Vi^{2}CLR: Video and Image for Visual Contrastive Learning of Representation

Ali Diba[1,2], Vivek Sharma[5,6], Reza Safdari[2], Dariush Lotfi[2], M. Saquib Sarfraz[3,4],
Rainer Stiefelhagen[3], Luc Van Gool[1,2,7],

[1] KU Leuven, [2] Sensifai, [3] Karlsruhe Institute of Technology, [4] Daimler TSS,

[5] Massachusetts Institute of Technology, [6] Harvard Medical School, [7] ETH Zurich

Abstract

In this paper, we introduce a novel self-supervised visual representation learning method which understands both images and videos in a joint learning fashion. The proposed neural network architecture and objectives are designed to obtain two different Convolutional Neural Networks for solving visual recognition tasks in the domain of videos and images. Our method called Video/Image for Visual Contrastive Learning of Representation(Vi^{2}CLR) uses unlabeled videos to exploit dynamic and static visual cues for self-supervised and instances similarity/dissimilarity learning. Vi^{2}CLR optimization pipeline consists of visual clustering part and representation learning based on groups of similar positive instances within a cluster and negative ones from other clusters and learning visual clusters and their distances. We show how a joint self-supervised visual clustering and instance similarity learning with 2D (image) and 3D (video) CovNet encoders yields such robust and near to supervised learning performance.

We extensively evaluate the method on downstream tasks like large scale action recognition, image and object classification on datasets like Kinetics, ImageNet, Pascal VOC'07 and UCF101 and achieve outstanding results compared to state-of-the-art self-supervised methods.

1. Introduction

Learning strong and discriminative representations is important for diverse applications in computer vision tasks such as image classification, object detection, image segmentation, activity recognition, video classification, medical imaging as well as natural language processing. More recently unsupervised or self-supervised representation learning has received a lot of attention as these methods are not dependent on manually curated ground-truth labels but rather utilize the supervision coming from the data itself and still rapidly close the performance gap with the supervised training. Most recent state-of-the-art methods are largely driven by instance [77, 12, 29] or prototype [44, 11] discrimination tasks. These discrimination methods rely on combination of two key components: (a) contrastive loss and (b) image [77, 29, 44, 12, 11] or video [28, 27, 56] augmentation. The contrastive loss [25] encourages small distances by pulling samples from the same label together and pushing far apart at least by the margin for the samples of different labels in feature space. The current contrastive loss functions are in the form of noise contrastive estimator [24] to compare instances (InfoNCE [51]), prototypes (ProtoNCE [44]), instances that include samples with the same semantic labels (UberNCE [28]) or complementary views, and multiple instances (Multi-Instance InfoNCE [45]). The data augmentation or transformation can be categorized into two types on the basis of the datatype, namely for images, and for videos. For instance discrimination [12] each sample of the dataset is treated as a class and enforce the augmented version of the same sample to be more similar, while in case of prototypes [44, 11] enforcing the augmented version of the samples to be closer to the prototype. Data augmentation plays a crucial role in the images and videos contrastive representation learning. In particular, for images [12, 11] the most popular augmentation methods are color transformation, geometric transformation and multi-crop; and for videos [28, 27, 56] randomly mining clips from the same video as positives, temporally consistent spatial augmentation, and mining complementary information from different views of the RGB-stream/optical-flow data are the transformation methods. See Section 2 in the related work for an extended review.

In this paper, we propose to extend the self-supervised training of ConvNets for solving visual recognition tasks both in videos and images simultaneously. Our contribution named as Vi^{2}CLR is a method that jointly optimizes two ConvNets for Videos and Images for Visual Contrastive Learning of Representation as a multi-task learning problem. We achieve this by learning both dynamic and static visual cues simultaneously in an end-to-end learning pipeline.

