IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 September 2023, accepted 14 October 2023, date of publication 19 October 2023, date of current version 26 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325842

RESEARCH ARTICLE

Self-Supervised Visual Representation Learning via Residual Momentum

TRUNG XUAN PHAM[1], (Student Member, IEEE), AXI NIU[2], (Student Member, IEEE), KANG ZHANG[1], (Student Member, IEEE), TEE JOSHUA TIAN JIN[1], (Student Member, IEEE), JI WOO HONG[1], (Member, IEEE), AND CHANG D. YOO[1], (Senior Member, IEEE)

[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Chang D. Yoo (cd_yoo@kaist.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant 2022R1A2C2012706; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT), through the Development of Uncertainty-Aware Agents Learning by Asking Questions under Grant 2022-0-00951.

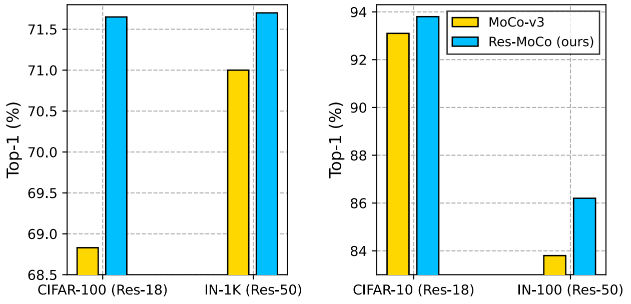
ABSTRACT Self-supervised learning (SSL) has emerged as a promising approach for learning representations from unlabeled data. Momentum-based contrastive frameworks such as MoCo-v3 have shown remarkable success among the many SSL methods proposed in recent years. However, a significant gap in encoder representation exists between the online encoder (student) and the momentum encoder (teacher) in these frameworks, limiting the performance on downstream tasks. We identify this gap as a bottleneck often overlooked in existing frameworks and propose "residual momentum" that explicitly reduces the gap during training to encourage the student to learn representations closer to the teacher's. We also reveal that a similar technique, knowledge distillation (KD), to reduce the distribution gap with cross-entropy-based loss in supervised learning is useless in the SSL context and demonstrate that the intra-representation gap measured by cosine similarity is crucial for EMA-based SSLs. Extensive experiments on different benchmark datasets and architectures demonstrate the superiority of our method compared to state-of-the-art contrastive learning baselines. Specifically, our method outperforms MoCo-v3 0.7% top-1 in ImageNet, 2.82% on CIFAR-100, 1.8% AP, and 3.0% AP75 on VOC detection pre-trained on the COCO dataset; it also improves DenseCL with 0.5% AP (800ep) and 0.6% AP75 (1600ep). Our work highlights the importance of reducing the teacher-student intra-gap in momentum-based contrastive learning frameworks and provides a practical solution for improving the quality of learned representations.

INDEX TERMS Contrastive learning, residual momentum, representation learning, self-supervised learning, knowledge distillation, teacher-student gap.

I. INTRODUCTION

Self-Supervised contrastive learning (SSL) has proven to be highly successful in the field of natural language processing (NLP) [1], [2] over the past few years. Recently, it has also emerged as a critical research paradigm in computer vision, owing to its unique advantage of not requiring expensive human labeling, as in the case with supervised learning frameworks [3], [4], [5], [6], [7]. In fact, SSL has outperformed traditional supervised pretraining methods in learning representations for a wide range of downstream tasks, including classification, segmentation, and object detection [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Without the ground-truth label, the core of most SSL methods lies in learning an encoder using augmentation-invariant representation [13], [18], [19], [20], [21], [22], [23], [24]. Amongst them, contrastive learning frameworks based on exponential moving averages (EMA or momentum) have attracted much attention. MoCo [25], MoCo-v2 [26], BYOL [20], DINO [27], ReSSL [28], DenseCL [29], and the

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.



(a) CIFAR-100 & ImageNet-1K (b) CIFAR-10 & ImageNet-100

FIGURE 1. Linear classification comparison on four datasets. We use ResNet-18 for CIFAR-10 and CIFAR-100; ResNet-50 for ImageNet-100 and ImageNet-1K. Models are trained 200ep for ImageNet-1K, and 1000ep for other datasets.

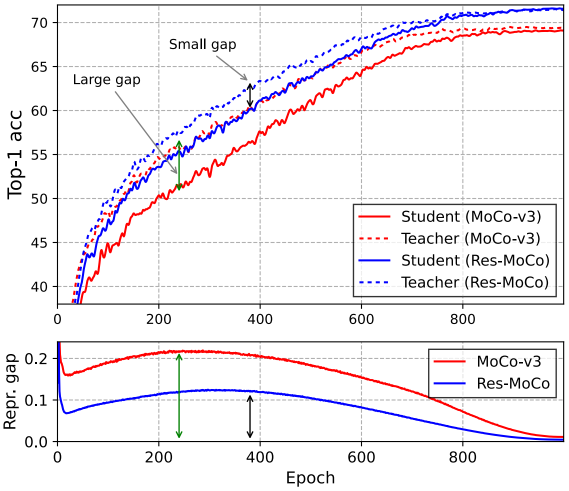


FIGURE 2. Intra-representation gap (IRAG) between teacher and student on CIFAR-100 (measured by Eq.5) and their corresponding performance gap. Our Res-MoCo significantly reduces the IRAG, narrowing their performance gap.

recent MoCo-v3 [9] are examples of the momentum-based frameworks that achieved great success in self-supervised visual representation learning. These frameworks use EMA to construct two branches of a Siamese architecture where one branch is with momentum encoder (called "teacher" [27] or "target" encoder [20]) and the other branch without it (called "student" or "online" encoder [20]). Specifically, MoCo [25] is a milestone in SSL that introduced a slow-moving average network (momentum encoder) to maintain consistent representations of negative pairs in a large memory bank using EMA. Without negative samples, BYOL [20] uses a moving average network to produce prediction targets to stabilize the bootstrap step and a simple cosine similarity loss to reduce the distance between the two distorted versions of an image. DINO [27] utilizes knowledge distillation (KD) to predict the output of a teacher constructed by an EMA encoder.

MoCo-v3 [9], on the other hand, has incorporated the best practices in the field to create a more powerful contrastive learning framework. Zhao et al. [13] proposed GLNet exploiting jointly global and local information under the EMA paradigm to improve MoCo-v3. Momentum-based approaches have considerably enriched the field of self-supervised learning with the teacher-student (TE-ST) formula. Prior works have carefully exploited the importance of TE-ST discrepancy, such as various knowledge distillation techniques in **traditional supervised learning** [15], [30], [31], [32], [33], [34] or **semi-supervised learning** [35], [36]. However, in **self-supervised learning** paradigm, this behavior, *i.e.* TE-ST gap, has not been adequately discovered. To fill this space, we investigate **if the ST-TE gap exists in the SSL context and how to solve it**.

We find that existing EMA-based SSLs remain a substantial gap between the two encoders. More importantly, we find that EMA-based SSLs concentrate solely on reducing the distance between representations of two different augmented views (**inter-view**) to learn *augmentation-invariant representation* and overlook the significance of reducing the distance between the representations of the same augmented view for the teacher and the student (**intra-view**). In order to address the potential TE-ST's representation gap in EMA-based SSLs, we focus on the intra-representation gap (IRAG) using the same view (Fig.3d) instead of the inter-representation gap (ITRG) with different views used in existing SSLs (Fig.3c).

Our investigation revealed a significant IRAG in existing SSLs during training (Fig.2 bottom), which leads to a large TE-ST's performance gap (Fig.2 top), hindering the student's capability. We show that the gap becomes more severe in large-scale datasets such as ImageNet or COCO. To address this issue, we propose "residual momentum" to directly reduce the IRAG and ultimately improve the performance of baseline models (Fig.1). Our contributions are as follows:

- We propose to focus on the representational gap between teacher and student within EMA-based SSL frameworks. Such a difference is carefully considered in supervised learning with KD but is overlooked in SSL contexts (Fig.3).
- We show that such a disparity (IRAG) can cause a substantial discrepancy in performance between the two models during training SSL, hindering students' ability to learn better representations.
- To address this issue, we introduce a residual momentum (referred to as "intra-momentum") during training, explicitly reducing the representation gap between the teacher and student in EMA-based SSLs. Our approach narrows their performance gap and significantly improves the student model's performance, and provides a complete picture of supervised learning and self-supervised learning from the perspective of the TE-ST gap.
- We evaluate the effectiveness of our approach on challenging benchmark datasets and various network architectures. Our experimental results demonstrate that our approach outperforms state-of-the-art CL baselines and achieves superior performance.

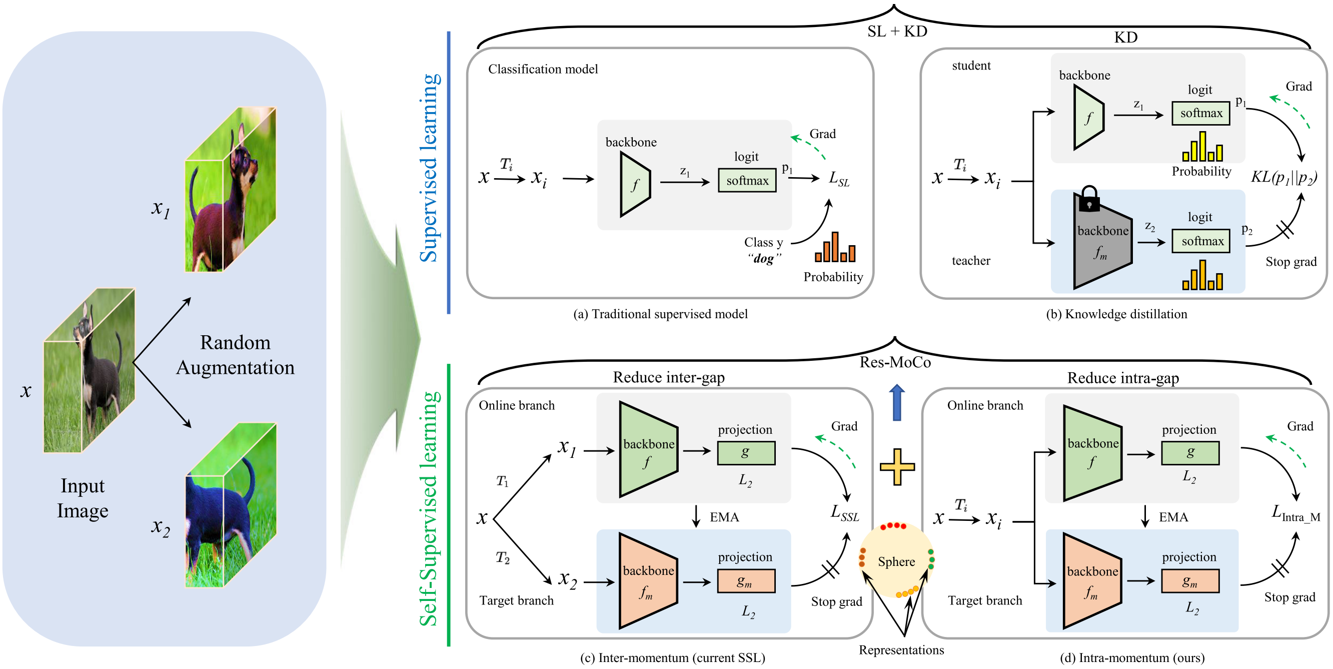


FIGURE 3. KD in supervised learning shares some similarities with intra-momentum in self-supervised learning but differs in some aspects, as discussed in the main text. Besides, the comparison of the existing CL frameworks that use Inter-momentum and the proposed method with Intra-momentum. T_i with $i \in \{1, 2\}$ is a random transformation to the given input image x. As the term suggests, existing SSL with Inter-M uses different inputs, *i.e.* x_1 and x_2 for student and teacher, respectively. By contrast, a model with Intra-momentum uses the same input x_i with $i \in \{1, 2\}$ for both the student and teacher model and uses $\mathcal{L}_{\text{Intra\_M}}$ to minimize their representation gap. Our final method, Res-MoCo, jointly optimizes the intra- and inter-momentum to achieve the best learning capability, provides a complete picture with supervised learning in the TE-ST perspective.

