the improvement is severely constrained by computational resources. Following the idea of using multiple neighbours, we introduce an alternative proposal that does not require multiple-loss computations. We compile the relevant information from the extracted k neighbours to create a pair of representations, defined as "centroids" that contain contextual information about all the neighbours and pull them together in the feature space applying a variation of the InfoNCE [28] objective function. This way, the generalization of the model is improved without contrasting multiple neighbours one by one.

Once Z^1 and Z^2 are computed from the pair of distorted images, we calculate the cosine similarity between each Z (query representation) and the Support set Q, a queue that stores the computed representations [14]. Next, we extract a sequence with the K most similar representations ($nn_i^1 = KNN(z_i^1, Q)$) for each representation in both batches. Then, for each Z^1 and Z^2, we obtain their respective batch of sequences of NNs, NN^1 and NN^2. As we try to avoid contrasting the neighbours one by one, we introduce a new element to the pipeline: a transformer encoder, ψ. Given NN^1 and NN^2, we input ψ with each sequence nn_i^1 to compute the self-attention of the sequences.

Given a sequence of neighbour representations nn_i^1, we obtain a single representation c_1 that contains as much information as possible about the input sequence nn_i^1. When computing self-attention [33], we mix the representations of the input sequence in a weighted manner so that a new enriched vector of representations is returned. Each element of this enriched vector contains contextual information about all the neighbours in the sequence. During training, for each sequence in NN^1, the process is made up of the following steps: (i) for each sequence Seq_i in NN^1, we add sinusoidal positional encoding [33]; (ii) then, we feed the transformer encoder ψ with Seq_i; (iii) inside the transformer encoder, self-attention is computed and a new sequence is returned Seq_i^c; (iv) finally, we select the first representation Seq_{i1}^c in the returned sequence Seq_i^c as our centroid c_i as we aim to contrast a single representation

that contains context information from the rest of the neighbours. After selecting the first representation on all sequences, we obtain a batch of representations defined as C^1.

On the online branch, a slightly different process is followed. A second MLP similar to the projector, the predictor, is used to change the feature space of Z^2 batch and obtain $P^{c2} = q_\theta^c(Z^2)$ batch of transformed representations. More concretely, we pass Z^2 through the centroid predictor q_θ^c. Then, we replace the fifth neighbour in each sequence nn_i^2 of NN^2 by p_i^{c2}. Finally, we reorder each sequence so p_i^{c2} is the first element. We define this process as $Shift$ operation. This is done to introduce the distorted image in the sequence, thereby impacting the back-propagation. More information about the *Shift* operation is provided in Section 1 of the supplementary material. Once modified NN^2 is created, we pass it through the transformer encoder to obtain C^2 by following the previously explained process. Finally, we contrast C^1 and C^2 using a variation of the InfoNCE [28] loss function aiming to bring the neighbour centroids together (see Figure 2). For each centroid pair, the centroid loss can be defined as:

$$L_i^{centroid} = -log\left(\frac{exp(c_i^1 \cdot c_i^2/\tau)}{\sum_{n=1}^{n} exp(c_i^1 \cdot c_n^2/\tau)}\right) \quad (2)$$

3.3. Reducing Redundancy: Feature Contrast

The application of the redundancy reduction principle to increase the independence of the features is one of the most successful approaches in the SSL SoTA [38]. The Barlow Twins' main idea is that, instead of focusing on the images, to directly contrast the features by computing a cross-correlation matrix C_{ij}. Then, it aims to increase the invariance of the features by equating the diagonal elements c_{ii} of C_{ij} to one (invariance term) while also decreasing the redundancy between the features by reducing the correlation between different features (i^{th} and j^{th} features). Inspired by this approach, we increase the richness of our framework by introducing a feature contrast objective function that measures the correlation of the features and aims to increase their independence.