Vi^{2}CLR optimization utilizes clustering as supervision for learning an effective visual (2D ConvNets) and video (3D ConvNets) representations. We believe clustering of-

fers an ability to bring together a diverse set of samples from images/videos across the whole dataset, which in turn provides variability and diversity which is a good way to learn representations and is an important factor for the increased performance shown in Section 4. For learning an effective representation we considered two aspects, they are: (a) all image or video instances in a given cluster are considered positive pairs, and negative pairs are mined from the batch minimizing Multi-Instance InfoNCE loss; and (b) all joint image-video representations in a given cluster are enforced to be closer to the cluster centroid, and negatives are the centroids of the other clusters minimizing our centroid InfoNCE loss. We name this loss *CenterNCE*.

We validate our Vi^{2}CLR based 2D and 3D ConvNets by fine-tuning them on downstream tasks and evaluating them on several standard downstream video and image classification benchmarks. For 3D ConvNets they are fine-tuned on target action recognition datasets, and for 2D ConvNets we use the learned features without fine-tuning and rather only employ an MLP projection head (i.e. linear classifier) on top of the frozen features, following [11]. Our 3D ConvNet is evaluated on three challenging benchmark action recognition datasets namely UCF101, HMDB51 and Kinetics-400. We experimentally show that our Vi^{2}CLR achieves state-of-the-art performance on UCF101 (88.9%), HMDB51 (55.7%) and Kinetics-400 (71.2%) outperforming all current video contrastive learning methods [28, 27, 56]. Our 2D ConvNet is evaluated using the ImageNet linear evaluation protocol. We have also presented that our Vi^{2}CLR outperforms SimCLR [12], SwAV [11] with achieving 74.6% top-1 accuracy on ImageNet.

2. Related work

This section discusses self-supervised image and video representation learning.

Popular contrastive learning methods. We aim to learn a representation that exhibits small distances between samples from the category, and large distances from different category in feature space, using contrastive loss function [25] one could achieve that by pulling samples from the same class closer and pushing samples from a different class further away. Of more relevance here is the line of research using contrastive learning in images [77, 29, 44, 12, 11] and videos [28, 27, 56]. The memory bank [77, 46] method accumulates the previously computed instance class representation, and then use that to form positive and negative pairs. They use noise contrastive estimator [24] to compare instances, which is a special form of contrastive learning [31, 51]. The end-to-end [81, 69, 12] method generates distinct representations of the same sample within the current mini-batch, replacing the memory bank. The momentum encoder [29] method uses a momentum-updated encoder which acts as a dynamic dictionary lookup for encoding samples on-the-fly. The contrastive clustering [44, 11] methods such as Prototypical Contrastive Learning (PCL) [44], or Swapping Assignment between multiple views (SwAV) [11] aims to learn the class prototype and enforce for samples belonging to a cluster to stay close.

Memory-augmented Dense Predictive Coding (MemDPC) [27] utilizes compressed memory (i.e. memory bank) for self-supervised video representation learning from RGB frames, or unsupervised optical flow. Contrastive Video Representation Learning (CVRL) [56] builds upon SimCLR [12] and perform temporally consistent spatial augmentation to train the network to pull clips from the same video and push clips from different videos in the feature space. In contrast to these works, in our work, we jointly optimize two ConvNets for Videos and Images for Visual Contrastive Learning of Representation as a multi-task learning problem using a clustering objective as supervision for learning. Next, the various pretext tasks and pseudo-labels strategies discussed below are based on contrastive learning.

Pretext tasks based on self-supervised learning. Self-supervised representation learning is increasing in popularity in recent years. This learning paradigm obtains supervision by exploiting the structure within the data, and thus removes the need for an often costly labeling effort.