II. RELATED WORKS

A. MOMENTUM-BASED SELF-SUPERVISED LEARNING

Exponential moving average (EMA or momentum) has been deeply studied for smoothing the original sequence signal [37], [38], optimization [39], [40], [41], reinforcement learning [42], [43], [44], knowledge distillation [45], [46], and recent semi-supervised learning [36], [47], [48]. Recently, EMA has also been applied in modern self-supervised learning frameworks. Amongst the seminal works, MoCo [9], [25], [26], DINO [27], ReSSL [28], and BYOL [20] are examples that use the EMA in the target encoder to prevent model collapse [20], [27], obtain consistent negative samples [25], [26], or boost performance as in [9] and [13]. Seyfi et al. [21] proposed XMoCo to extend momentum contrast to regularize the consistency and improve MoCo-v2 with a notable margin. Cheng et al. [49] found that constructing distributed augmentation invariance in an EMA-based SSL can improve the learned feature quality.

In practice, EMA-based SSL approaches have demonstrated state-of-the-art performance for downstream tasks compared to EMA-free SSL counterparts [19], [21], [49], [50], [51]. The EMA-based frameworks contain two branches (as shown in Fig.3): the first branch is an encoder that allows back-propagation [52] during training, which refers to as *online* encoder (student). The second branch is *target* encoder (teacher), which is constructed by a momentum encoder whose parameters are dynamically updated via the *online* encoder using Eq.1. Different from all the above research that only focuses on augmentation-invariant representation learning tasks, *our work explores and addresses the disparity in representation between the teacher and student to improve EMA-based SSL frameworks such as MoCo-v3.*

B. TEACHER-STUDENT GAPS

There have been early works trying to reduce the distribution gap between the student and teacher networks to maximize the performance of the student model [53]. In Knowledge Distillation (KD) [53], the knowledge from a larger and better performing model (teacher) is used to generate the soft targets for a smaller student model, hence reducing the distribution gap between the two models (the teacher model is often a pre-trained/fixed model as Fig.3b).

For KD, however, earlier works show the sharpness gap [31] (with adaptive temperature), confidence gap [54] (with normalized logits), and capacity gap [30] (with gradient similarity) between teacher and student, preventing the student model's capability. The other works of self-knowledge distillation (self-KD) try to use students themselves as teachers. In [55], the self-KD model is trained to reduce the distance between features extracted from two separate distorted versions of an image by a KL divergence loss.

The self-training in [36] may be close to our work where the distribution distance between outputs of the teacher and student is minimized to improve generalizability. There are three points that the mean teacher model in [36] is different from our work. First, the mean teacher approach works in a supervised learning paradigm where our proposed *intra-momentum* works in a self-supervised manner, *i.e.* without any labels involved. Second, [36] minimizes the distribution distance or prediction of labels with MSE loss on softmax outputs. By contrast, *intra-momentum* is trained to minimize the representation gap between teacher and student with cosine similarity loss (*i.e.* no softmax applied). And third, in [36], the teacher and student have different inputs with injected noises η and η'. By contrast, our proposed *intra-momentum* uses the same input for both student and teacher models (see Fig.3).

A key difference between KD in *supervised learning* and *self-supervised contrastive learning* is the output of the latent vectors. Specifically, SL and KD produce the softmax to obtain the **probability distribution** of each labeled class, while SSL produces the **representations** that lie on some **hyperspheres** (without labels) [56] as clearly illustrated in Fig.3.

As discussed in [27], momentum-based SSL frameworks [9], [13], [20], [21], [25], [27] have a form of KD.

For these SSLs, we notice that two different augmented images (positive pair) are fed separately into the *teacher* and *student* to learn the augmentation-invariant representation. During training, a loss function (either cross-entropy [27], contrastive loss [9], [25], [28], or a simple cosine similarity [20]) is applied in their outputs to minimize their gap. We refer to this gap as *inter-representation gap* (ITRG) since they use different inputs and refer to the momentum in these frameworks as *inter-momentum*.

In contrast to the existing inter-momentum used in SSLs, in this work, we focus on the representation gap between the teacher (TE) and the student (ST) but for a given the same input image. We refer to this gap as *intra-representation gap* (IRAG), which can be minimized by the proposed *intra-momentum*. Compared to the existing *inter-momentum*, the *intra-momentum* has unique properties: *First*, it is augmentation agnostic. *Second*, different from the traditional inter-momentum integrated into the existing SSL loss, the proposed *intra-momentum* is decoupled from the existing SSL loss and can be flexibly plugged into other CL frameworks (*e.g.* Eq.9 and Eq.10).

Motivated by earlier KDs [30], [31], [32], [33], [34], [35], which reduce various TE-ST gaps mentioned above that achieve great success in supervised learning. We notice that *"no existing work shows if the KD losses can work in SSL beyond the supervised loss"*. This work provides comprehensive empirical results showing that **"minimizing the distribution gap as these KD techniques are useless in the SSL context"** (Section V-A). Our intra-momentum is proposed to reduce the "representation gap" in self-supervised learning by an explicit distance function, *i.e.* cosine similarity with ℓ_2-norm (Eq.5) that significantly helps the baselines (see Section V-A). The comparison between the traditional KD and our method is visualized in Fig.3.

In the following sections, we present the effectiveness of our proposed method in comparison to strong EMA-based CL seminal frameworks, including MoCo-v3 for classification pretraining and DenseCL for dense prediction pretraining.

III. METHOD

In this section, we first present the background of SSL. After that, we analyze how *inter-momentum* is applied in the existing SSL frameworks, since MoCo-v3 is the latest version of MoCo series, which employed the best practices in the area, we choose mainly MoCo-v3 as our CL baseline. Then, we introduce the preliminaries of the proposed *intra-momentum* to narrow the "intra-representation gap". Finally, we formulate the objective function of the final method, which consists of both "inter-momentum" and "intra-momentum".

A. BACKGROUND

Given an unlabeled image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, two random augmentations $\mathbf{x}_1 \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{x}_2 \in \mathbb{R}^{H\times W\times 3}$ are generated and formed as a positive pair, the augmentations from the other images in the current mini-batch are treated as the negative samples [19], [26]. The two augmented images $\mathbf{x}_1$ and $\mathbf{x}_2$ are separately fed into two different encoders, *i.e. student* and *teacher* to train the *student* model to learn the augmentation-invariant representation.[1] We consider an online encoder (student), *i.e.* $\mathbb{E}$ with a backbone f (*e.g.* ResNet-50), projector g [19], [25] and may be followed by a predictor q [9], [20]. The target encoder (teacher), *i.e.* $\mathbb{E}_m$ (f_m, g_m, q_m) has the same architecture as the student network. The subscript m beside each character denotes momentum. The standard backpropagation [52] updates the parameters of the student network while the teacher's parameters are momentum updated as follows [9], [20]:

$$\xi \leftarrow \beta\xi + (1-\beta)\theta, \tag{1}$$

where θ and ξ are the parameters of the student $\mathbb{E}$ and teacher $\mathbb{E}_m$, respectively. A constant $\beta \in (0, 1)$ is the momentum coefficient which is often chosen with a value of 0.99 for short training (200 epochs) [20], [25], [50]. In longer training, *i.e.* 800 / 1000 epochs for full convergence, SSL methods often use the higher momentum value, *i.e.* $\beta = 0.996$ [9], [20], [27]. The presented framework is commonly used in most EMA-based SSL approaches. We specify the MoCo-v3 design as a baseline in the next section.

B. INTER-MOMENTUM TO MINIMIZE THE INTER-REPRESENTATION GAP IN SSL

MoCo [25] is the first work to introduce EMA for self-supervised contrastive learning and has become a ground-breaking and highly recognized framework. Its latest version, *i.e.* MoCo-v3 [9], employs the best practices in the SSL area. Specifically, MoCo-v3 consists of the online encoder $\mathbb{E}$ with (f, g, q) and the target encoder $\mathbb{E}'_m$ with (f_m, g_m). Note that there is no predictor on the target encoder. Two crops $\mathbf{x}_1$ and $\mathbf{x}_2$ are embedded by $\mathbb{E}$ and $\mathbb{E}'_m$ to have the outputs q (query) and k_m (key). The objective function of MoCo-v3 is adopted

[1] To avoid ambiguity, we use following concepts interchangeably: (*online encoder*, *student*); (*target encoder*, *momentum encoder*, *teacher*).

by InfoNCE loss [57]:

$$\mathcal{L}_{\text{ctr}} = -\log \frac{\exp(q{\cdot}k_m^{+}/\tau)}{\exp(q{\cdot}k_m^{+}/\tau) + \sum\limits_{k_m^{-}} \exp(q{\cdot}k_m^{-}/\tau)}, \tag{2}$$

where $(\cdot)$ denotes cosine similarity, k_m^{+} is the output of $\mathbb{E}'_m$ for the augmentation of a same image as the query q (positive sample), k_m^{-} is the negative samples of q. Symbol τ is a temperature hyper-parameter [19], q and k are l_2-normalized [9]. For every sample $\mathbf{x}$ in the current mini-batch, the above loss is symmetrized as follows MoCo-v3 [9]:

$$\mathcal{L}_{\text{Inter-M}} = \frac{1}{2}\left(\mathcal{L}_{\text{ctr}}(q_1, k_{2,m}) + \mathcal{L}_{\text{ctr}}(q_2, k_{1,m})\right). \tag{3}$$

Here we put "Inter-M" to the subscript to denote that momentum used in MoCo-v3 (and all existing SSL frameworks) to construct the teacher and student that uses two different augmented images as inputs. Obviously, "inter-momentum" (Inter-M) in Eq.3 is designed to minimize the distance between two different inputs (inter-representation gap), *i.e.* making $\mathbf{x}_1$ and $\mathbf{x}_2$ closer to learn the invariant-augmentation representations during the training process.

