Temporal Knowledge Consistency for Unsupervised Visual Representation Learning

Weixin Feng[1][*]   Yuanjiang Wang[2][*][†]   Lihua Ma[2]   Ye Yuan[2]   Chi Zhang[2]

Beijing University of Posts and Telecommunications [1]   Megvii Technology[2]

fengweixin@bupt.edu.cn, wangyuanjiang@megvii.com

{malihua, yuanye, zhangchi}@megvii.com

Abstract

The instance discrimination paradigm has become dominant in unsupervised learning. It always adopts a teacher-student framework, in which the teacher provides embedded knowledge as a supervision signal for the student. The student learns meaningful representations by enforcing instance spatial consistency with the views from the teacher. However, the outputs of the teacher can vary dramatically on the same instance during different training stages, introducing unexpected noise and leading to catastrophic forgetting caused by inconsistent objectives. In this paper, we first integrate instance temporal consistency into current instance discrimination paradigms, and propose a novel and strong algorithm named Temporal Knowledge Consistency (TKC). Specifically, our TKC dynamically ensembles the knowledge of temporal teachers and adaptively selects useful information according to its importance to learning instance temporal consistency. Experimental result shows that TKC can learn better visual representations on both ResNet and AlexNet on linear evaluation protocol while transfer well to downstream tasks. All experiments suggest the good effectiveness and generalization of our method. Code will be made available.

1. Introduction

The rise of Deep Convolutional Neural Networks (DCNN) [23, 28, 46] has led to significant success in computer vision benchmarks [8, 13, 33]. The excellent performance of supervised DCNN always relies on a large quantity of manually labeled data, which is costly to collect [18, 51]. Unsupervised representation learning has been attracted more and more interest, for it can learn a good
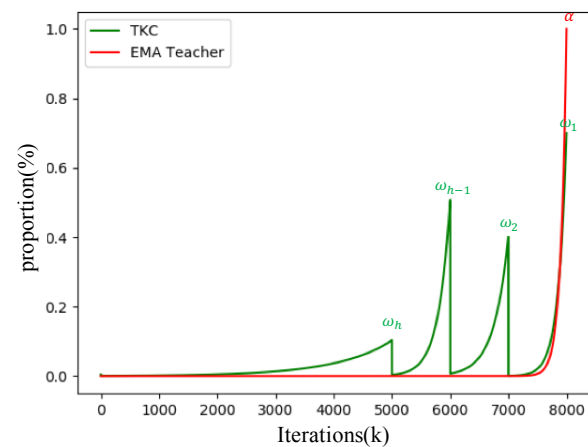

Figure 1. Mainstream unsupervised methods adopt the teacher-student framework, where the teacher is an EMA ensemble of previous student encoders. This figure illustrates the proportion of previous students in the teacher with respect to training steps. The red curve shows that the EMA teacher ensembles the previous encoders by a predesigned factor α, where only alomst encoders in the very close steps are ensembled. Our TKC (the green curve) reuses the early models and adaptively learns the importance ω for each of them, thus leads to temporal consistent representations.

representation without human annotations. These methods are generally to manually design a pretext task to learn representations, such as image in-painting [42], colorization [9, 60, 30, 31], rotate predicting [18, 6, 14] and clustering [2, 62, 4]. All these pretext tasks are based on specific domain knowledge, which has poor generation on various downstream tasks. Recently, instance discrimination [51, 21, 5, 19, 36] paradigm has led to remarkable progress in unsupervised representation learning and even surpasses the supervised pre-training on extensive downstream tasks [36, 21].

The instance discrimination paradigm treats each sample itself as its own category and trains the CNN to sep-

---

arate all the different samples from each other. The current paradigm can be formulated as a teacher-student framework enforcing the instance spatial consistency of two networks, which are the student network and the EMA teacher network [7, 19, 5]. The instance spatial consistency constrains the similarity of different spatial views from the same instance, and its ultimate goal is to learn instance-discriminative and spatial-invariant representations. One of the key points in these instance discrimination works is the EMA teacher. For instance, MoCo [21] uses the EMA teacher to output consistent negative samples for the student; BYOL [19] trains a student to mimic the representations from the EMA teacher; SimCLR [5] maintains a real-time EMA teacher of the student.

However, we argue that the current EMA teacher is suboptimal as illustrated in Fig. 1: (1) the EMA teacher only ensembles the rare knowledge of recent encoders by a handcraft proportion, which means that it only concentrates on instance spatial consistency while the instance temporal consistency is ignored. As a consequence, the outputs of the same sample can vary dramatically among different training stages, which can introduce unexpected noise and finally lead to catastrophic forgetting [34, 61]. (2) The EMA manner can't leverage the importance of different encoders. It assumes that the outputs of later models are largely more important than the earlier ones, despite that the benefits of previous epochs have been observed in previous works [29, 61].