For example, one may learn image representations via in-painting [52], patch prediction [16], solving the jigsaw puzzle [49], colorization [84], geometric transformation [18], learning to generate an accurate distribution of real images [33], predicting the future in egocentric videos [85], learning the steadiness of visual change that temporally close frames exhibit only small differences in feature space [32] or to recognize complex, long-term activities [42], learned through a proxy task of inferring the temporal ordering of a set of unordered videos in a timeline [64], predicting the geometric transformation that is applied to the image [22] or video [35], counting the number of visual primitives in the image wrt. scaling and tiling [50], predicting the future frames [15, 27, 71], predicting the speed [6, 19, 73], correspondence between video-and-narrations [45] or frames-and-audio [1, 2, 3, 38, 53, 55, 5], predicting optical-flow [48], transfer knowledge between flow-and-rgb network [67] or image-and-video [14], more recently predicting similar images invariant to their multiple data augmentation (data transformation) [12].

Pseudo-labels based on self-supervised learning. Pseudo-labels based discriminative representation learning has attracted quite some attention because it removes the need for an often costly labeling effort. For example, some pretext tasks form pseudo-labels by posing the problem as a non-parametric classification problem at the instance-level [77], augmenting the data by applying a random set of transformations to each patch and the considering it as a unique la-

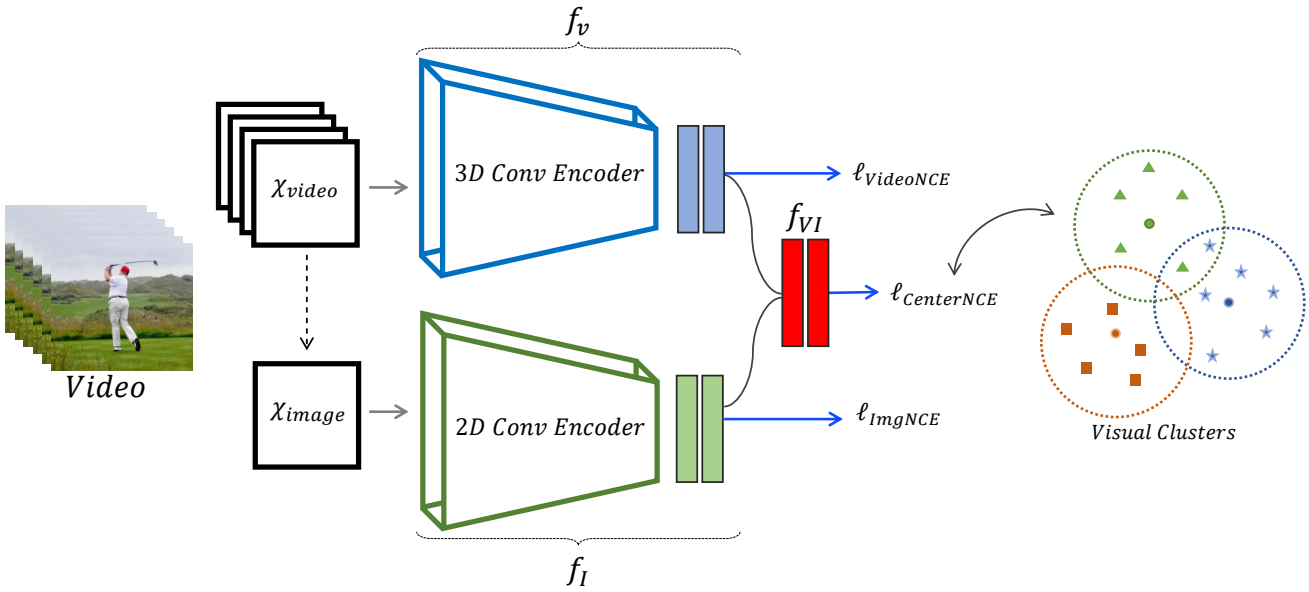

Figure 1. **Vi^{2}CLR training pipeline**. Illustration of our approach for jointly learning image and video representations. Vi^{2}CLR optimization utilizes a combination of three contrastive losses optimizing cluster-based, video and image sample-based contrastive learning based on positive and negative sample similarities within visual clusters. f_V and f_I denote visual encoder of video and image streams.

bel (Exemplar-CNN) [18], using K-means clustering algorithm with a pre-determined number of clusters as a means of generating pseudo labels [9] or the FINCH clustering algorithm that automatically estimates the number of clusters [58, 63], jointly optimize clustering and representation learning [82, 9, 58, 10, 1], via maximising mutual information objective between the class assignments of each pair [34], by aligning the features to the noise as targets [7], or using optimization-based approach [70].