To do so, we construct the output of the momentum and online projectors as $(Z)_{ij}$ matrices where each element represents an exact feature j of a single augmented image representation, i. Then, we use L_2 normalization on both matrices in the batch dimension [38] and we compute the cosine similarity between the transposed matrix, Z^{1T} and Z^2 to obtain the cross-correlation matrix, CC^1. We compute this term symmetrically. This is done by swapping the branches and computing the similarities. Finally, two cross-correlation matrices CC^1 and CC^2 are obtained. Redundancy reduction feature contrast objective [2] is computed as follows:

$$L_{Red.} = \frac{1}{2}\sqrt{\frac{1}{2D}\sum_{i=1}^{D}((1-cc_{ii}^1)^2 + (1-cc_{ii}^2)^2)} + \frac{1}{2}\sqrt{\frac{1}{2D(D-1)}\sum_{i=1}^{D}\sum_{j\neq i}^{D}((cc_{ij}^1)^2 + (cc_{ij}^2)^2)} \quad (3)$$

As it can be noted, the first term increases the correlation between the elements that represent the same feature among the distorted images (diagonal elements), while the second term decreases the correlation between elements that represent different features (off-diagonal elements).

3.4. Final Objective: The All4One Objective

Once all objectives are computed, the final loss function is formed by summing the previously defined objectives. The All4One objective is defined as:

$$L_{All4One} = \sigma L_{NNCLR} + \kappa L_{Centroid} + \eta L_{Red} \quad (4)$$

where σ, κ and η are determined through loss progression as 0.5, 0.5 and 5, respectively. Depending on the dataset, giving more importance to each objective could increase the performance of the models. However, the common benchmarks used for validation are balanced, so we keep the three All4One objectives uniformly weighted. By combining different objectives, the All4One objective improves the learning of representations. We show the improvements over other methods below.

4. Experiments

In this section, we first describe the implementation details of All4One and its training. Then, we evaluate it using the common image classification linear evaluation pipeline on different datasets: CIFAR-10 [23], CIFAR-100 [23], ImageNet-1K [24] and ImageNet-100, a reduced ImageNet of 100 classes. We extensively study the different components of our proposal and discuss the various design decisions in detail. Finally, we also extend our proposal to a transformer-based backbone using ViT [13] and also show the efficacy of our proposal on other downstream tasks.

4.1. Implementation Details

Architecture. All4One follows a momentum instance discrimination pipeline (Figure 3). The momentum branch includes the usual momentum projector, g_ξ [15]. The online branch, on the contrary, includes a projector, g_θ and double predictor, q_θ^{nn} and q_θ^c. MLP projectors are formed by 3 fully connected layers of size [2048, 2048, 256], while the MLP predictor uses 2 fully connected layers with a dimensionality of [4096, 256]. Similar to NNCLR [14], all

fully-connected layers, except the last ones, are followed by batch-normalization [19].

MLP components of a SSL method filter the features generated by the encoder, keeping those that are useful for the downstream task [1]. In All4One, there are multiple objectives that are applied in completely different representations, so the predictors involved should filter the features independently. For this reason, we propose the use of two different predictors that work separately for each objective. Finally, our approach introduces a small Transformer encoder, ψ, that is applied for both branches and a shared $KNN(\cdot,\cdot)$ operator that extracts image representations from the Support Set, Q, given a query. The final objective of our proposal follows Eq. (4), which brings together both contrasting approaches.

Training. We train All4One on CIFAR-10, CIFAR-100, ImageNet100 (using a ResNet-18 backbone) and the complete ILSVRC2012 ImageNet (using a ResNet-50 backbone) without any class label or annotation. During the training, all backbones are initialized with default Solo-learn [11] initialization. All MLP components use the PyTorch default initialization. The transformer encoder uses three transformer encoder layers with 8 heads each [33]. Following SimCLR [6], lr is adapted using $lr * batchsize/256$ formula and readapted for each layer using LARS [37] (only for CNNs). For ResNet-18 and ResNet-50, we use the SGD optimizer and on the ViT-Small backbone, we use AdamW. We find it optimal to use 1.0, 0.00015 and 0.1 as base learning rates for the CNN backbones, ViT and the internal transformer encoder respectively, which are gradually reduced by using a warm-up cosine annealing scheduler. When using AdamW, we detach the lr of the transformer encoder by setting it to a constant of 0.1. For the complete ILSVRC2012 ImageNet experiments, we adapt the learning rates to 1.5 and 0.45 for All4One and NNCLR respectively. All experiments are run with the same batch size and for the same amount of epochs. Finally, the queue or Support set size is set to 98304, following the settings of NNCLR [14]. The rest of the hyperparameters of the model are directly extracted from NNCLR [14]. All the training processes are done on a single NVIDIA RTX 3090 GPU, except for ImageNet experiments where the evaluations are done on 4xNVIDIA V100 GPUs. Training curve comparison is shown in Section 4 of the supplementary material.