Next, we introduce a novel residual momentum ("intra-momentum"), which is trained to have representations of the teacher and student as close as possible.

C. INTRA-MOMENTUM TO MINIMIZE THE INTRA-REPRESENTATION GAP IN SSL

Previous works, *e.g.* MoCo-v3 in the last section, only focus on Inter-M to minimize the gap between the teacher and student's outputs from two different augmented images while ignoring the potential representation gap between the teacher and student for the same augmented view.

As shown in Fig.2, such an unaware discrepancy causes a big gap in their performance, which prevents the student model from learning good representation. To this end, we propose to measure the intra-representation gap between teacher and student models using the same input via the lens of intra-momentum (Intra-M).

Without changes in the architecture of MoCo-v3, we consider the online encoder $\mathbb{E}(f, g, q)$ and the momentum encoder $\mathbb{E}_m(f_m, g_m, q_m)$. MoCo-v3 uses the output of g_m as the target (asymmetry), but Intra-M uses the output of q_m (symmetry). We define the intra-representation gap as follows:

$$\mathcal{L}_{\text{Intra-gap}} = \mathcal{D}(q, q_m), \tag{4}$$

where $\mathcal{D}$ is a distance function. In this paper, we consider three choices for the distance function to narrow the gaps between the two vectors (either the distribution gap or the representation gap). *First*, we use the negative cosine similarity function (*default* to measure representation gap) as follows [20]:

$$\mathcal{D}(q, q_m)_{\text{cosine}} = \|q - q_m\|_2^2 = 2 - 2 \cdot (q \cdot q_m), \tag{5}$$

where $\|.\|$ denotes the ℓ_2-norm. Vectors q and q_m are ℓ_2-normalized. *Second*, we ablate the other choice of distance function with the entropy function H as used in KD for supervised learning (distribution gap) [27], [53]:

$$\mathcal{D}(q, q_m)_{\text{CE}} = H(q, q_m) = -P(q)\log(P(q_m)), \tag{6}$$

where $P(x)$ is the softmax output of a vector $x \in \mathbb{R}^K$: $P(x)^{(i)} = \frac{\exp(x)^{(i)}/\tau_s}{\sum_{k=1}^{K}\exp(x)^{(k)}/\tau_s}$, τ_s is a temperature parameter which the common choices in KD are {3,4,5} [53]. Here, q and q_m are not ℓ_2-normalized. And *third*, we verify the usage of the mean square error (MSE) to the softmax outputs of the student and teacher (distribution gap) as in [36]:

$$\mathcal{D}(q, q_m)_{\text{MSE}} = \frac{1}{2}(q - q_m)^2, \tag{7}$$

where q and q_m are not ℓ_2-normalized but softmax-normalized. Note that q and q_m are the outputs of the online and momentum encoder from the same image augmentation input. Our experiments show that Intra-M using cosine similarity performs the best compared to MSE or CE in top-1 linear accuracy. This suggests that cosine similarity for *intra-momentum* is more suitable in contrastive learning. We also tried the asymmetric design for Intra-M, *i.e.* $\mathcal{L}_{\text{Intra-gap}} = \mathcal{D}(q, g_m)$ however, it performs worse than $\mathcal{L}_{\text{Intra-gap}} = \mathcal{D}(q, q_m)$. This suggests that the asymmetry does not benefit Intra-M. Finally, following $\mathcal{L}_{\text{Inter-M}}$ (Eq.3), we symmetrize loss for Intra-M as follows:

$$\mathcal{L}_{\text{Intra-M}} = \tfrac{1}{2}\left(\mathcal{L}_{\text{Intra-gap}}(q_1, q_{1,m}) + \mathcal{L}_{\text{Intra-gap}}(q_2, q_{2,m})\right). \tag{8}$$

1) DISCUSSIONS

For the Inter-M, x_1 and x_2 are forwarded two times to the student and teacher encoder, resulting in a total of 4 forwards to the backbone and projector (symmetric loss). Our Intra-M uses the same outputs g_m of Inter-M; the additional costs come only from an MLP q_m and the backward pass, which are negligible. As shown in Tab.12, Intra-M causes only 5s slower than MoCo-v3 when pre-trained on ImageNet (the overhead is less than 1%).

Eq.5 and Eq.7 are used to minimize the representation gap between vectors q and q_m, but Eq.6 is widely used in KD [53] to match the probability distribution between two vectors, we included for completeness and comparisons.

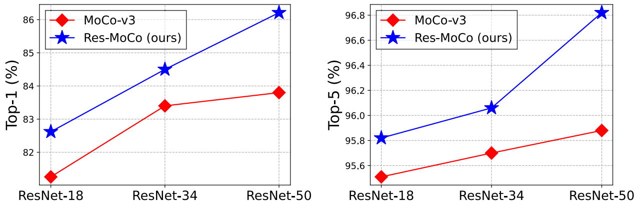
D. FINAL OBJECTIVE FUNCTIONS

1) Res-MoCo

We adopt MoCo-v3 [9] as our baseline for the image classification pretraining. Our final method is named Res-MoCo (residual momentum contrastive learning) Fig.3, whose loss is a combination of the two momentum losses (Eq.3 and Eq.8) for joint optimization as follows:

$$\mathcal{L}_{\text{Res-MoCo}} = \mathcal{L}_{\text{Inter-M}} + \lambda\mathcal{L}_{\text{Intra-M}}. \tag{9}$$

Here, λ is a hyperparameter used to control the effect of the EMA of Intra-M. By default, Res-MoCo uses cosine similarity (Eq.5) for $\mathcal{L}_{\text{Intra-M}}$. Eq.9 shows that Res-MoCo is



(a) Top-1 linear accuracy (%) (b) Top-5 linear accuracy (%)

FIGURE 4. Impact of Intra-M on different backbone architectures. We compare MoCo-v3 and the proposed Res-MoCo (adding Intra-M) on IN-100, pretraining for 1000ep.

designed to narrow the overall representation gap between the student and teacher, including *inter-representation gap* (existing MoCo-v3) and *intra-representation gap* (this work). We analyze the impact of each component in the ablation section. Compared to MoCo-v3, our Res-MoCo achieves superior performance for all architectures and datasets on downstream tasks as highlighted in Fig.1 and Fig.4.

To explore the effect of the proposed residual momentum, we consider also DenseCL [29] (a CL baseline for dense prediction pretraining) to conduct experiments. Our objective function, in this case, is as follows:

$$\mathcal{L}_{\text{Res-DenseCL}} = \mathcal{L}_{\text{DenseCL}} + \lambda \mathcal{L}_{\text{Intra-M}}. \tag{10}$$

We set λ=1 by default. In optimal setting [29], $\mathcal{L}_{\text{DenseCL}}$ has two equal-weighted loss terms (w/Inter-M): global term $\mathcal{L}_q$, and dense term $\mathcal{L}_r$. Intra-M has been added to each term to minimize the corresponding teacher-student representational gap for dense and global representations.

IV. EXPERIMENTS

A. SETUP AND EVALUATION

1) DATASETS

We use CIFAR-10/100 (10/100 classes) [58], ImagetNet-100 (IN-100, 100 classes) [59], and ImageNet-1K (IN-1K, 1000 classes) [60] for classification and VOC07+12 [61]/ COCO2017 [62] for object detection. We consider different backbones for Res-MoCo, such as ResNet-18, ResNet-34, ResNet-50 [3], and ViT-S [63].

2) PRE-TRAINING SETUP

The experiments use the SSL library [64], [65] with their optimal tuned hyper-parameters, keeping the same settings for MoCo-v3/Res-MoCo and DenseCL/Res-DenseCL. The encoder is trained without labels on the training set of each dataset. We trained ResNet-18 for 1000ep on CIFAR-10/100 and IN-100, while ResNet-50 is used for larger datasets IN-1K and COCO train2017, trained for 200ep/800ep and 800ep/1600ep, respectively.[2]

3) EVALUATION

We evaluate the proposed SSL framework, following prior works [12], [29], in linear classification and transfer learning for object detection. Pre-trained models are assessed by training a linear classifier on frozen representations using the test set [9], [65]. Object detection tasks utilize ResNet-50 pre-trained by Res-MoCo on IN-1K to initialize Faster R-CNN and fine-tune with a standard 2x schedule [29], [50] for VOC07+12 and COCO datasets. Finally, we evaluate the transferability of the COCO pre-trained Res-DensCL/Res-MoCo models to VOC07+12.

TABLE 1. CIFAR-10. Comparison of MoCo-v3 and Res-MoCo. All methods are trained in 1000ep using the same settings on ResNet-18. We employ the results for other SSL methods from the official solo-learn library [65].

Method	Inter-M	Intra-M	Top-1 (%)	Top-5 (%)	KNN-1 (%)
W-MSE [66]	-	-	88.67	99.68	-
SwAV [67]	-	-	89.17	99.68	-
SimCLR [19]	-	-	90.74	99.75	-
SimSiam [50]	-	-	90.51	99.72	-
VICReg [68]	-	-	92.07	99.74	-
NNCLR [69]	-	-	91.88	99.78	-
Barlow Twins [51]	-	-	92.10	99.73	-
ReSSL [28]	✓	-	90.63	99.62	-
DINO [27]	✓	-	89.52	99.71	-
BYOL [20]	✓	-	92.58	99.79	88.84
MoCo-v3 [9]	✓	-	93.10	99.80	89.12
Res-MoCo (ours)	-	✓	**93.53** +0.43	**99.88** +0.08	**90.78** +1.66
Res-MoCo (ours)	✓	✓	**93.81** +0.71	**99.84** +0.04	**90.78** +1.66

TABLE 2. CIFAR-100. Comparison of MoCo-v3 and Res-MoCo. All methods are trained in 1000ep using the same settings on ResNet-18. We employ the results for the other SSL methods from the official solo-learn library [65].