In this paper, we integrate instance temporal consistency into the instance discrimination paradigm and propose a novel and strong algorithm, namely Temporal Knowledge Consistency(TKC), which contains the *temporal teacher* and the *knowledge transformer*. Specifically, *temporal teacher* supplies instance temporal consistency via introducing the temporal knowledge from previous models. And the *knowledge transformer* dynamically learns the importance of different temporal teachers, then adaptively ensembles the useful information according to their importance, to generate instance temporal consistency objective. In addition, we provide a computation-economical implementation, which can provide temporal knowledge without preserving multiple previous models.

Our experimental results on different tasks and benchmarks have demonstrated that TKC can learn a better visual representation with excellent transferability and scalability. Concretely, we achieve state-of-the-art performance on ResNet and AlexNet backbones on linear evaluation protocol. Moreover, we evaluate representations learned by TKC on many downstream tasks and architectures. All results suggest the effectiveness of TKC. Overall, the main contributions in this work include:

- We are the first to integrate instance temporal consistency into the current EMA teacher in the instance dis-
crimination paradigm.

- We propose a novel and strong algorithm, named Temporal Knowledge Consistency (TKC), which can dynamically ensemble the knowledge from different temporal teachers.

- Extensive experiments are conducted on several benchmarks and architectures, which shows the superior performance on mainstream benchmarks and the scalability of TKC.

2. Related Works

Unsupervised Pretext Tasks. Unsupervised representation learning aims to learn meaningful representations from large amounts of data samples via constructing a wide range of pretext tasks without human labels. These pretext tasks usually vary in different forms. Among them, one family of these typical pretext works are generative-based which rely on auto-encoder [44] or GAN [35, 12], such as colorization [9, 60, 30, 31] and image in-painting [42]. And the others are discriminative-based, like predicting rotation or augmentation [18, 6, 14] of the image, solving jigsaw puzzles [38], locating relative patch [10, 11], ordering video frames [15, 54, 50], matching corresponding audio [41, 17, 40, 27], and clustering [2, 3, 62, 58, 1, 52, 56, 55, 16]. All the pretext methods are based on specific domain knowledge, fail to generalize to different downstream tasks. Recent progress in unsupervised representation learning mainly benefits from instance discrimination and attracts widespread attention from researchers.

Instance Discrimination. Instance discrimination methods [51, 21, 7, 5, 19, 39, 14, 36] have dominated the unsupervised learning field in the few years, which treat each sample itself as its own category and train the CNN to separate all the different samples from each other. This paradigm commonly includes a teacher model to provide a supervised signal, and a student model to learn the embedded knowledge from the former. Wu *et al*. [51] is the first to propose instance discrimination in unsupervised learning, which regards the student model in the last epoch as the teacher model. It learns meaningful representations by means of the classic InfoNCE loss [20, 39] and the target generated by the teacher. MoCo [21, 7] takes the EMA ensemble of the student as the teacher model to provide consistent and robust objectives, and brings a breakthrough by solving the knowledge out-of-date problem with the help of the EMA teacher. It also maintains a queue of negative samples and keeps them fresh. SimCLR [5] builds symmetrical architecture between the student and the teacher, while uses stronger data augmentation to enforce network to learn instance spatial consistency. BYOL [19] also implements the teacher with the EMA ensemble of the student, and makes use of L2 loss to pull the embedding features of positives pairs while

removing explicitly negative samples. Our TKC explicitly integrates instance temporal consistency into the instance discrimination paradigm, making the targets generated by the teacher more accurate and stable.

Temporal Knowledge. Temporal knowledge is widely used in both semi-supervised learning and optimization. In the field of semi-supervised learning, plenty of the proposed works adopt the EMA ensemble to take advantage of the knowledge of the previous training stage, to learn the time-consistent representations. Temporal Ensemble [29] ensembles the output of different epochs to yield better predictions. Mean teacher [47] instead ensembles the previous student as a teacher to prevent incorrect target and outweighs the cost of misclassification. Tian *et al.* [61] points out the catastrophic forgetting problems in semi-supervised, and solves it by measuring the time consistency of samples and filtering the inconsistent ones. In the field of optimization, temporal knowledge is integrated by different advanced optimization strategies during training. SGD only uses the gradient computed by a mini-batch to backpropagate, which is noisy and inaccurate. Momentum [43] and NAG [37] instead use the gradient in a mini-batch by the momentum of the gradient to accelerate the convergence of model training and suppress shocks. Adam [26] is another momentum updating strategy which further introduces the second momentum to leverage different channels. All these works use temporal knowledge to reduce noise and accelerate convergence.

3. Method

In this section, we first point out the limitation of the current EMA teacher in Sec 3.1. Secondly, we propose *temporal teacher* to improve it in Sec 3.2. Thirdly, we introduce the *knowledge transformer* to dynamically leverage the importance of different models in Sec 3.3. Then we propose a *temporal loss* to learn instance temporal consistency in Sec 3.4. At last, we describe our overall framework and the algorithm in Sec 3.5.