4.2. All4One Evaluation

CIFAR and ImageNet100 Linear Evaluations. We compare our approach against the current SoTA SSL frameworks. We first show the evaluations on the CIFAR datasets and the ImageNet-100 dataset. ImageNet-100 is a reduced ImageNet version with 100 classes and the images are 224x224 (as compared to CIFAR which has 32x32).

	CIFAR-10		CIFAR100		ImageNet100	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BYOL	92.58	99.79	70.46	91.96	80.16	95.02
DC V2	88.85	99.58	63.61	88.09	75.36	93.22
DINO	89.52	99.71	66.76	90.34	74.84	92.92
MoCoV2+	92.94	99.79	69.89	91.65	78.20	95.50
MoCoV3	93.10	99.80	68.83	90.57	80.36	95.18
ReSSL	90.63	99.62	65.92	89.73	76.92	94.20
SimCLR	90.74	99.75	65.78	89.04	77.64	94.06
Simsiam	90.51	99.72	66.04	89.62	74.54	93.16
SwAV	89.17	99.68	64.88	88.78	74.04	92.70
VIbCReg	91.18	99.74	67.37	90.07	79.86	94.98
VICReg	92.07	99.74	68.54	90.83	79.22	95.06
W-MSE	88.67	99.68	61.33	87.26	67.60	90.94
BT	92.10	99.73	70.90	91.91	80.38	95.28
NNCLR	91.88	99.78	69.62	91.52	79.80	95.28
All4One	**93.24**	**99.88**	**72.17**	**93.35**	**81.93**	**96.23**

Table 1: **Linear evaluation results on CIFAR-10, CIFAR-100 and ImageNet100.** The results are extracted from Solo-learn Self-supervised learning library [11].

We report Top-1 and Top-5 linear accuracies for all the datasets. As can be seen from Table 1, our approach clearly outperforms the previous SoTA approaches, including the ones that inspired our own approach. We gain 1.36%, 2.55% and 2.13% over NNCLR on CIFAR-10, CIFAR-100 and ImageNet-100 respectively, and similarly improve by 1.14%, 1.27% and 1.55% over Barlow Twins. This emphasizes the fact that we are able to outperform both the feature contrast approach and the neighbour contrast approach by a considerable margin by combining them and adding the new contrastive centroid loss.

Linear Evaluation on ImageNet. In table 2a, we compare NNCLR [14] and All4One on the complete ILSVRC2012 ImageNet [12] dataset for linear evaluation task. Considering the computational resources for ImageNet, we perform the linear evaluation only for 100 epochs with an effective batch size of 1024. We are able to outperform NNCLR on the larger ImageNet. This highlights the improvements in the performance of our approach on larger datasets.

Linear evaluation on Transformer Backbone. We study the All4One behaviour on a different backbone by replacing the ResNet with a Transformer backbone. We use a ViT-Small on CIFAR-100 for this study. In Table 2b, we see how All4One outperforms NNCLR verifying the independence of our approach on backbones. We have an improvement of 1.15% compared to the ViT version of NNCLR.

Semi-supervised ImageNet100 and ImageNet Evaluations. We perform semi-supervised evaluation following the experiments in NNCLR [14]. We fine-tune the Ima-

Method	Top-1	Top-5
NNCLR	65.74	86.90
All4One	**66.60**	**87.51**

(a) **Linear evaluation on ILSVRC2012 ImageNet.** For NNCLR, all hyperparameters except for the batch size (we use 1024 for both approaches) are the ones recommended in the original paper [14].