Finally, it is worth noting the work using temporal continuity or ordering as a signal for supervision. For example, one may obtain pseudo-labels via shuffling frames [47], finding an odd frame [21], by sorting distances [61, 60], using tracking information [74, 62], predicting whether a video flows forwards or backwards [54, 76], inferring the temporal ordering [80, 64, 43], by temporal coherence [40, 41, 72, 75]. In contrast to these, our work differs substantially in scope and technical approach. Our contribution is a method that utilizes clustering assignments as supervision for learning an effective visual and video representation together. To the best of our knowledge, the Vi^{2}CLR is the first of its kind self-supervised neural network to tackle both video and image recognition task simultaneously by only using one source of data.

3. Methodology

Given tremendous successes with deep learning, creating an effective image or video representation seems not far fetched via self-supervised representation learning anymore. We aim to learn representations that embody semantic information relating to images to video and vice-versa. Our goal is to learn solving visual recognition tasks both in videos and images simultaneously from unlabeled videos in an end-to-end pipeline. We propose a method that jointly optimizes a 3D ConvNet for Videos and a 2D ConvNet for Images. Furthermore, our method utilizes clustering as supervision for representation learning in a multi-task learning setup.

We first introduce the preliminaries in Section 3.1 on self-supervised learning with InfoNCE used by [12] and Multi-Instance InfoNCE used by [28].Finally in Section 3.2, we introduce our video/image contrastive learning of representation (Vi^{2}CLR) pipeline using clustering and instance discrimination objective in a contrastive learning setup. Algorithm 1 sketches the steps of the proposed Vi^{2}CLR. Note that, our method yields two trained deep ConvNet models, one for videos (3D ConvNets) and one for images (2D ConvNets) at the end trained with an only single source of unlabeled videos in end-to-end learning.

3.1. Self-Supervised Learning by InfoNCE

Assume we have a set of unlabeled video clips with N samples, $X = \{x_1, x_2, .., x_N\}$. One can train a self-supervised representation neural network $f(.)$ with the InfoNCE [12, 28] objective loss function. The objective of InfoNCE works as an instance discriminator, that pulls positive sample representations closer while it repels negative samples apart in the feature space. This approach of self-supervised training network can be utilized in other tasks for video understanding like action recognition, video captioning, retrieval, etc. Let's assume the query sample representation be $r_i = f(x_i)$, the InfoNCE contrastive loss

function ($\mathcal{L}_{\text{InfoNCE}}$) is defined as:

$$\sum_{i=1}^{n} -\log \frac{\exp\left(r_i \cdot r_p/\tau\right)}{\exp\left(r_i \cdot r_p/\tau\right) + \sum_{n\in\mathcal{N}_i} \exp\left(r_i \cdot r_n/\tau\right)} \quad (1)$$

where r_p is a positive sample feature representation, for instance, x_i which is and augmented set of the original sample, while $\mathcal{N}_i$ conversely refers to an associated set of negative samples where r_n is a negative sample representation in the mini-batch, and τ is the temperature hyper-parameter.