Method	Inter-M	Intra-M	Top-1 (%)	Top-5 (%)	KNN-1 (%)
W-MSE [66]	-	-	61.33	87.26	-
SwAV [67]	-	-	64.88	88.78	-
SimCLR [19]	-	-	65.78	89.04	58.52
SimSiam [50]	-	-	66.04	89.97	59.01
VICReg [68]	-	-	68.54	90.83	-
NNCLR [69]	-	-	69.62	91.52	-
Barlow Twins [51]	-	-	70.84	92.04	62.35
ReSSL [28]	✓	-	65.92	89.91	59.05
DINO [27]	✓	-	66.76	88.63	56.58
BYOL [20]	✓	-	70.06	92.12	61.83
MoCo-v3 [9]	✓	-	68.83	90.07	60.75
Res-MoCo (ours)	-	✓	**69.85** +1.02	**92.47** +2.40	**62.41** +1.66
Res-MoCo (ours)	✓	✓	**71.65** +2.82	**92.32** +2.25	**64.27** +3.52

B. MAIN RESULTS

1) LINEAR CLASSIFICATION

In Tab.1 and Tab.2, we report the results on CIFAR-10/100 (size of 32 × 32) compared Res-MoCo to its baseline MoCo-v3 and the other state-of-the-art methods. On CIFAR-10, Res-MoCo outperforms MoCo-v3 for all metrics with 93.81% (+0.71%), 99.84% (+0.04%), and 90.78% (+1.66%) on top-1, top-5, and KNN-1 accuracy, respectively, and surpassing other state-of-the-art methods.

On CIFAR-100, the improvement is more significant when Res-MoCo outperforms baseline MoCo-v3 with +2.82%, +2.25%, and +3.52% on top-1, top-5, and KNN-1, respectively. Tab.3, Tab.4, Tab.9 report the results for the large-scale dataset (224 × 224), IN-100 and IN-1K. Res-MoCo shows a clear improvement over MoCo-v3, *i.e.* +0.7% and +2.41% accuracy on IN-1K and IN-100, respectively, demonstrating the effectiveness of *Intra-M*.

[2] As a common practice, we use the starting momentum $\beta = 0.99$ for 200ep and $\beta = 0.996$ for 800-1600ep. The value β increases to 1.0 using a cosine schedule [9], [20].

TABLE 3. Linear Classification accuracy ImageNet-1K vs. Object Detection. All SSL methods used ResNet-50 pre-trained on IN-1K. Detection results of MoCo-v3/ Res-MoCo are run three times to get an average. Our Res-Moco outperforms all competitors with a visible margin.

Method	ImageNet-1K			VOC07+12			COCO		
	Epoch	Bsz	Lin.	AP	AP50	AP75	AP	AP50	AP75
Random Init.	-	-	-	33.8	60.2	33.1	32.8	50.9	35.3
Pre-Train Sup.	200	256	**76.2**	54.2	81.6	59.8	38.2	58.2	41.2
MoCo (CVPR20) [25]	200	256	60.6	55.9	81.5	62.6	38.5	58.3	41.6
SimCLR (ICML20) [19]	1000	4096	69.3	56.3	81.9	62.5	37.9	57.7	40.9
SimSiam (CVPR21) [50]	800	256	71.3	56.4	82.0	62.8	37.9	57.5	40.9
MoCo-v2 (Arxiv20) [26]	800	256	71.1	57.4	82.5	64.0	39.3	58.9	42.5
SwAV (NeurIPS20) [67]	1000	4096	71.8	56.1	82.6	62.7	38.4	58.6	41.3
MoCo-v2 (Arxiv20) [26]	200	256	67.5	-	-	-	-	-	-
DenseCL (CVPR21) [13]	200	256	63.6	-	-	-	-	-	-
XMOCO (TCSVT22) [21]	200	256	65.0	-	-	-	-	-	-
GLNet (TCSVT22) [13]	200	256	70.5	-	-	-	-	-	-
BDAI (TCSVT23) [49]	200	4096	71.4	56.4	82.6	62.8	37.9	57.9	40.9
DimCL (Access23) [70]	200	512	70.8	56.1	82.1	62.9	38.5	57.5	41.9
MoCo-v3 (ICCV21) [9]	200	1024	71.0	56.2	82.4	62.9	39.1	58.8	42.2
MoCo-v3 (ICCV21) [9]	800	1024	72.4	56.7	82.5	63.6	39.3	59.1	42.5
Res-MoCo (ours)	200	1024	**71.7**	**56.5**	**82.6**	**63.0**	**39.5**	**59.2**	**42.9**
Res-MoCo (ours)	800	1024	**73.1**	**57.2**	**83.0**	**64.1**	**39.7**	**59.5**	**43.1**

TABLE 4. ImageNet-100. Compare baselines and Res-MoCo. Methods are trained for 1000ep (ResNet-18).

Method	Inter-M	Intra-M	Top-1 (%)	Top-5 (%)	KNN-1 (%)
SimCLR [19]	-	-	78.46	-	73.22
MoCo-v2 [26]	✓	-	79.98	-	74.56
BYOL [20]	✓	-	81.46	95.26	75.34
MoCo-v3 [9]	✓	-	81.26	95.51	75.28
Res-MoCo (ours)	-	✓	**81.66**$_{+0.40}$	**95.86**$_{+0.35}$	**76.04**$_{+0.76}$
Res-MoCo (ours)	✓	✓	**82.62**$_{+1.36}$	**95.82**$_{+0.31}$	**76.14**$_{+0.86}$

TABLE 5. Transfer to VOC07+12 object detection. All SSL methods are pretrained on COCO dataset with ResNet-50.

Method	Inter-M	Intra-M	800ep			1600ep		
			AP	AP50	AP75	AP	AP50	AP75
Rand. Init.	-	-	33.8	60.2	33.1	33.8	60.2	33.1
Super. IN-1K	-	-	54.2	81.6	59.8	54.2	81.6	59.8
MoCo-v3	✓	-	53.6	80.6	59.0	53.6	80.7	58.4
Res-MoCo (ours)	✓	✓	**55.1**	**81.3**	**60.8**	**55.4**	**81.8**	**61.4**
MoCo-v2	✓	-	54.8	-	-	55.0	-	-
DenseCL	✓	-	56.5	81.8	63.0	57.1	82.1	63.3
Res-DenseCL (ours)	✓	✓	**57.0**	**82.0**	**63.5**	**57.4**	**82.6**	**63.9**

2) TRANSFER LEARNING IN-1K TO COCO/VOC

In Tab.3, the quality of representations is evaluated by transferring them to object detection. The pre-trained models are finetuned end-to-end in the target datasets using the public code [25], [65]. Res-MoCo shows competitive results among the leading methods, outperforms MoCo-v3 for all metrics as well as all considered datasets **in the same setting batch size**, and surpasses all other COCO competitors for AP, AP50, and AP75. This demonstrates Intra-M help learn better quality representations for downstream object detection beyond classification tasks.

3) TRANSFER LEARNING COCO TO VOC

Tab.5 presents a comparison of transfer learning performance from COCO [62] to VOC [61] across Res-MoCo and Res-DenseCL, along with their respective baselines. The results demonstrate the superiority of our methods, as both Res-MoCo and Res-DenseCL outperform baselines in all settings.

Specifically, Res-MoCo achieves a 1.5-1.8% improvement in AP and AP75 over MoCo-v3, while Res-DenseCL outperforms DenseCL by 0.3-0.6% on these metrics. Notably, MoCo-v3 trained on COCO performs worse than supervised IN-1K (53.6% vs. 54.2% AP, -0.6%), while Res-MoCo outperforms it with 55.1% (+0.9%) (800ep) and 55.4% (+1.2%) (1600ep). These results indicate the importance of narrowing the TE-ST gap in SSL, and highlight the superior quality of learned representations in our models.

V. ABLATION STUDY AND ANALYSIS

We have provided the most important results for small-scale (CIFAR, IN-100) and large-scale (IN-1K, COCO). We follow recent SSL papers, such as ECCV22 [71] and CVPR22 [12], to mainly perform ablation studies with CIFAR-10/100 or IN-100 for efficiency in our hardware resource.

A. GAP REDUCTION FUNCTIONS

Our first choice to reduce the representation gap between the teacher and student is *cosine similarity* (CS) loss. We also ablate the other choices of the distance function such as *cross-entropy* (CE) or mean square error (MSE). We find that in SSL, CS loss performs the best compared to CE in KD [53], or MSE in self-training [36]. From Tab.6, we obtained the following interesting insights between SL and SSL:

- Reducing the probability distribution gap, the widely-used technique with KL divergence or cross-entropy

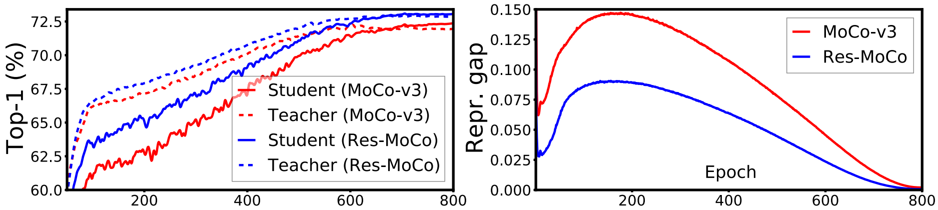


FIGURE 5. Training on large-scale IN-1K dataset. We show linear accuracy (left) and the corresponding intra-representation gap (right). In this setting, the existing work MoCo-v3 suffers from a severe mismatch between the online (student) and target (teacher) models, hindering the learning ability. By contrast, Intra-M helps reduce the gap to improve learning in Res-MoCo.

TABLE 6. Choices of $\mathcal{L}_{\text{Intra-M}}$. Linear classification accuracy on CIFAR-100, when trained 1000ep when all models are fully converged. Cosine function performs the best.

Intra-Gap Type	Method	Top 1	Top-5	KNN-1
N/A	MoCo-v3	68.83	90.07	60.75
Distribution (KD softmax)	Res-MoCo w/ MSE (Eq.7)	68.73	90.92	60.61
Distribution (KD softmax)	Res-MoCo w/ CE (Eq.6)	68.92	90.85	62.14
Representation (L_2-norm)	**Res-MoCo w/ CS (Eq.5)**	**71.65**	**92.32**	**64.27**

loss in KD [30], [32], [33], [72], [73] (very successful for **supervised learning**) is unsuitable for SSL. This makes sense because, in supervised learning, the model is trained to match the input data to some class distributions via a cross-entropy loss to produce the class probability.

- In the context of **self-supervised learning**, the target is to train a model to learn some useful representations from input data, normally with contrastive loss or negative cosine similarity. The encoder is trained to produce representations of images that lie into some hyperspheres with alignment and uniformity [56]. Therefore, reducing the representational gap with cosine similarity function as *intra-momentum* is more suitable for learning representations in contrastive learning frameworks.