3.1. Limitation of EMA teacher

Instance discrimination paradigm always involves two encoders, the teacher encoder T and the student encoder S. For a training sample x, the augmentation from augmentation distribution $\mathcal{T}$ is applied twice to obtain two augmented sample x_0, x_n. The teacher output $r_n^T = T(x_n)$ as the target to provide instance spatial knowledge. The student network takes the other sample x_0 then outputs $r_0^S = T(x_0)$, and learn knowledge by constraining its similarity with r_n^T. In this teacher-student framework, the teacher encoder has the same architecture with the student, and its parameters are updated by an exponential moving average (EMA) of the models:

$$T^{n+1} = \alpha T^n + (1-\alpha)S^n \qquad (1)$$

where n is the training step, α is to control the updating speed of the teacher. We name the teacher as EMA teacher. In current training step $n+1$, the teacher is ensembled by the last teacher T^n with ratio α, and the last student S^n with ratio $1-\alpha$. The last teacher T^n is also an ensemble of previous students. In order to explore the temporal knowledge in EMA teacher, we expand T^n in Eq 1 as following:

$$
\begin{aligned}
T^{n+1} &= (1-\alpha) \cdot \sum_{m=0}^{n-1}(\alpha^m S^{n-m}) + \alpha^n \cdot T_1 \\
&\approx (1-\alpha) \cdot [S^n + \alpha S^{n-1} + ... + \alpha^n S^0]
\end{aligned}
\qquad (2)
$$

where S^m means the student model at step m, α is the updating factor. In Eq 2, we can find out that current teacher T^{n+1} is an ensemble of a sequence of student S from step 0 to step n.

However, we note that the EMA teacher can only preserve the knowledge from the latest encoders. On the one hand, as m goes to infinity $m \to \infty$, the weight of student S^{n-m} approaches 0, for α is lower than 1. When training MoCo [21] on ImageNet, only student models within an epoch can provide the knowledge, as illustrated in Fig. 1. This knowledge from only near steps is insufficient, which can cause the dramatically changes among different training stages and prevent the student to learn instance temporal consistency.

On the other hand, the strategy of EMA is also too simple. It assumes that the importance of earlier models is decreased exponentially with time, even though the earlier models can provide useful information to mitigate the catastrophic forgetting. In a summary, these two flaws prevent the instance discrimination paradigm from making full use of temporal knowledge and learning instance temporal consistency.

3.2. Temporal Teachers

EMA teacher in Eq.2 only attaches importance to recent models. However, the output of these models is smooth and similar due to the low learning rate and momentum optimizer. As a consequence, they fail to supply instance temporal consistency to lighten the dramatic changes of the models, which can easily lead to training failure and catastrophic forgetting. We claim that jointly utilizing the knowledge from previous models can provide a more consistent and robust target. To achieve that, we propose to take out the previous models, which have few proportions in the EMA teacher, to build our ***temporal teacher***. Then we make full use of them to alleviate catastrophic forgetting and learn instance temporal consistency in the instance discrimination paradigm.

We explicitly preserve a group of previous teachers as *temporal teacher* to reuse the knowledge from previous encoders. To formulate our proposal, we use T_n to denote
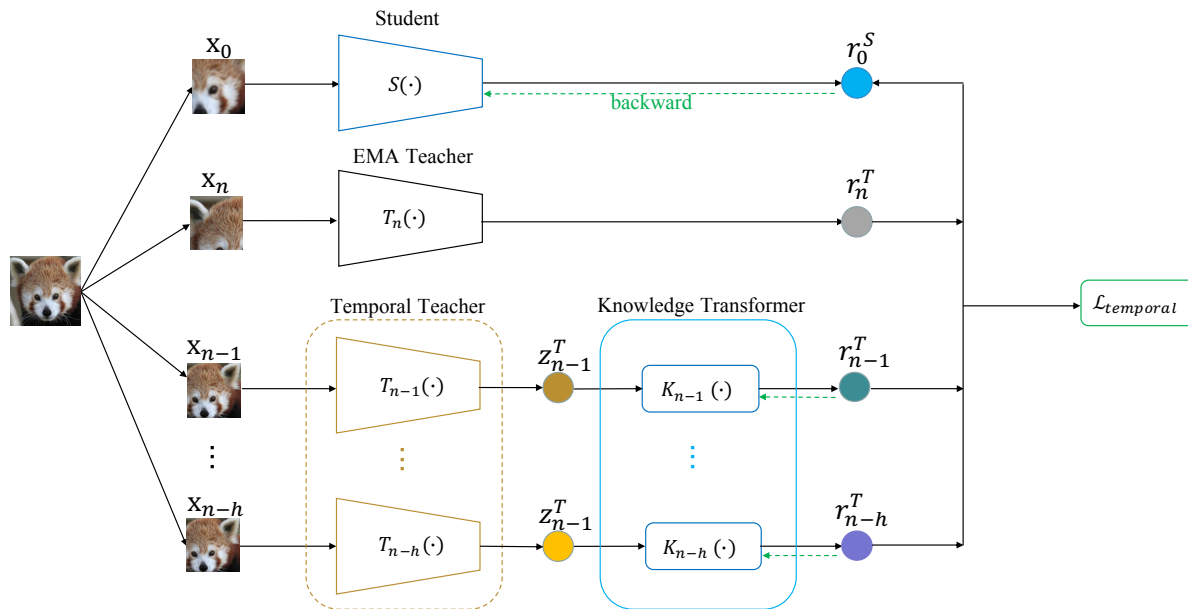
Figure 2. **The overall framework of our TKC.** For each training image x, TKC generate a target from the EMA teacher, and h targets from the *temporal teachers*. *Temporal teachers* is a set of encoders from previous training stages. The *knowledge transformer* is appended behind the *temporal teacher* to dynamically leverage their importance. Every teacher in TKC frameworks can provide a supervised signal, which then feeds into the *temporal loss* and backward to update the student and the *knowledge transformer*. The green dotted line means backpropagate.