Multi-Instance InfoNCE. Assume we have a set of multiple positive instances for a query instance, here we address the Multi-Instance InfoNCE (MIL-NCE) [45] objective loss function. The objective of MIL-NCE works similar to InfoNCE while considering one-or-more actual positives within the $\mathcal{P}$. Based on a positive sample set $\mathcal{P}$ and a negative sample set $\mathcal{N}$, the MIL-NCE objective loss function ($\mathcal{L}_{MIL-NCE}$) is defined as:

$$\sum_{i=1}^{n} -\log \frac{\sum_{p\in\mathcal{P}_i} \exp\left(r_i \cdot r_p/\tau\right)}{\sum_{p\in\mathcal{P}_i} \exp\left(r_i \cdot r_p/\tau\right) + \sum_{n\in\mathcal{N}_i} \exp\left(r_i \cdot r_n/\tau\right)} \quad (2)$$

The positive set may contain multiple positive samples for a given query plus its own augmentations, and the negative set contains all other samples in the mini-batch, and their augmentations. Because of MIL-NCE ability to handle multiple instances, recently the objective has been used for training self-supervised contrastive methods in the domain of images [45] or videos [28] in different setups.

3.2. Vi^{2}CLR

This section describes the proposed approach to train our self-supervised Vi^{2}CLR. The Vi^{2}CLR optimization utilizes clustering as supervision for learning an effective visual (2D ConvNets) and video (3D ConvNets) representation. More specifically, we train both the video and image data stream together via their respective 3D and 2D ConvNets simultaneously. The training objective is to contrast between multiple positive and negative instances for a given query via comparing their cluster assignments. The Vi^{2}CLR training routine proceeds in two steps: first learning representation and second clustering of samples. Algorithm 1 sketches the steps of the proposed Vi^{2}CLR. The clusters are constructed from joint 2D/3D learned feature representations (discussed later) to exploit both dynamic and static visual cues. The total objective function is a combination of three contrastive losses optimizing cluster-based, video and image sample-based contrastive learning to capture higher-level semantic knowledge utilizing global and local similarities and dissimilarities. Figure 1 illustrates our approach.

Our Vi^{2}CLR learns two objective functions: $f_V(.)$ and $f_I(.)$, where $r_V = f_V(x)$ and $r_I = f_I(\hat{x})$ refer to the representation of the 3D ConvNet (video) and 2D ConvNet (image) encoder, where x is the video clip and $\hat{x}$ is an image. As 2D ConvNets expect images as input which are 2D (spatial) in nature, we extract a frame from middle of the video clip and feed it as input to the 2D ConvNets. The video clip itself (spatial + temporal dimension) forms the input to the 3D ConvNets. Formally, $f_V : x \rightarrow r_V, r_V \in \mathbb{R}^{d_1}$ and $f_I : \hat{x} \rightarrow r_I, r_I \in \mathbb{R}^{d_2}$, where d_1 and d_2 denotes dimensionality of the encoded video and image embedding space, respectively. As mentioned earlier, we perform clustering on joint 2D/3D representations. For that, after we have extracted the video and image encoded representations, we first concatenate the feature maps and then feed it to a non-linear layer to obtain $r_J \in \mathbb{R}^d$ where d denotes the encoded feature dimension, which we refer to as joint 2D/3D representation. In each epoch of Vi^{2}CLR training, we first extract the features of the entire dataset, form a joint 2D/3D representation R_J and then perform clustering [4, 58] on the features to obtain cluster assignments. More details on clustering can be found in the experimental section.

CenterNCE loss: The joint 2D/3D representation is also used for computing the *CenterNCE* loss. After obtaining joint 2D/3D representation based cluster assignments for each instance, we compute and store the cluster centroids for each cluster. In practice during training, for a given query sample x_i, the sample is enforced to be closer to the cluster centroid it belongs to, and negatives are the centroids of the other clusters in the mini-batch. In the similar spirit of other contrastive learning research like [11, 44], our centroid based *CenterNCE* loss ($\mathcal{L}_{\text{CenterNCE}}$) is as follows:

$$\sum_{i=1}^{n} -\log \frac{\exp\left(r_{Ji} \cdot c_s/\phi_s\right)}{\sum_{j=1}^{k} \exp\left(r_{Ji} \cdot c_j/\phi_j\right)} \quad (3)$$

where $r_J = f_{VI}(x_i)$ is the joint 2D/3D representation, n is the batch-size, k is the #clusters in the dataset, c_s is the centroid of the cluster that x_i belongs to, and ϕ denotes concentration estimate of each cluster to ensure learning of more balanced cluster [44].