B. TE-ST'S REPRESENTATION GAPS IN SSL

We present learning curves on large-scale datasets to demonstrate the intra-representation gap in Res-MoCo with and without residual momentum (Intra-M). The results of ImageNet-1K in Fig.5 indicate that Res-MoCo significantly reduces the representation disparity between the teacher and student (MoCo-v3, max magnitude gap $\approx$ 0.15). This leads to a corresponding improvement of the student model performance during most epochs of training.

In the non-object-centric datasets such as COCO (Fig.6), the TE-ST's representation discrepancy is more serious for MoCo-v3, with a max magnitude gap of $\approx$ 0.3; Res-MoCo effectively narrows this gap, resulting in significantly better AP ($\approx$2%, left chart). DenseCL [29] is optimized for dense prediction pretraining tasks; however, our experiments show that there is still a big TE-ST gap in both global and dense projection when training on COCO, which negatively impacts the student's capability.

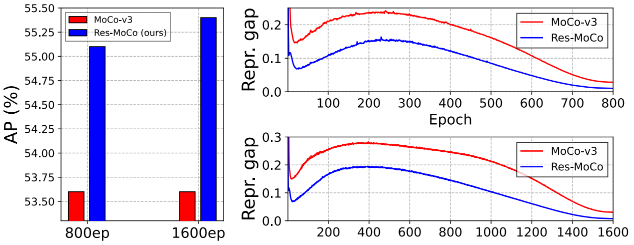


FIGURE 6. Performance on VOC07+12 (left) vs. Repr. gap when trained on COCO with 800ep and 1600ep (right).

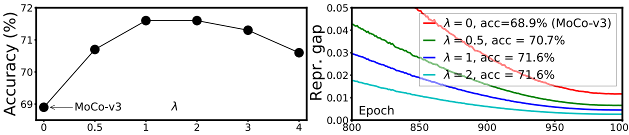


FIGURE 7. Impact of Intra-M on CIFAR-100, λ in Eq. 9.

TABLE 7. Impact of embedding and batch size, top-1 (%). Results are performed using CIFAR-100, trained 1000ep.

Method	Batch size				Embedding size				
	128	256	512	1024	128	256	512	1024	4096
MoCo-v3	68.5	68.9	69.1	69.0	68.6	68.9	68.8	69.0	68.7
Res-MoCo	**71.2**	**71.6**	**71.3**	**71.1**	**71.5**	**71.6**	**71.2**	**71.4**	**71.3**
Gain (%)	**+ 2.7**	**+ 2.7**	**+ 2.2**	**+ 2.1**	**+ 2.9**	**+ 2.7**	**+ 2.4**	**+ 2.4**	**+ 2.6**

C. EFFECT OF INTRA-M WITH DIFFERENT λ

Fig.7 shows effect of Intra-M (Eq.9). Compared to MoCo-v3, λ>0 helps mitigate significantly the teacher-student difference, which corresponds to higher linear classification accuracy. The good trade-off is λ=1, which we used for all other experiments except other notices. *What might be the impact of gradually diminishing the importance of $\mathcal{L}_{Intra\text{-}M}$ in the overall optimization process?*

We tried decreasing weight λ from 1 to 0 (Eq.9) via a cosine decay along the training. And we observe the top-1 linear accuracy is 1% lower than the case keeping λ fixed.

D. EFFECT ON BATCH AND EMBEDDING SIZE

We show its behavior for varying batch size and dimensions of the projection output in Tab.7. Overall, Intra-M consistently improves MoCo-v3 in top-1 linear accuracy for all considered settings, demonstrating the important role of the proposed Intra-M in our framework Res-MoCo.

TABLE 8. Impact of Intra-M on ViT models, top-1 (%). Results are performed using CIFAR-100, trained 1000ep.

Arch.	Method	CIFAR-10			CIFAR-100		
		top-1	top-5	KNN-1	top-1	top-5	KNN-1
ViT-Tiny	MoCo-v3	76.21	98.64	76.13	49.07	77.07	45.66
	Res-MoCo	**78.89**	**98.81**	**78.24**	**50.76**	**78.43**	**47.72**
ViT-S	MoCo-v3	79.79	98.82	78.30	52.82	80.39	48.07
	Res-MoCo	**81.31**	**99.15**	**80.14**	**54.73**	**81.71**	**49.25**

TABLE 9. Different Backbone Architectures. Comparison of MoCo-v3 and Res-MoCo (ours) on ImageNet-100. All experiments are trained for 1000 epochs. With the presence of *intra-momentum*, models learn better quality representations.

Method	Inter-M	Intra-M	Arch.	Top-1	Δ_{top-1}	Top-5	Δ_{top-5}	KNN-1	Δ_{knn}
MoCo-v3	✓	-	ViT-S	79.66	-	95.40	-	76.64	-
Res-MoCo	✓	✓	ViT-S	**80.74**	+ 1.08	**95.78**	+ 0.38	**77.46**	+ 0.82
MoCo-v3	✓	-	ResNet-18	81.26	-	95.51	-	75.28	-
Res-MoCo	✓	✓	ResNet-18	**82.62**	+ 1.36	**95.82**	+ 0.31	**75.84**	+ 0.56
MoCo-v3	✓	-	ResNet-34	83.40	-	95.70	-	78.82	-
Res-MoCo	✓	✓	ResNet-34	**84.50**	+ 1.10	**96.06**	+ 0.36	**79.74**	+ 0.92
MoCo-v3	✓	-	ResNet-50	83.80	-	95.88	-	72.80	-
Res-MoCo	✓	✓	ResNet-50	**86.21**	+ 2.41	**96.82**	+ 0.94	**74.46**	+ 1.66

E. EFFECT ON DIFFERENT ARCHITECTURES

Here we verify the importance of reducing the intra-representation gap in SSL with various architectures, including CNNs (ResNet) and ViT models (Tab.8 and Tab. 9).

F. DIFFERENT TEMPERATURES IN CL LOSS

Temperature plays a crucial role in self-supervised contrastive learning [19], [56]. Tab.10 further remarks on the importance of the intra- momentum's presence. We find that with higher temperatures, our model performs much better than the model without intra-momentum in the KNN metric, where the baseline without the guide of the teacher's output itself with higher temperature, makes the teacher-student mismatch more serious. Note that KNN is designed to measure the similarity of a sample to its neighbors.

G. ON THE IMPORTANCE OF TEACHER-STUDENT SIMILARITY

We find that in EMA-based SSL such as MoCo-v3: "**High Similarity Outputs of the Teacher and Student Gives Better Accuracy.**"

To this end, we transform the representation gap from Eq.5 into cosine similarity for a more friendly view when comparing their performance on linear evaluation. To this end, we report Tab. 11 for ResNet-18 It can be seen that over the training, the more similar the representation model has, there better linear accuracy and KNN the model obtains.

This demonstrates Res-MoCo with the proposed *intra-momentum* highly encourages the student model to match the teacher's output, improving student learning capability.

H. IMPORTANCE OF INTRA MOMENTUM

We ablate the importance of the proposed *intra momentum* by injecting *intra momentum* into MoCo-v3 for every 8000 steps during training. We use CIFAR-100 and ResNet-18 to train the SSL model for 1000 epochs when all models fully converge. As shown in Fig. 8, whenever adding *intra-M* to the model MoCo-v3, the representation gap (measured by cosine similarity) between the *student* and *teacher* decreases (blue line) in Fig. 8 which strongly corresponds to the performance significantly boost to the level of Res-MoCo (green line).

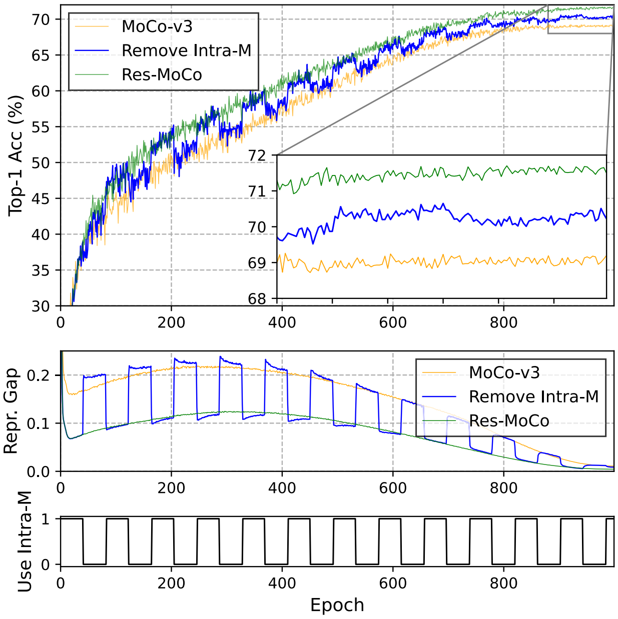


FIGURE 8. The use of Intra-M, use if value 1; otherwise, not use. We remove Intra-M loss for every 8000 steps (∼ 40 epochs). Experiments were done on CIFAR-100 for 1000 epochs for all models. Note that when removing Intra-M, Res-MoCo becomes the baseline.

By contrast, if removing *intra momentum*, we can observe that the accuracy of the student models quickly drops to the MoCo-v3 baseline orange line. This indicates the crucial role of the proposed *intra momentum* in the momentum-based SSL framework.

I. TRAINING TIME

Compared to the MoCo-v3 baseline, adding Intra-M loss makes the training time overhead of our Res-MoCo less than 1%, as shown in Tab.12, demonstrating the efficiency of the proposed method.

J. IRAG GAP WITH DIFFERENT MOMENTUM SPEEDS

In this part, we ablate the potential intra-representation gap (IRAG) between the student and teacher encoder in different momentum update speeds $\beta = \{0.9, 0.99, 0.996\}$. As shown in Fig.9, for 1000 epochs, $\beta = 0.996$ gives the best performance for MoCo-v3; however, it also shows the most significant repr. gap. Although the network architectures of the teacher and student are identical, the EMA updates the weight from the student to the teacher with a prolonged momentum speed, *i.e.* $\beta = 0.996$ [20], [27], [50]; therefore their weights are still quite different. In Eq.1, with $\beta = 0.996$, only 0.4% of the student's weight is counted to compute the new weight for the teacher.

Fig.9 shows a considerable gap in their produced representations of MoCo-v3, which we identify potentially prevents the SSL model from learning good representations. The

TABLE 10. Different temperatures. Comparison of MoCo-v3 and Res-MoCo on ImageNet-100. All methods are trained for 1000 epochs using the same settings run on 2 GPUs, batch size 256. It shows that Res-MoCo outperforms MoCo-v3 for both temperature settings.