current EMA teacher, and $\{T_{n-1}, T_{n-2}, ...\}$ to represent the temporal teachers. The lower subscript means earlier. Each of these teachers is saved for each s training step, including the knowledge mainly from this training step. Note that the subscript n means different with the superscript in Eq.1. The distance between T_{j-1} and T_{j-2} is s training steps.

The teacher far away from now is too out-of-date, whose knowledge can be inconsistent and noisy for the current teacher. Hence we only preserve adjacent teachers as *temporal teacher*, while throwing away the previous one. We use h to represent the number of temporal teachers, and denote the *temporal teachers* by $\{T_{n-1}, T_{n-2}, ..., T_{n-h}\}$.

We illustrate the *temporal teacher* in the brown dotted box in Fig.2. For a sample x in the training set, we apply h times data augmentations from the augmentation distribution $\mathcal{T}$, to obtain $x_j, j \in [n-h, n-1]$. The *temporal teachers* are set stop-gradient, and take the augmented views as input to yield representations $z_j^T = T_j(x_j), j \in [n-h, n-1]$ as the temporal predicting target. The subscript j of z_j^T indicate that the target is corresponding to the teacher T_j.

In the implementation, we propose a more efficient way to achieve *temporal teacher*. Instead of getting the target from previous teachers, we preserve the representations of all the training data in the previous h stages in a memory named **history bank**. For each training sample x, we can get z_j^T from the *history bank* instead of from the teacher

T_j. *History bank* is an approximate implementation of the *temporal teachers*, for both of them can provide temporal knowledge. In this way, the computational cost is largely reduced and the additional GPU memory allocation is negligible. The detail can be seen in the supplementary material.

3.3. Knowledge Transformer

In EMA teacher, the weights of different ensemble models are decreased over time exponentially. However, the importance of different models may not be in line with the EMA rule. In this section, we propose to dynamically predict the importance of the *temporal teachers*'s knowledge by **knowledge transformer**.

The *knowledge transformer* is illustrated in the blue box in Fig.2. It takes the rough target $z_j^T, j \in [n-1, n-h]$ from the teacher T_j as input, and then transfer the knowledge of them to leverage their importance. The formulation is as follows:

$$r_j^T = K_j(z_j^T) \tag{3}$$

where r_j^T denotes the target after leveraging the importance, which has thrown the harmful information from it and only preserves temporal consistent knowledge. This strategy can adaptively learn and adjust the importance of the *temporal teachers* in the early or later encoders, which is better than coupled it to the handcraft proportion in the EMA teacher.

In the implementation, we use an MLP with one hidden layer to transfer the knowledge for each temporal teacher. During training, the *knowledge transformer* is training simultaneously with the student.

3.4. Temporal Loss

Algorithm 1 Temporal knowledge consistency

---

input: $S(\cdot), T_n(\cdot), K(\cdot)$
hyperparameters: α, h, s
 1: **for** each sample x **do**
 2:     draw h+2 augmentations
 3:     # *the original models*
 4:     $r_0^S = S(x_0)$
 5:     $r_n^T = T_n(x_n)$
 6:     **for all** $j \in \{n-1, n-h\}$ **do**
 7:         # *temporal teacher*
 8:         $z_j^T = T_j(x_j)$
 9:         # *knowledge transformer*
10:         $r_j^T = K_j(z_j^T)$
11:     **end for**
12:     # *temporal loss*
13:     compute the loss in Eq. 4
14:     backward to update S and K
15:     update the T_n by Eq. 1
16: **end for**
17: **return** $S(\cdot)$

Different from previous works that only maximize the mutual information (MI) of the student output r_0^S and the target r_n^T from the EMA teacher, we propose to combine maximal the MI between r_0^S and each $r_j^T, j \in [n-h, n-1]$. This is in the intuitive that we hope the student can synchronously learn instance temporal consistency from temporal knowledge. Our objective is as follows:

$$\mathcal{L}_{tem} = \max_{r_0^S}(I(r_0^S; r_n^T) + \sum_{j=n-h}^{n-1} I(r_0^S; r_j^T)) \quad (4)$$