Method	Top-1	Top-5
NNCLR	68.55	90.94
All4One	**69.7**	**91.65**

(b) **Linear evaluation on CIFAR-100 using ViT-Small backbone.** Same hyperparameter settings are used for both methods.

	ImageNet100		ImageNet	
Method	1%	10%	1%	10%
NNCLR	54.14	75.49	37.51	58.74
All4One (Ours)	**58.73**	**76.95**	**38.96**	**60.14**

(c) **Semi-supervised learning results (Top-1 linear accuracy) on ImageNet100 and ImageNet.**

	Food-101	Caltech-101	Dogs	Pets
NNCLR	69.52	90.15	67.47	**83.34**
All4One	**71.16**	**91.10**	**68.07**	81.57

(d) **Transfer learning evaluation (Top-1 linear accuracy).**

Table 2: **All4One linear evaluation experiments.**

geNet pre-trained model (ResNet-50 on 100 epochs) on 1% and 10% subsets of the datasets. The results are presented in Table 2c. As can be seen, All4One generalizes better than NNCLR, outperforming it for both ImageNet100 and ImageNet on 1% and 10% subsets of the data.

Transfer Learning. Finally, we evaluate All4One and NNCLR on transfer learning downstream tasks for Food101 [20], Caltech-101 [36], Dogs [39] and Pets [29] dataset. For all datasets, we freeze the ImageNet pre-trained model (ResNet-50 on 100 epochs) and train a single linear classifier on top of it for 90 epochs on train splits while performing a sweep over the lr to obtain the best-performing one. Then, we evaluate the performance on the validation split for the Food101 dataset and test split for the rest of the datasets. The results are shown in Table 2d. As can be seen, All4One outperforms NNCLR in 3 out of 4 datasets, further validating the increased generalization capabilities of All4One.

4.3. Ablation Study

First, we show the importance of each objective defined by our approach. Then, we analyse the dimensionality and augmentation robustness of our model. Finally, we present some design choices such as the number of layers used by the transformer encoder and the number of extracted neighbours. All the ablations are done following exactly the same settings defined in Section 4.1 if not stated otherwise.

Objective Importance. As explained, our approach introduces redundancy reduction and novel neighbour contrast objectives. For this reason, we find it interesting to analyse, one by one, the importance of each of them. In Table 3, we report the performance of each objective of the framework. In addition, we also study the importance of EMA ($v2$ vs $v3$). We see that the addition of EMA boosts the overall performance of the model.

As can be seen with the different versions of All4One, all three objectives are important regarding the overall performance. Intuitively, both NNCLR [14] and Centroid-based objectives focus on contrasting neighbour image representations, so it is possible that, during the training, both objectives may partially overlap. This is not the case for the redundancy reduction, as it focuses on the features instead, causing its removal to be more critical than the others. Moreover, we designed the redundancy objective to use the representations z_i^1 and z_i^2 to compute the loss rather than using the neighbours. This fact adds more richness to the final loss function, as a total of three different representations (original neighbour nn_i^1, centroid derived from the neighbours c_i^1 and image representation z_i^1) are used for the unified objective.

On the other hand, we check that the Centroid objective is the one that increases the most on the NN retrieval accuracy, reaching 86.16% when combined with the NNCLR [14] and EMA architecture (experiment $v3$). As expected, contrasting contextual information from multiple neighbours encourages the model to create better representations easier to distinguish just by using a simple KNN operator. However, the introduction of different objectives, such as the redundancy reduction objective, forces the model to generalize more and, consequently, perform better on the downstream tasks.