Method	Inter-M	Intra-M	Arch.	Top-1	Δ_{top-1}	Top-5	Δ_{top-5}	KNN-1	Δ_{knn}
MoCo-v3 [9] $\tau = 0.2$	✓	-	ResNet-50	83.80	-	95.88	-	72.80	-
Res-MoCo (ours) $\tau = 0.2$	✓	✓	ResNet-50	**86.21**	+ 2.41	**96.82**	+ 0.94	**74.46**	+ 1.66
MoCo-v3 [9] $\tau = 1.0$	✓	-	ResNet-50	82.54	-	95.50	-	66.28	-
Res-MoCo $\tau = 1.0$	✓	✓	ResNet-50	**84.64**	+ 2.10	**96.28**	+ 0.78	**71.12**	+ 4.84

TABLE 11. Similarity of the student and teacher of each image itself on ImageNet-100, ResNet-18. During training, Res-MoCo shows a much higher similarity between teacher and student, which strongly corresponds to a performance boost in both linear top-1 (%) and KNN accuracy (%). We monitor similarity (sim) by taking the average cosine similarity of each training batch.

Method	Epoch	Inter-M	Intra-M	Sim (%)	Δ_{sim}	Top-1	Δ_{top-1}	KNN-1	Δ_{knn}
MoCo-v3 [9]	100	✓	-	89.65	-	54.62	-	54.98	-
Res-MoCo (ours)	100	✓	✓	**93.63**	+ 3.98	**60.21**	+ 5.59	**60.42**	+ 5.54
MoCo-v3 [9]	200	✓	-	88.89	-	60.38	-	59.06	-
Res-MoCo (ours)	200	✓	✓	**92.39**	+ 3.50	**64.08**	+ 3.70	**63.32**	+ 4.26
MoCo-v3 [9]	400	✓	-	89.63	-	63.35	-	60.64	-
Res-MoCo (ours)	400	✓	✓	**94.43**	+ 4.8	**69.98**	+ 6.63	**67.48**	+ 6.84
MoCo-v3 [9]	800	✓	-	96.72	-	78.14	-	72.67	-
Res-MoCo (ours)	800	✓	✓	**98.77**	+ 2.05	**80.41**	+ 2.27	**75.03**	+ 2.36
MoCo-v3 [9]	1000	✓	-	99.76	-	81.26	-	75.28	-
Res-MoCo (ours)	1000	✓	✓	**99.94**	+ 0.17	**82.62**	+ 1.36	**75.84**	+ 0.56

TABLE 12. Pretraining time. It is measured per epoch using the 4 GPUs machine (A6000), using the SSL library in [65].

Method	COCO	ImageNet-1K
MoCo-v3	$2'3''$	$18'40''$
Res-MoCo	$2'4''$	$18'45''$

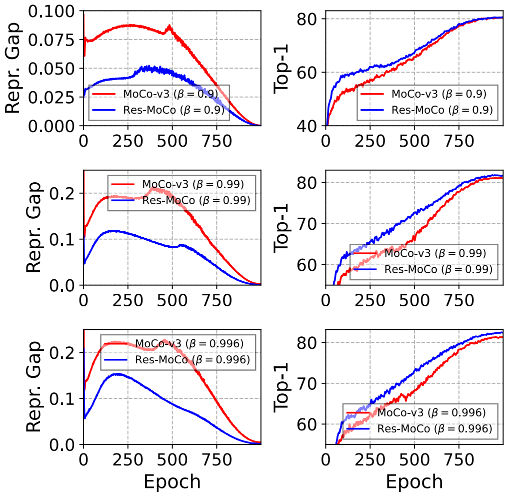


FIGURE 9. Representation gap between teacher and student encoders of MoCo-v3 and Res-MoCo with different momentum update speeds β on ImageNet-100.

repr. gap keeps increasing and only narrows at the end of training. At this point, with a nearly zero learning rate (cosine decay), the model converged, and the overall weight was insignificantly updated. The momentum coefficient also becomes 1 due to the cosine schedule (β increases from 0.996 to 1 [20], [25]) ($\beta = 1$ means that the student continues to update, but the teacher is kept unchanged). Applying *intra momentum* loss, this gap is significantly reduced and brings a performance boost for both values of β, and Res-MoCo achieves the highest performance gap compared to MoCo-v3 with $\beta = 0.996$.

Reducing the representation gap with *intra-momentum* makes an apparent effect of forcing the student model to learn to perform as closely as the teacher possible during training, hence improving both models together. Note that the teacher model is dynamically updated with EMA from the student; therefore, the better the student model learned, the better teacher models are updated, and vice versa.

K. VISUALIZATIONS

We analyze the pre-trained models with different aspects to understand how Res-MoCo consistently outperforms MoCo-v3 for most downstream tasks. First, we compare the feature maps learned by each method. Second, we use GradCAM heat map, a powerful tool in deep learning for model interpretation [74]. The ResNet-50 trained on IN-100 for 1000ep in Tab.9 is chosen for comparisons.

1) LEARNED FEATURE MAPS

Qualitative results of our Res-MoCo method compared to MoCo-v3 are shown in Fig. 10. The first four columns demonstrate that Res-MoCo produces cleaner feature maps by removing background noise, in contrast to MoCo-v3.

Discarding irrelevant information and focusing solely on the objects is crucial in achieving high-performance object recognition and detection [75], [76], [77], [78].

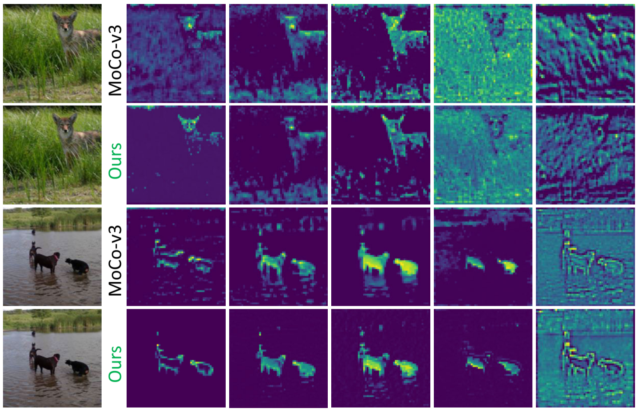


FIGURE 10. Features learned by MoCo-v3 [9] and Res-MoCo (ours) for images on the test set of IN-100. We visualize the five most meaningful feature maps in the last CONV layer in the ResNet-50's first block. It is best viewed in color.

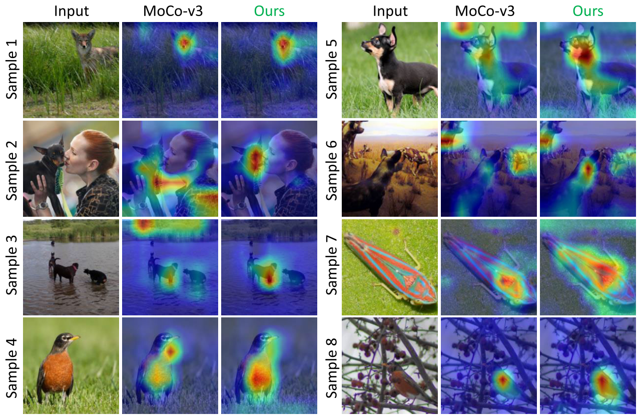


FIGURE 11. GradCAM generated by Res-MoCo (ours) and MoCo-v3, using images from the IN-100 test set, demonstrates that Res-MoCo produces a more accurate heat map.

2) GRADCAM ATTENTION

Fig. 11 demonstrates that our proposed Res-MoCo method generates more accurate object-focused attention heatmaps compared to MoCo-v3 for various examples. In the second image ("Sample 2"), for instance, MoCo-v3's attention heatmap highlights not only the dog but also other irrelevant parts, while Res-MoCo's heatmap precisely focuses on the dog's face. Fig.12 shows the evolution of each model training for 800 epochs where our models generated better heatmaps.

These visualizations provide compelling evidence of Res-MoCo's significant improvements over the strong baseline MoCo-v3 in object recognition and detection tasks.

VI. DISCUSSIONS ON INTRA-M

A. INTUITION OF INTRA-M

Chen et al. [9] found that EMA (Inter-M) improves MoCo-v3 performance by 2.2% (top-1 linear accuracy on IN-1K) compared to the version without it, despite EMA not being necessary to work. Our Intra-M, which also uses the EMA encoder, demonstrates similar performance improvements as Inter-M (see Tab.13). Both methods (case (b) and (c)) leverage EMA as a mean teacher model to generate more reliable targets (than case (a)) by ensembling past models [36], [79]. Given augmentations x_1 and x_2, Inter-M is trained to make the outputs of student f and teacher f_m closer $f(x_1) = f_m(x_2)$ on the different view. Intra-M trains the student encoder f to match the teacher encoder's reliable representations on the same view $f(x_1) = f_m(x_1)$, minimizing the gap and promoting better learning.

Combining them (case (d)), the student learns from the teacher's reliable representations of another crop and itself, resulting in improved performance. In EMA-based SSLs, the teacher is updated dynamically from the student, leading to a mutually beneficial relationship where both models improve (see Fig.2,5).

Compared to VICReg [68] and Barlow Twins (BT) [51]: BT prevents collapse by using a particular term as cross-correlation to regularize "redundancy" while VICReg regularizes "variance" and "covariance". Intra-M is designed not to prevent collapse but to regularize the teacher-student "representation gap" to improve EMA-based SSLs.

Compared to KD: The KD was initially proposed on *supervised learning* (SL) to minimize the TE-ST probability "distribution gap". Their teacher model often uses an offline pre-trained model instead of an EMA update as in SSL. By contrast, we investigate the "representation gap" in *self-supervised learning* (SSL) via the lens of Intra-M to narrow such a gap for improving SSLs, and we show that reducing the distribution gap technique of KD is useless in SSLs (see Table 6).

B. IMPACT OF INTER-M AND INTRA-M

1) CONTRASTIVE LEARNING WITH INTER-M

When removing the momentum encoder in Eq.2, the loss function for self-supervised contrastive learning becomes as follows:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(q{\cdot}k^{+}/\tau)}{\exp(q{\cdot}k^{+}/\tau) + \sum_{k^-} \exp(q{\cdot}k^{-}/\tau)}. \tag{11}$$

We emphasize that the key k_m (with subscript m) in Eq.3 comes from the EMA encoder, but in Eq.11, the key $k^{\{+/-\}}$ comes from the shared online encoder with *stop gradient* [50] (*i.e.* teacher ← student).

As shown in Tab.13 for IN-100, CL (w/o any momentum) yields 78.71% top-1, MoCo-v3 (CL+Inter-M) gives 81.26%. It shows that Inter-M helps to boost CL + 2.55% improvement on IN-100 and +2.82% on CIFAR-100.