The first term maximizes the MI in the current phase, like previous works do [39, 24, 21, 19], which can only learn spatial consistent representations between different views. The second term maximizes the MI with previous knowledge to encourage the temporal consistency representations between different training stages, to mitigate the oscillation and catastrophic forgetting. Because the mutual information is notoriously hard to estimate, we instead maximizing the lower bound of it by the InfoNCE [39, 21, 5]:

$$\mathcal{L}_{tem}^{(1)} = \sum_{j=n-h}^{n} -\log \frac{sim(r_0^S \cdot r_j^T)}{sim(r_0^S \cdot r_j^T) + \sum_{r_j^-} sim(r_0^S \cdot r_j^-)} \quad (5)$$

where r_j^- presents the representation of other samples from the same teacher T_j, and $sim(r_0^S \cdot r_j^T)$ means their cosine

similarity as following:

$$sim(r_0^S \cdot r_j^T) = exp(r_0^S \cdot r_j^T / \tau) \quad (6)$$

where τ is temperature coefficient. In Eq. 5, the term $j = n$ estimates the MI with the current target, the other terms estimate the MI with temporal targets. InfoNCE is relied on the negative samples to estimate the probability distributions. Furthermore, our methods can also work on the methods without negative samples like BYOL [19]. We minimize the L2 distance to maximize the MI for these works:

$$\mathcal{L}_{tem}^{(2)} = \sum_{j=n-h}^{n} ||r_0^S - r_j^T||^2 \quad (7)$$

3.5. Overall Framework

As previous works do, TKC also introduce a student $S(\cdot)$ amd an EMA teacher $T_n(\cdot)$. For a training sample x from the data distribute, we obtain r_0^S from S and r_n^T from T_n. To learn consistent knowledge, we also get targets from the *temporal teachers* as $z_j^T, j \in [n-h, n-1]$. These targets should transport to the *knowledge transformer* to filter important knowledge as $r_j^T, j \in [n-h, n-1]$. Then all the representations are fed into the *temporal loss* in Eq. 4. During training, all the teachers are set stop-gradient. The loss will be back-propagated to update the student S and the *knowledge transformer* $K_j, j \in [n-h, n-1]$. Algorithm 1 summarizes the algorithmic flow of the TKC procedure.

4. Experiments

In this section, we evaluate the quality of feature representation learned by our proposed TKC on several unsupervised benchmarks. We first follow standard linear evaluation protocol to assess the learned representations on ImageNet [8]. Then we transfer the pre-trained features to different downstream tasks, including object detection, instance segmentation, and semi-supervised classification. Finally, we perform a set of analysis studies to give an intuition of its performance. For brief-expression, all the experiments are based on MoCo v2 [7] framework and ResNet-50 [23] backbone unless otherwise stated.

4.1. Evaluation on Linear Classification

We implement our TKC based on MoCo v2, which is composed of a standard ResNet-50 [23] backbone and an MLP layer in the teacher-student framework. And the number of temporal teachers h is set to 2. We train TKC model on 8 NVidia-1080ti GPUs with a mini-batch size of 256 and set α as 0.999, τ as 0.2. Moreover, we set the base learning rate lr as 0.3, weight decay as 0.0001, and introduce a warm-up stage in the first 10 epochs, where linearly increase the learning rate from 0.01 to 0.03. All other hyperparameters, training settings on pretext task and linear eval-

Method	architecture	epochs	Top-1	Top-5
Random	-	200	5.6	-
Supervised	-	200	75.5	-
200 epoch training				
LA [62]	R50	200	60.2	-
CMC [48]	R50(2x)	200	64.4	88.2
CPC v2 [24]	R50	200	63.8	85.3
MoCo [21]	R50	200	60.6	-
MoCHi [25]	R50	200	68.0	
CO2 [49]	R50	200	68.0	
MoCo v2 [7]	R50	200	67.5	-
TKC	R50	200	**69.0**(+1.0)	**88.7**
400 epoch training				
SwAV [4]	R50	400	70.1	-
TKC	R50	400	**70.8**(+0.7)	**89.9**
800 & 1000 epoch training				
PIRL [36]	R50	800	63.6	-
MoCo v2 [7]	R50	800	71.1	-
SimCLR [5]	R50	1000	69.3	89.0
TKC	R50	1000	**72.1**(+1.0)	**90.6**

Table 1. Top-1 and top-5 accuracy under the linear classification protocol on ImageNet with the MoCo framework and ResNet-50 backbone. We report our results of different epochs.

uation are strictly kept aligned with the implementations in [7].

Table 1 summaries the top-1 and top-5 accuracy of our method. We report our results for different epochs pretrained and also list top-performing methods. TKC improves MoCo v2 by 1.5 % on 200 epochs results, which indicates that *temporal teachers* can provide more accurate targets to learn consistent representations. Our results are also superiors to previous works on different pretext tasks, including all other instance discrimination paradigms. This demonstrates that temporal knowledge can benefit from stable training while mitigates the effect of catastrophic forgetting.

In order to verify the scalability of TKC, we respectively conduct our TKC on BYOL [19] baseline, and AlexNet [28] backbone. The number of teachers h is changed to 3 for AlexNet. Specifically, We use a PyTorch implementation of BYOL in Momentum$_2$ Teacher [32] as BYOL baseline and train the model for 100 epochs with 128 batch size on 8 Nvidia-1080ti GPUs. As for AlexNet, we adopt the implementation in Deep Clustering [2], where we train the network with a mini-batch of 1024 on 4 NVidia-1080ti GPUs, and the learning rate is initialized by 0.24 with a cosine decay schedule for 200 epochs. More detail can be seen in the supplementary material.