Since we have access to cluster assignments for each sample, we can take advantage of mining promising groups of positive and negative sets based on the clustering. For learning an effective representation, we thus considered mining positive and negative pairs using cluster assignments. In specific, for both $r_V = f_V(x)$ in video space (3D) and $r_I = f_I(\hat{x})$ in image space (2D) streams, we have separate MIL-NCE objective functions, defined as $\mathcal{L}_{VideoNCE}$ and $\mathcal{L}_{ImgNCE}$ and are based on Eq. 2. The total loss function for Vi^{2}CLR with two ConvNets (2D ConvNet and 3D ConvNet) encoder is given as:

$$\mathcal{L}_{\text{Vi}^2\text{CLR}} = \mathcal{L}_{\text{CenterNCE}} + \mathcal{L}_{\text{VideoNCE}} + \mathcal{L}_{\text{ImgNCE}} \quad (4)$$

where equal weight was given to each loss.

Positive/Negative Samples: For each instance in a given cluster, we randomly mine samples from the same cluster

as positive pairs, and negative samples are mined from other clusters in the mini-batch during training.

Note that, as the training proceeds, in each epoch the network gains stronger representations, thus leading to mine better positive pairs. We believe, this in turn, progressively leads to semantically-meaningful clusters that accounts for improving the performance of the 2D and 3D ConvNet encoders. After the encoders are trained, we use them for performing various downstream tasks, such as action recognition and image classification.

Algorithm 1 Vi^{2}CLR Training.

Input: Video Clips $X = \{x_1, x_2, .., x_N\}$, Video encoder $f_V(x)$, Image encoder $f_I(\hat{x})$. Where x is the video clip and $\hat{x}$ is the middle frame of x

while *Not* $MaxEpoch$ **do**

  $R_J = f_{VI}(X)$

    `// Joint 2D, 3D representation`

  $C = Clustering(R_J)$

    `// Clustering Assignment`

  for x *in Batch(X)* **do**

    $r_V = f_V(x), r_I = f_I(\hat{x}), r_J = f_{VI}(x)$

    $\mathcal{L}_{\text{CenterNCE}}(r_J, C)$

    $\mathcal{L}_{\text{VideoNCE}}(r_V, \mathcal{P}_{C,x})$

    $\mathcal{L}_{\text{ImgNCE}}(r_I, \mathcal{P}_{C,x})$

    $\mathcal{L}_{\text{Vi}^2\text{CLR}} = \mathcal{L}_{\text{CenterNCE}} + \mathcal{L}_{\text{VideoNCE}} + \mathcal{L}_{\text{ImgNCE}}$

  end

end

4. Experiments

In this section, we first introduce the datasets used for Vi^{2}CLR training and downstream tasks (image and video) datasets for evaluation. Followed by the implementation details. Finally, we compare to state-of-the-art self-supervised methods on image classification, video classification and video retrieval tasks.

4.1. Datasets

To train our Vi^{2}CLR 2D/3D model, we use the Kinetics-400 [36] training set consisting of ∼250K video clips with a maximum duration of 10 seconds. For the downstream video recognition task, we benchmark on the Kinetics, UCF101 [66] and HMBD51 [39] and for image recognition task, we benchmark on the ImageNet ILSVRC-2012 [57] and Pascal VOC2007 [20] datasets.