Dimensionality Robustness. Redundancy reduction approaches such as Barlow Twins [38] highly depend on the dimensionality of the embeddings. Other approaches such as Opt-SSL [2] required high dimensional embeddings to provide SOTA results. Our approach, however, manages to outperform Barlow Twins with much lower dimensional embeddings as it does not only depend on the redundancy reduction. Even if the dimensionality of the embedding is low, the symbiosis formed avoids the decrease in performance, as can be seen in Table 4a. Another factor of consideration to use low dimensions is that $KNN(\cdot,\cdot)$ suffers

Method	NNCLR obj.	Cen. obj.	Feat. obj.	EMA	Top-1	k-NN Top-1	NN Top-1
All4One$_{v0}$	✓				69.62	62.16	68.8 (77.8*)
All4One$_{v1}$		✓			67.4	59.61	82.8
All4One$_{v2}$	✓	✓			71.02	63.21	85.28
All4One$_{v3}$	✓	✓		✓	71.08	63.83	**86.16**
All4One$_{v4}$		✓	✓	✓	71.31	63.72	80.6
All4One$_{v5}$	✓		✓	✓	71.64	64.58	78.8
All4One$_{v6}$	✓	✓	✓	✓	**72.17**	**64.84**	**82.16**

Table 3: **All4One objective function ablation study (using CIFAR-100).** All4One$_{v0}$ is equal to vanilla NNCLR. NN retrieval accuracy marked by * represents the NN retrieval obtained by increasing the queue size from 65503 to 98304 [14]

	Top-1	k-NN Top-1
Barlow Twins (2048)	71.21	63.11
Barlow Twins (256)	62.14	54.64
All4One (256)	**72.17**	**64.84**

(a) **Dimensionality analysis using CIFAR-100 dataset.**

Method	Top-1	k-NN Top-1
Barlow Twins [38]	39.66	30.89
NNCLR [14]	35.39	27.64
All4One (Ours)	**44.7**	**33.99**

(b) **Augmentation analysis using CIFAR-100.**

Layer number	Top-1	k-NN Top-1
3	72.17	64.84
6	71.86	64.50
9	71.75	64.52

(c) **Number of transformer layers.**

Number of NN	Top-1	k-NN Top-1
5	72.17	64.84
10	72.00	64.54
15	71.92	64.63
20	71.79	64.6

(d) **Number of NNs extracted.**

Table 4: **All4One ablation experiments.** Evaluated on CIFAR-100 for linear and k-NN classification.

from the curse of dimensionality, which decreases its efficiency when high dimensional embeddings are used.

Augmentation Dependency. SSL frameworks depend heavily on the augmentations used by the Pretext task generator to create hard positive samples. However, neighbour contrastive approaches [14, 21] empirically proved that they are naturally more robust to augmentation removals. NNCLR [14] proposes the removal of all the augmentations except random crop augmentation to compare NNCLR frameworks robustness with SimCLR [6] and BYOL [15] frameworks. We follow the same to check the robustness of All4One compared to NNCLR and Barlow Twins. As it can be seen in Table 4b, our approach proves to be more robust than Barlow Twins [38] and NNCLR [14] to augmentation removal. Intuitively, adding a term that uses multiple neighbours increases, even more, the richness of the loss function and makes the augmentations less relevant.

Number of Transformer Layers. We increase the transformer encoder complexity by increasing the number of encoder layers (Table 4c). We infer that more complex encoders decrease the overall performance of the model. In SSL paradigms, the main goal is to train an encoder or the backbone. We hypothesize that simplifying the transformer encoder encourages the backbone to produce more useful features, rather than letting the transformer encoder create them with their feed-forward layers, thus decreasing performance.

Number of Neighbours. We study the effect of the neighbours' number on All4One (Table 4d). Similar to previous works [14, 21], All4One is robust regarding the number of neighbours extracted, obtaining the best results when 5 of them are used. As we increase the number of neighbours, the contextual information obtained from the Self-Attention operations performed by the transformer encoder may get more sparse, decreasing slightly the model performance while also increasing the required computation time.

4.4. Discussions

Dimensionality of the Embeddings. Feature contrast approaches such as Barlow Twins [38] state high dimensional embeddings as a requirement for their frameworks. However, we prove that combining the basic ideas behind feature contrast with another instance discrimination strategy drastically reduces this dependency. Usually, different SSL approaches tend to be applied individually and obtain improvements. We show that by combining and complementing these ideas could lead to higher improvements.