Method	architecture	epochs	Top-1	Top-5
BYOL[†] [19]	R50	100	70.1	90.6
BYOL[†] + TKC	R50	100	**72.4**(+2.3)	**91.7**(+1.1)

Table 2. Top-1 and top-5 accuracy under the linear classification protocol on ImageNet with BYOL framework. [†] denotes the results from unofficial re-implementations.

Table 2 shows our results on BYOL [19] baseline. We find that TKC can bootstrap BYOL for 2.3%, which shows

that temporal knowledge can also benefit different instance discrimination methods via maximizing the mutual information for temporal targets. The results of BYOL are incompatible with the official ones because we use an unofficial reproduction of BYOL. We conduct this experiment only to prove that TKC can improve different instance discrimination methods. For more details about this reproduction, please refer to the supplementary material.

Method	conv1	conv2	conv3	conv4	conv5
Random	11.6	17.1	16.9	16.3	14.1
Supervised	19.3	36.3	44.2	48.3	50.5
Jigsaw [38]	19.2	30.1	34.7	33.9	28.3
Rotation [18]	18.8	31.7	38.7	38.2	36.5
DeepCluster [2]	12.9	29.2	38.2	39.8	36.1
NPID [51]	16.8	26.5	31.8	34.1	35.6
AET [59]	19.2	32.8	40.6	39.7	37.7
LA [62]	14.9	30.1	35.7	39.4	40.2
ODC [58]	19.6	32.8	40.4	41.4	37.3
Rot-Decouple [14]	19.3	33.3	40.8	41.8	**44.3**
TKC	**20.3**(+1.1)	**34.2**(+0.9)	**42.6**(+1.8)	**46.2**(+4.4)	44.0

Table 3. Top-1 accuracy under the linear classification protocol on ImageNet with the AlexNet backbone. We fine-tune a fc layer from the top of different layers.

For AlexNet, as shown in Table 3, TKC achieves state-of-the-art top-1 accuracy on conv1 to conv4, which outperforms all self-supervised methods on this track. Despite that TKC from conv5 underperforms Rot-decouple [14] by 0.3%, our best result is from conv4, which surpasses the best of Rot-decouple by 1.9%. The results show that TKC is also a leading method on AlexNet linear classification benchmark. We notice that TKC has more improvement on AlexNet than ResNet-50. This might be because the dropout layer in AlexNet can provide various temporal knowledge, which could be more effective in learning instance temporal consistency.

4.2. Transfer to Downstream Tasks

The primary goal of self-supervised learning is to learn good representations that transfer well on downstream tasks. In this subsection, we transfer the representations of 200 epoch TKC to three benchmarks: object detection, instance segmentation, and semi-supervised learning. We show that TKC learns better transferable representations on all three downstream tasks.

Object Detection. We both transfer to VOC [13] and COCO [33] dataset to evaluate our representations. As for Pascal VOC, We use Faster R-CNN [45] with ResNet50 backbone as the detector. We fine-tune the candidate pretrained model for 48k iterations with a min-batch size of 8 on Pascal VOC [13] training set. The learning rate is initialized from 0.001 and then decayed at 36k and 44k iterations. The weight decay is set to 0.0001, and training image scales range between 480 to 800. We use AP_{50}, AP, AP_{75} as evaluation metric on VOC test2007 set.

For COCO [33] dataset, we train a Mask R-CNN [22] to learn the object detection and instance segmentation tasks

pre-train	AP_{50}	AP	AP_{75}
random-init	60.2	33.8	33.1
supervised	81.3	53.5	58.8
NPID++ [51]	79.1	52.3	56.9
PIRL [36]	80.7	54.0	59.7
MoCo v2 [21]	81.5	55.9	62.6
TKC	**81.8**(+0.3)	**56.5**(+0.6)	**62.8**(+0.2)

Table 4. Object detection fine-tuned on PASCAL VOC with Faster-RCNN.

	pre-train	AP_{50}	AP	AP_{75}
	supervised	59.8	40.2	43.8
fine-tune	MoCo v2	60.0	40.1	43.4
	TKC	**60.1**(+0.1)	**40.4**(+0.3)	**43.9**(+0.5)
	supervised	54.3	34.3	36.5
freeze	MoCo v2	48.1	29.2	30.8
	TKC	**54.2**(+6.1)	**34.7**(+5.5)	**37.1**(+6.3)

Table 5. Object detection on COCO. The detection framework is Mask R-CNN. We report the results for both fine-tune and freezing the backbone.

synchronously. We train it for 180k iterations and decay the learning rate by 0.1 at 120k and 160k iterations. The input image size is between 640 and 800 on the training stage and 800 on the test stage. All hyper-parameters of this fine-tuning protocol are consistent with the MoCo v2 baseline.