2) CONTRASTIVE LEARNING WITH INTRA-M

We conduct an ablation study by replacing Inter-M with Intra-M in the loss function, which consists of the *intra-momentum* and *CL* terms as follows:

$$\mathcal{L}_{\text{CL-Intra\_M}} = \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{Intra-M}}. \tag{12}$$

Tab.13 on IN-100 shows that Intra-M improves CL to 81.66% (+2.95%), which is slightly better than using Inter-M (81.26%) (+2.55%). Combining Inter-M and Intra-M results in a boost to 82.62%. This behavior is also observed on

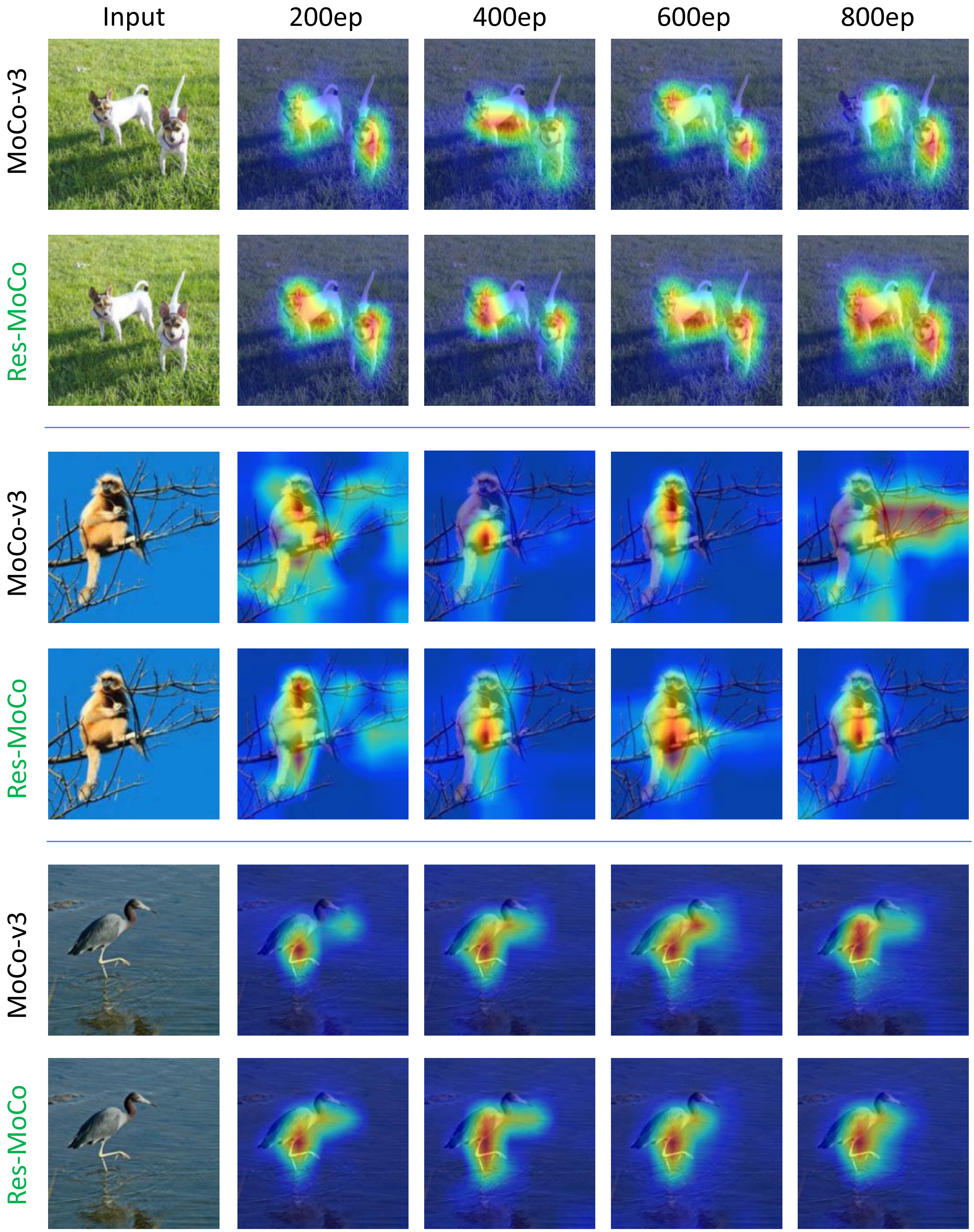


FIGURE 12. Comparison of GradCAM learned by MoCo-v3 and Res-MoCo (ours) for different epochs. The image samples are from the test set of ImageNet-100. The first row is the original image, the second row is the heat map by MoCo-v3, and the third row is the heat map produced by Res-MoCo (ours). It clearly shows that the heat map produced by Res-MoCo is much more accurate than that of MoCo-v3. It is best viewed in color.

TABLE 13. Impact of Inter-M and Intra-M in contrastive learning framework (baseline is MoCo-v3 with case b)). Performance on CIFAR-100/IN-100, pretrained 1000ep using ResNet-18 backbone.

No.	Case	CIFAR-100		IN-100	
		Top 1	KNN-1	Top 1	KNN-1
1	(a) CL	66.05	58.62	78.71	73.24
2	(b) CL + Inter-M	68.83	60.75	81.26	75.28
3	(c) CL + Intra-M	69.85	62.41	81.66	76.04
4	(d) CL + Inter-M + Intra-M	**71.65**	**64.27**	**82.62**	**76.14**

CIFAR-100, where Intra-M performs on par Inter-M, and their combination yields the best performance with 71.65%.

VII. CONCLUSION

This paper presents a simple yet nontrivial approach to address a problem in momentum-based SSL frameworks such as MoCo-v3 that is overlooked: the representation gap between the teacher and student models during training. Our investigation reveals that this often overlooked gap can significantly impede the models' ability to learn high-quality representations. To bridge this gap, we propose intra-momentum, which systematically reduces the representation gap by training the student model to closely match the teacher's output. We have shown that techniques like KD to reduce the distribution gap in *supervised learning* is not applicable to *self-supervised learning* and proposed reducing the representation gap instead. Exhaustive experiments on challenging image datasets demonstrate that our method significantly outperforms other CL baselines. Our findings underscore the importance of considering the representation gap between the teacher and student models in EMA-based SSL frameworks as MoCo and DenseCL.

VIII. DISCUSSION AND FUTURE WORKS

The proposed method significantly reduces the teacher-student gap to improve the SSL with a visible margin, yielding a complete and unified picture of traditional supervised learning with knowledge distillation. However, we have mainly conducted extensive experiments with CNN backbones, *i.e.* ResNet family. It is interesting to investigate our method for the large model of vision transformer ViT [63] or Swin Transformer [80]. The transformers-based models are powerful and have potential; however, it is very computationally costly, it is challenging with our current hardware resources. Some of our experiments with ViT-S (same size as ReNet-50) (Tab.8 and Tab.9) demonstrated that the proposed method works well for transformers. Until very recently, CNN has also been focused again on where CNN outperforms ViT as shown in [81]. We believe CNN-based SSL is still adequate to evaluate future SSLs.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Language Technol.*, 2019, pp. 1–16.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 1–9.

[5] T. X. Pham, J. W. Choi, R. J. L. Mina, T. X. Nguyen, S. R. Madjid, and C. D. Yoo, "LAD: A hybrid deep learning system for benign paroxysmal positional vertigo disorders diagnostic," *IEEE Access*, vol. 10, pp. 113995–114007, 2022.

[6] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo, "Modality shifting attention network for multi-modal video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10103–10112.

[7] A. Niu, Y. Zhu, C. Zhang, J. Sun, P. Wang, I. S. Kweon, and Y. Zhang, "MS2Net: Multi-scale and multi-stage feature fusion for blurred image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5137–5150, Aug. 2022.

[8] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "Image BERT pre-training with online tokenizer," in *Proc. ICLR*, 2022, pp. 1–29. [Online]. Available: https://openreview.net/forum?id=ydopy-e6Dg

[9] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.

[10] H. Zhang, G. Zhang, Y. Chen, and Y. Zheng, "Global relation-aware contrast learning for unsupervised person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8599–8610, Dec. 2022.

[11] C. Zhang, K. Zhang, C. Zhang, T. X. Pham, C. D. Yoo, and I. S. Kweon, "How does SimSiam avoid collapse without negative samples? A unified understanding with self-supervised contrastive learning," in *Proc. ICLR*, 2022, pp. 1–18.

[12] C. Zhang, K. Zhang, T. X. Pham, A. Niu, Z. Qiao, C. D. Yoo, and I. S. Kweon, "Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying MoCo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14421–14430.

[13] W. Zhao, C. Li, W. Zhang, L. Yang, P. Zhuang, L. Li, K. Fan, and H. Yang, "Embedding global contrastive and local location in self-supervised learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2275–2289, May 2023.

[14] Y. Wei, L. Yang, Y. Han, and Q. Hu, "Multi-source collaborative contrastive learning for decentralized domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2202–2216, May 2023.

[15] Q. Zhou, S. He, H. Liu, T. Chen, and J. Chen, "Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2176–2189, May 2023.

[16] Y. Liu, H. Ge, L. Sun, and Y. Hou, "Complementary attention-driven contrastive learning with hard-sample exploring for unsupervised domain adaptive person Re-ID," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 326–341, Jan. 2023.

[17] R. Guo, J. Sun, C. Zhang, and X. Qian, "A contrastive graph convolutional network for toe-tapping assessment in Parkinson's disease," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8864–8874, Dec. 2022.

[18] P. Bachman, R. Devon Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 2019, *arXiv:1906.00910*.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, and M. G. Azar, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. NIPS*, 2020, pp. 21271–21284.

[21] M. Seyfi, A. Banitalebi-Dehkordi, and Y. Zhang, "Extending momentum contrast with cross similarity consistency regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6714–6727, Oct. 2022.

[22] Y. Zhu, H. Shuai, G. Liu, and Q. Liu, ‘‘Self-supervised video representation learning using improved instance-wise contrastive learning and deep clustering,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6741–6752, Oct. 2022.

[23] B. Fang, X. Li, G. Han, and J. He, ‘‘Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning,’’ *IEEE Access*, vol. 11, pp. 45547–45558, 2023.

[24] H. Yang, X. Li, and W. Pedrycz, ‘‘Learning from crowds with contrastive representation,’’ *IEEE Access*, vol. 11, pp. 40182–40191, 2023.

[25] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, ‘‘Momentum contrast for unsupervised visual representation learning,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[26] X. Chen, H. Fan, R. Girshick, and K. He, ‘‘Improved baselines with momentum contrastive learning,’’ 2020, *arXiv:2003.04297*.