NN Retrieval Increase and Comparison. In neighbour contrast approaches, the number of times the KNN operator retrieves a neighbour from the same semantic class as the query (NN retrieval accuracy) has been defined as critical. However, we prove that increasing this accuracy does

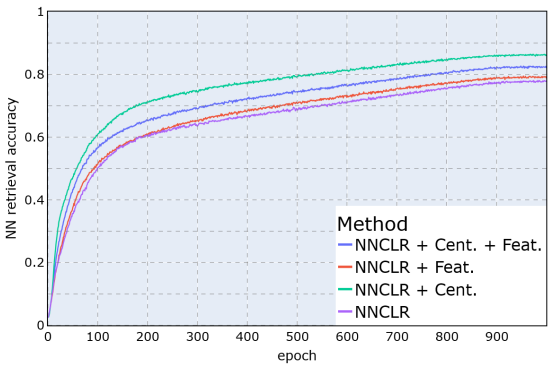

Figure 4: **Top-1 NN retrieval accuracy comparison.**

Figure 5: **NN extractions performed by All4One.**

not imperatively lead to a high overall accuracy increase. In fact, the best-performing version of All4One is not the one that obtains the highest NN retrieval accuracy. Hyperparameters such as the Support Set size or even the pretext task defined affect directly this accuracy. In Figure 4, we show how adding different pretext tasks affects the original NNCLR [14] method regarding the NN retrieval accuracy. As can be seen, adding the feature contrast strategy to vanilla NNCLR slightly boosts this performance. Also, combining NNCLR with a centroid pretext task provides a NN retrieval accuracy of 86%. However, when the three of them are combined, the NN retrieval accuracy does not surpass the 86% mark, even if it is the best overall performing version of All4One. This shows that the NN retrieval accuracy is not as critical as it is stated. In fact, as can be seen in Figure 5, neighbour contrast frameworks aim to bring together images that are similar so, even if the retrieved neighbour does not belong to the same semantic class, bringing it together with a very similar image on the feature space would not bring down the overall performance of the model. More NN extractions can be found in Section 5 of supplementary material.

4.5. Limitations

Even though All4One produced promising results, we identify some limitations.

Computation Efficiency. All4One, due to the introduction of three different objective functions, is more efficient than NN approaches that use multiple neighbours, but less efficient than NNCLR [14], which only contrasts a single neighbour. This is a limitation on low computation constraints. We provide a computation complexity analysis in the supplementary material.

Increased Number of Hyperparameters. The final All4One objective function introduces 3 additional hyperparameters to tune. Also, the transformer encoder uses a different learning rate, which also adds an extra hyperparameter to the overall framework.

Transformer Encoder Parameters. Several advances are available in terms of training transformers. The stability of the transformer encoder when using different settings compared to that of the other components is not known.

5. Conclusions

We propose a symbiotic approach that leverages NN contrastive learning by contrasting contextual information from multiple neighbours in an efficient way via self-attention. Also, we integrate a feature contrast objective function beneficial to the overall framework. All4One proves to generalize better and provide richer representations, outperforming previous SoTA contrastive approaches thanks to the integration of its different objectives. This highlights its exceptional performance in low data regimes, low dimensionality scenarios and weak augmentation settings. In the future, we plan to extend All4One to more complex backbones and investigate its application in diverse downstream tasks such as Object Detection and Instance Segmentation.

Acknowledgements

This work was partially funded by the Horizon EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00), PID2022-141566NB-I00 (AEI-MICINN), and CERCA Programme / Generalitat de Catalunya. B. Nagarajan acknowledges the support of FPI Becas, MICINN, Spain.

References

[1] Srikar Appalaraju, Yi Zhu, Yusheng Xie, and István Fehérvári. Towards Good Practices in Self-supervised Representation Learning, Dec. 2020. arXiv:2012.00868 [cs]. 6

[2] Nil Ballús, Bhalaji Nagarajan, and Petia Radeva. Opt-SSL: An Enhanced Self-Supervised Framework for Food Recognition. In Armando J. Pinho, Petia Georgieva, Luís F. Teixeira, and Joan Andreu Sánchez, editors, Pattern Recognition and Image Analysis, Lecture Notes in Computer Science,

pages 655–666, Cham, 2022. Springer International Publishing. 3, 5, 7

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021. 3

[4] H. Barlow. Redundancy reduction revisited. Network: Computation in Neural Systems, 12(3):241–253, Jan. 2001. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/net.12.3.241.253. 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In Advances in Neural Information Processing Systems, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, pages 1597–1607. PMLR, Nov. 2020. ISSN: 2640-3498. 2, 3, 6, 8