As shown in Table 4, our TKC achieves 81.8 AP on the PASCAL VOC dataset, which outperforms all pre-trained models from competitors include the supervised ones. Our TKC shows consistent improvement for both AP_{50}, AP, AP_{75}, which shows that TKC indeed learns more consistent and transferable representations than MoCo v2. The upper part of the table 5 shows the results on COCO, TKC as well surpass the MoCo v2 AP_{75} by 0.5% . The results on these two datasets indicated that comprehensive temporal knowledge can lead to transferable representations and learn better representations on different scenes and tasks.

pre-train	AP_{50}	AP	AP_{75}
random-init	24.6	11.6	9.7
supervised	80.2	51.4	55.5
MoCo v2 [21]	79.0	51.7	56.2
TKC	**80.9**(+1.9)	**52.7**(+1.0)	**57.6**(+1.4)

Table 6. Object detection on PASCAL VOC by freezing the backbone and only training the detection head of Faster-RCNN.

We also evaluate TKC on detection in another way. We freeze the Faster R-CNN backbone and only train from the detection head to challenge it. This is somewhat like linear classification. Table 6 shows the results on VOC dataset. For both AP, AP_{50} and AP_{75}, TKC surpass MoCo v2 baseline by more than 1.0%, and also surpass the supervised counterpart. Table 5 shows that on COCO dataset, the improvement is even more than 5.5 %. Train on the frozen backbone can better reflect the pre-trained model's representation because that the trained head is more dependent

on what the pretext task learns. The results on frozen backbone show that TKC does learn better semantics representations. The training detail can be seen in the supplementary material.

Instance Segmentation. We evaluate the instance segmentation on the COCO dataset, following the same setting as COCO detection. Table 7 shows both the results by fine-tune and freezing. The finetune results gains 0.5% AP_{75} on MoCo v2 baseline, indicates the temporal consistency can better locate the target to improve the IOU of instances. Moreover, the gain is further expanded to 1.9 % when only train the segmentation head. We note in this way $AP_{5}0$ is improved by 3.0%, show that TKC can also learn better representations on a simple task.

dataset	pre-train	AP_{50}	AP	AP_{75}
	supervised	56.7	34.9	37.1
fine-tune	MoCo v2	**56.8**	35.0	37.2
	TKC	**56.8**	**35.2**(+0.2)	**37.7**(+0.5)
	supervised	51.1	30.6	31.7
freeze	MoCo v2	48.1	29.2	30.8
	TKC	**51.1**(+3.0)	**30.9**(+1.7)	**32.7**(+1.9)

Table 7. Instance segmentation on COCO. The detection framework is Mask R-CNN. We report the results for both fine-tune and freezing the backbone.

Method	Model	Epochs	Label fraction	
			1%	10%
Supervised	R50v2		48.4	80.4
NPID [51]	R50	200	39.2	77.4
PIRL [36]	R50	800	57.2	83.8
MoCo v1[21]†	R50	200	61.3	84.0
SimCLR [5]†	R50	200	64.5	82.6
MoCo v2 [7]‡	R50	200	61.7	84.6
TKC	R50	200	**72.1**(+10.4)	**86.2**(+1.6)

Table 8. **Semi-supervised Learning on ImageNet.** We finetune the model with 1% and 10% labels. Center-crop top-5 accuracy is reported to compare with previous methods. † indicates that the score is from this work [53]. ‡ means that we implement under the same strategy using the officially released pre-trained model.

Semi-supervised Learning. We then evaluate the utility of TKC in a data-efficient setting by performing semi-supervised learning on ImageNet. In this benchmark, We follow the experimental setup of [5, 36]. The dataset is sampled of 1% and 10% from the labeled ImageNet-1k training data in a class-balanced way. We finetune the TKC pre-trained model on these two labeled subsets and validate it on the whole ImageNet validation data. In order to compare with previous works, we report the top-5 accuracy. The supervised baseline from [57] is trained only using 1% and 10% labels, with a stronger architecture of ResNet50-v2, trained for 1000 epochs. Table 8 shows that our TKC surpasses all the previous methods trained for 200 epochs. When only 1% of data is labeled, TKC surpasses

our MoCo v2 baseline by a large margin of 9.6%, indicating that the temporal knowledge is more beneficial when lacking labeled data.

In addition, the mainstream semi-supervised learning methods adopt a consistent regularization to learn smooth manifold. The intuition in this field is similar to us, where they consider that consistent representation between similar samples can bring up accuracy classification boundary. Similarly, TKC also encourages consistent representations between different training stages to get a smoother manifold. The significant improvement on semi-supervised benchmarks shows we indeed learn temporal consistent representations.

4.3. Analysis

Ablation study. Our method introduces two new hyper-parameters, the step interval of each teacher s, and amount of teachers h. We use s as the steps among an epoch and do not tune it. For h, we take an ablation on the AlexNet backbone. In Table 9, the first column $h = 1$ means only the EMA teacher is used, which is an implementation of MoCo v2. We see that the *temporal teacher* can boost accuracy from 39.9 to 42.2 when introduced only one temporal teacher. TKC achieves the best performance when maintaining three teachers, for this setting can acquire the most temporal knowledge to stable the target. When increasing h even more, the result is declined unexpectedly. This might be because when involving the too old teachers, their representations are changed too much. It is hard to learn consistently with these teachers. Nonetheless, this confirms our motivation again that the inconsistency between different training stages has alleviated convergence.

h	1	2	3	4	6
Top-1	39.9	42.2	43.5	41.9	41.8

Table 9. Ablation study on the effect of teacher numbers.