[27] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, ‘‘Emerging properties in self-supervised vision transformers,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[28] M. Zheng, S. You, F. Wang, C. Qian, C. Zhang, X. Wang, and C. Xu, ‘‘ReSSL: Relational self-supervised learning with weak augmentation,’’ in *Proc. NIPS*, 2021, pp. 2543–2555.

[29] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, ‘‘Dense contrastive learning for self-supervised visual pre-training,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3023–3032.

[30] Y. Zhu and Y. Wang, ‘‘Student customized knowledge distillation: Bridging the gap between student and teacher,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5037–5046.

[31] J. Guo. (2022). *Reducing the Teacher–Student Gap Via Adaptive Temperatures*. [Online]. Available: https://openreview.net/forum?id=h-z_zqT2yJU

[32] T. Su, Q. Liang, J. Zhang, Z. Yu, Z. Xu, G. Wang, and X. Liu, ‘‘Deep cross-layer collaborative learning network for online knowledge distillation,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2075–2087, May 2023.

[33] K. Zhang, C. Zhang, S. Li, D. Zeng, and S. Ge, ‘‘Student network learning via evolutionary knowledge distillation,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2251–2263, Apr. 2022.

[34] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and J. C. SanMiguel, ‘‘Attention-based knowledge distillation in scene recognition: The impact of a DCT-driven loss,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4769–4783, Sep. 2023.

[35] X. Cui, C. Wang, D. Ren, Y. Chen, and P. Zhu, ‘‘Semi-supervised image deraining using knowledge distillation,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8327–8341, Dec. 2022.

[36] A. Tarvainen and H. Valpola, ‘‘Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,’’ in *Proc. NIPS*, 2017, pp. 1–10.

[37] M. J. Pring, *Technical Analysis Explained: The Successful Investor’s Guide to Spotting Investment Trends and Turning Points*. New York, NY, USA: McGraw-Hill, 2002.

[38] F. Klinker, ‘‘Exponential moving average versus moving exponential average,’’ *Mathematische Semesterberichte*, vol. 58, no. 1, pp. 97–107, Apr. 2011.

[39] D. P. Kingma and J. Ba, ‘‘Adam: A method for stochastic optimization,’’ in *Proc. ICLR*, 2015, pp. 1–15.

[40] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, ‘‘On the importance of initialization and momentum in deep learning,’’ in *Proc. ICML*, 2013, pp. 1139–1147.

[41] J. Ma and D. Yarats, ‘‘Quasi-hyperbolic momentum and Adam for deep learning,’’ in *Proc. ICLR*, 2018, pp. 1–38.

[42] N. Vieillard, B. Scherrer, O. Pietquin, and M. Geist, ‘‘Momentum in reinforcement learning,’’ in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2529–2538.

[43] E. Korkmaz, ‘‘Nesterov momentum adversarial perturbations in the deep reinforcement learning domain,’’ in *Proc. ICML*, 2020, pp. 1–6.

[44] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, ‘‘Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,’’ in *Proc. ICML*, 2018, pp. 1861–1870.

[45] Y. Tian, D. Krishnan, and P. Isola, ‘‘Contrastive representation distillation,’’ in *Proc. ICLR*, 2019, pp. 1–19.

[46] K. Lee, ‘‘Prototypical contrastive predictive coding,’’ in *Proc. ICLR*, 2022, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=8la28hZOwug

[47] Z. Cai, A. Ravichandran, S. Maji, C. Fowlkes, Z. Tu, and S. Soatto, ‘‘Exponential moving average normalization for self-supervised and semi-supervised learning,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 194–203.

[48] Z. Li, S. Liu, and J. Sun, ‘‘Momentum^2 teacher: Momentum teacher with momentum statistics for self-supervised learning,’’ 2021, *arXiv:2101.07525*.

[49] H. Cheng, H. Li, Q. Wu, H. Qiu, X. Zhang, F. Meng, and T. Zhao, ‘‘Disturbed augmentation invariance for unsupervised visual representation learning,’’ *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 3, 2023, doi: 10.1109/TCSVT.2023.3272741.

[50] X. Chen and K. He, ‘‘Exploring simple Siamese representation learning,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[51] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, ‘‘Barlow twins: Self-supervised learning via redundancy reduction,’’ in *Proc. ICML*, 2021, pp. 12310–12320.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, and S. Chintala, ‘‘PyTorch: An imperative style, high-performance deep learning library,’’ in *Proc. NIPS*, 2019, pp. 1–12.

[53] G. Hinton, O. Vinyals, and J. Dean, ‘‘Distilling the knowledge in a neural network,’’ 2015, *arXiv:1503.02531*.

[54] J. Guo, M. Chen, Y. Hu, C. Zhu, X. He, and D. Cai, ‘‘Reducing the teacher–student gap via spherical knowledge disitllation,’’ 2020, *arXiv:2010.07485*.

[55] T.-B. Xu and C.-L. Liu, ‘‘Data-distortion guided self-distillation for deep neural networks,’’ in *Proc. AAAI*, 2019, pp. 5565–5572.

[56] T. Wang and P. Isola, ‘‘Understanding contrastive representation learning through alignment and uniformity on the hypersphere,’’ in *Proc. ICML*, 2020, pp. 9929–9939.

[57] A. van den Oord, Y. Li, and O. Vinyals, ‘‘Representation learning with contrastive predictive coding,’’ 2018, *arXiv:1807.03748*.

[58] A. Krizhevsky, V. Nair, and G. Hinton. (2010). *CIFAR-10 (Canadian Institute for Advanced Research)*. [Online]. Available: http://www.cs.toronto.edu/kriz/cifar.html

[59] Y. Tian, D. Krishnan, and P. Isola, ‘‘Contrastive multiview coding,’’ in *Proc. ECCV*, 2020, pp. 776–794.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in *Proc. NIPS*, 2012, pp. 1–9.

[61] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ‘‘The PASCAL visual object classes (VOC) challenge,’’ in *Proc. IJCV*, 2010, pp. 303–338.

[62] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ‘‘Microsoft COCO: Common objects in context,’’ in *Proc. ECCV*, 2014, pp. 740–755.

[63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16×16 words: Transformers for image recognition at scale,’’ in *Proc. ICLR*, 2021, pp. 1–22.

[64] M Contributors. (2021). *MMSelfSup: Openmmlab Self-Supervised Learning Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmselfsup

[65] V. G. T. da Costa, E. Fini, M. Nabi, N. Sebe, and E. Ricci, ‘‘SOLO-Learn: A library of self-supervised methods for visual representation learning,’’ *J. Mach. Learn. Res.*, vol. 23, no. 56, pp. 1–6, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-1155.html

[66] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, ‘‘Whitening for self-supervised representation learning,’’ in *Proc. ICML*, 2021, pp. 3015–3024.

[67] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, ‘‘Unsupervised learning of visual features by contrasting cluster assignments,’’ in *Proc. NIPS*, 2020, pp. 9912–9924.

[68] A. Bardes, J. Ponce, and Y. LeCun, ‘‘VICReg: Variance-invariance-covariance regularization for self-supervised learning,’’ in *Proc. ICLR*, 2022, pp. 1–23.

[69] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, ‘‘With a little help from my friends: Nearest-neighbor contrastive learning of visual representations,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9568–9577.

[70] T. Nguyen, T. X. Pham, C. Zhang, T. M. Luu, T. Vu, and C. D. Yoo, ‘‘DimCL: Dimensional contrastive learning for improving self-supervised learning,’’ *IEEE Access*, vol. 11, pp. 21534–21545, 2023.

[71] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *Proc. ECCV* Cham, Switzerland: Springer, 2022, pp. 668–684.

[72] X. Zhu and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[73] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[75] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10595–10604.

[76] C. Zhang, T.-Y. Pan, Y. Li, H. Hu, D. Xuan, S. Changpinyo, B. Gong, and W.-L. Chao, "MosaicOS: A simple and effective use of object-centric images for long-tailed object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 407–417.

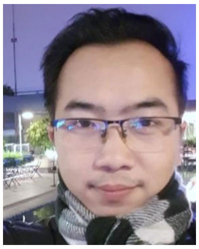
[77] L. A. Varga and A. Zell, "Tackling the background bias in sparse object detection via cropped windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2768–2777.

[78] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5794–5803.

[79] J. Huang and W. Zhou, "Re^{2}EMA: Regularized and reinitialized exponential moving average for target model update in object tracking," in *Proc. AAAI*, 2019, pp. 8457–8464.

[80] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[81] K. Tian, Y. Jiang, C. Lin, L. Wang, and Z. Yuan, "Designing BERT for convolutional networks: Sparse and hierarchical masked modeling," in *Proc. ICLR*, 2023, pp. 1–16.



TRUNG XUAN PHAM (Student Member, IEEE) received the B.S. degree from the School of Electronics and Telecommunications, Hanoi University of Science and Technology (HUST), in 2014. He is currently pursuing the Ph.D. degree with KAIST, under the supervision of Prof. Chang D. Yoo. His doctoral research interests include speech processing, self-supervised learning, and computer vision.

AXI NIU (Student Member, IEEE) received the B.S. and M.S. degrees from Henan University, Kaifeng, China, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. Her research interests include image processing and computer vision.

KANG ZHANG (Student Member, IEEE) received the B.S. degree from the Harbin Institute of Technology, in 2020. He is currently pursuing the Ph.D. degree with the Korea Advanced Institute of Science and Technology. His research interests include deep learning, self-supervised learning, and adversarial machine learning.

TEE JOSHUA TIAN JIN (Student Member, IEEE) received the B.S. degree from KAIST, in 2022, where he is currently pursuing the master's degree. His research interest includes deep learning.

JI WOO HONG (Member, IEEE) received the B.S. degree in mechanical engineering from Michigan State University, in 2019, and the M.S. degree in robotics program from the Korea Advanced Institute of Science and Technology, in 2022, where he is currently pursuing the Ph.D. degree. His research interests include 3D human pose and shape estimation and visual-language reasoning.

CHANG D. YOO (Senior Member, IEEE) received the B.S. degree in engineering and applied science from the California Institute of Technology, the M.S. degree in electrical engineering from Cornell University, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology. From January 1997 to March 1999, he was Senior Researcher at Korea Telecom (KT). Since 1999, he has been a Faculty Member with the Korea Advance Institute of Science and Technology (KAIST), where he is currently a tenured Full Professor with the School of Electrical Engineering and an Adjunct Professor with the Department of Computer Science. He also served as the Dean of the Office of Special Projects and the Office of International Relations. He is the Director of the Video Turing Test Research Center and the AI Fairness Research Center. His current research interests include machine learning, signal processing, computer vision, and audio processing. He is a member of Tau Beta Pi and Sigma Xi.