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. In Advances in Neural Information Processing Systems, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning, Mar. 2020. arXiv:2003.04297 [cs]. 3

[9] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. pages 15750–15758, 2021. 3

[10] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 539–546 vol. 1, June 2005. ISSN: 1063-6919. 2

[11] Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A Library of Self-supervised Methods for Visual Representation Learning. Journal of Machine Learning Research, 23(56):1–6, 2022. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, June 2009. ISSN: 1063-6919. 6

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 5

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a Little Help From My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. pages 9588–9597, 2021. 1, 2, 3, 4, 5, 6, 7, 8, 9

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In Advances in Neural Information Processing Systems, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 2, 3, 5, 8

[16] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–13, 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16000–16009, 2022. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. pages 9729–9738, 2020. 2, 3

[19] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Mar. 2015. arXiv:1502.03167 [cs]. 6

[20] Parneet Kaur, Karan Sikka, and Ajay Divakaran. Combining weakly and webly supervised learning for classifying food images. arXiv preprint arXiv:1712.08730, 2017. 7

[21] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean Shift for Self-Supervised Learning. pages 10326–10335, 2021. 2, 3, 4, 8

[22] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking Graph Transformers with Spectral Attention. In Advances in Neural Information Processing Systems, volume 34, pages 21618–21629. Curran Associates, Inc., 2021. 3

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012. 5

[25] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-Supervised Learning: Generative or Contrastive. IEEE Transactions on Knowledge and Data Engineering, 35(1):857–876, Jan. 2023. Conference Name: IEEE Transactions on Knowledge and Data Engineering. 1, 2

[26] Zhuang Ma and Michael Collins. Noise Contrastive Estimation and Negative Sampling for Conditional Models: Consistency and Statistical Efficiency, Sept. 2018. arXiv:1809.01812 [cs, stat]. 2

[27] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting Self-Supervised Learning via Knowledge Transfer. pages 9359–9367, 2018. 3

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, Jan. 2019. arXiv:1807.03748 [cs, stat]. 2, 3, 4, 5

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In 2012 IEEE conference on

computer vision and pattern recognition, pages 3498–3505. IEEE, 2012. 7

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems, 29, 2016. 2

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding, Dec. 2020. arXiv:1906.05849 [cs]. 2

[32] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? arXiv preprint arXiv:2201.05119, 2022. 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, Dec. 2017. arXiv:1706.03762 [cs]. 3, 4, 6

[34] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. KVT: k-NN Attention for Boosting Vision Transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, Computer Vision – ECCV 2022, Lecture Notes in Computer Science, pages 285–302, Cham, 2022. Springer Nature Switzerland. 3

[35] Lemeng Wu, Xingchao Liu, and Qiang Liu. Centroid Transformers: Learning to Abstract with Attention, Mar. 2021. arXiv:2102.08606 [cs, stat]. 3

[36] Xinxing Xu, Joey Tianyi Zhou, IvorW Tsang, Zheng Qin, Rick Siow Mong Goh, and Yong Liu. Simple and efficient learning using privileged information. arXiv e-prints, pages arXiv–1604, 2016. 7

[37] Yang You, Igor Gitman, and Boris Ginsburg. Large Batch Training of Convolutional Networks, Sept. 2017. arXiv:1708.03888 [cs]. 6

[38] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Proceedings of the 38th International Conference on Machine Learning, pages 12310–12320. PMLR, July 2021. ISSN: 2640-3498. 2, 3, 5, 7, 8

[39] Peisen Zhao, Lingxi Xie, Ya Zhang, and Qi Tian. Universal-to-specific framework for complex action recognition. IEEE Transactions on Multimedia, 23:3441–3453, 2020. 7

[40] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832, 2021. 3