Convergence comparison. In Section 3.4, We consider that TKC can combine maximize the mutual information with the target from the different stages, and therefore will enforce the network to learn temporal consistent representations and mitigate the catastrophic forgetting. To confirm our proposal, we use a kNN classifier to validate the model performance during training. As shown in Fig. 3, TKC has a lower accuracy in the earlier training, which is because that the model is more inconsistent and noisy in the earlier stage, resulting in a big difference between the *temporal teacher* and the current teacher. This difference prevents TKC from providing consistent signals. However, the TKC catches up MoCo v2 from the middle stage and finally surpasses it for 4.6% at the end of the training, which indicates that TKC can stably provide a consistent signal from the middle training. This consistent signal can guide a more

accurate training direct and accelerate convergence. Fig. 3 shows that the TKC at 160 epochs meets the accuracy of fully trained MoCo, reducing 80% training time by mitigating catastrophic forgetting.
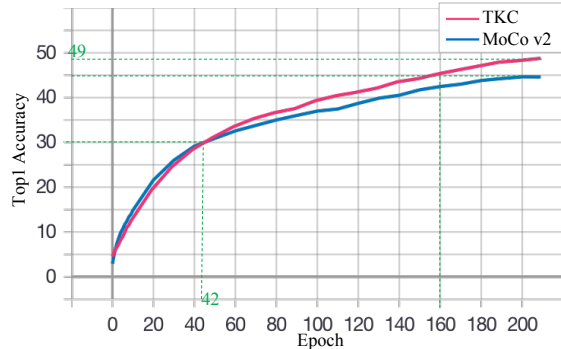

Figure 3. Comparison of validation accuracy between MoCo v2 and TKC. The top-1 accuracy is from a kNN classifier.

5. Conclusion

We summarize the existing instance discrimination methods into a teacher-student framework and note that the teacher can only provide instance spatial consistency. However, the output of the same instance can vary dramatically between different epochs when only spatial consistency is involved. We instead present a novel and strong method named Temporal Knowledge Consistency (TKC), which integrates the knowledge from previous teachers to improve the model's robustness and prevent possible catastrophic forgetting. TKC contains three modules. The *temporal teacher* introduces the instance temporal consistency from previous models, the *knowledge transformer* leverages the knowledge of these teachers, and the *temporal loss* reduces the MI between the student and the temporal teacher. Temporal teacher is an orthogonal improvement for different instance discrimination methods. Our experimental results show that TKC can improve different frameworks MoCo, BYOL, and architectures ResNet-50, AlexNet. It also provides transferable representations on downstream tasks such as object detection, instance segmentation, and semi-supervised learning. Moreover, we hope our study can draw much attention to solve the unstable in unsupervised learning and search for effective ways to generate stable output with no labels.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning

of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 1, 2, 6

[3] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2959–2968, 2019. 2

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 5, 6, 7

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019. 1, 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5, 6, 7

[8] 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. Ieee, 2009. 1, 5

[9] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 567–575, 2015. 1, 2

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[11] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 2

[12] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10542–10552, 2019. 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 6

[14] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10364–10374, 2019. 1, 2, 6

[15] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[16] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels, 2020. 2

[17] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 6

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 5, 6

[20] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 2

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 5, 6, 7

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 5

[24] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 5, 6

[25] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 2020. 6

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, pages 1–13, 2015. 3

[27] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 2

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1, 6

[29] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *International Conference on Machine Learning*, 2017. 2, 3

[30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic

colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 1, 2

[31] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 1, 2

[32] Zeming Li, Songtao Liu, and Jian Sun. Momentumˆ 2 teacher: Momentum teacher with momentum statistics for self-supervised learning. *arXiv preprint arXiv:2101.07525*, 2021. 6

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 6

[34] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[35] Lars Mescheder, S Nowozin, and Andreas Geiger. Adversarranzato2007unsupervisedial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *34th International Conference on Machine Learning (ICML)*, pages 2391–2400. PMLR, 2017. 2

[36] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 6, 7

[37] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983. 3

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 2, 6

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[40] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2

[41] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 2

[42] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 2

[43] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999. 3

[44] Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 6

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 1

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 3

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 6

[49] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. *International Conference on Learning Representations (ICLR)*, 2021. 6

[50] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 2

[51] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 6, 7

[52] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016. 2

[53] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *arXiv preprint arXiv:2008.11702*, 2020. 7

[54] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 2

[55] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020. 2

[56] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2016. 2

[57] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019. 7

[58] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised

representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6688–6697, 2020. 2, 6

[59] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019. 6

[60] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1, 2

[61] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Time-consistent self-supervision for semi-supervised learning. In *International Conference on Machine Learning*, pages 11523–11533. PMLR, 2020. 2, 3

[62] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 6