4.2. Vi^{2}CLR Implementation Details

We choose ResNet-50 and S3D [79] as our 2D and 3D ConvNet encoders for Vi^{2}CLR training, and which are then used for downstream tasks. Same as recent contrastive learning methods, SimCLR [12] and CoCLR [28], for both encoders, we attach a non-linear MLP projection head with 128-dimensions (i.e. $d_1 = 128$ and $d_2 = 128$). The concatenated output of the two encoders is fed to another MLP projection head of 128-dimensions (i.e. $d = 128$) resulting in joint 2D/3D representation for computing $CenterNCE$ loss. We remove the MLP projection heads for both 2D/3D encoders for downstream task evaluations, as done in SimCLR [12]. For Vi^{2}CLR 3D ConvNet training, we resize the video clips with a spatial resolution of 128×128, where we extract the middle frame of the video clip for 2D ConvNet. Note that while one can use a random frame from the video clip for the 2D input we empirically found that choosing

Clustering Method	Epoch #50	Epoch #100	Epoch #200
Vi^{2}CLR (**Kmeans**)	64.3	71.3	73.7
Vi^{2}CLR (**FINCH**)	65.5	72.9	74.3

Table 1. **Impact of Clustering** on Vi^{2}CLR. Top-1 accuracy for linear image classification task with frozen weights and a single classification layer trained on ImageNet using ResNet-50.

Method	ImageNet	VOC07
Supervised	76.5	87.5
Jigsaw [49]	45.7	64.9
Colorization [84]	39.6	55.6
BigBiGAN [17]	56.6	-
MoCo [29]	60.6	79.2
PIRL [46]	63.6	81.1
SeLa [82]	61.5	-
CPCv2 [30]	65.9	-
SimCLR [12]	61.9	-
SimCLR [12]	69.3	-
PCL [44]	67.6	85.4
MoCov2 [13]	71.1	-
SwAV [11]	74.2	88.9
Vi^{2}CLR	74.6	89.4

Table 2. **Linear classification** on ImageNet. Top-1 accuracy for linear classification task with frozen weights and a single classification layer trained on ImageNet. All the methods use ResNet-50 as backbone architecture with 24M parameters.

Method	k=1	k=2	k=4	k=8	k=16
Supervised	54.3	67.8	73.9	79.6	82.3
Jigsaw [49]	26.5	31.1	40.0	46.7	51.8
SimCLR [12]	32.7	43.1	52.5	61.0	67.1
MoCo [29]	31.4	42.0	49.5	60.0	65.9
PCL [44]	47.9	59.6	66.2	74.5	78.3
Vi^{2}CLR	49.1	62.2	68.4	76.8	80.6

Table 3. **Few-shot classification** on Pascal VOC07 dataset using linear SVMs trained on fixed representations. All the compared methods use ResNet-50 pre-trained on ImageNet for feature extraction.

Method	Learning Method	# Training Epoch	1% Labeled Images		10% Labeled Images	
			top-1	top-5	top-1	top-5
Supervised	-	-	25.4	48.4	56.4	80.4
UDA [78]	label-propagation	-	-	-	68.8	88.5
FixMatch [65]	label-propagation	-	-	-	71.5	89.1
Pseudolabels [83]	Semi-supervised	-	-	-	51.6	82.4
S^4L Exemplar [83]	Semi-supervised	-	-	-	47.0	83.7
S^4L Rotation [83]	Semi-supervised	-	-	-	53.4	83.8
PIRL [46]	Self-supervised	800	30.7	57.2	60.4	83.8
Jigsaw [49]	Self-supervised	90	-	-	45.3	79.3
SimCLR [12]	Self-supervised	200	-	-	56.5	82.7
MoCo [29]	Self-supervised	200	-	-	56.9	83.0
PCL [44]	Self-supervised	200	-	-	75.3	85.6
SwAV [11]	Self-supervised	**800**	53.9	78.5	70.2	89.9
SwAV [11]	Self-supervised	300	52.7	77.0	68.9	88.7
Vi^{2}CLR	Self-supervised	300	53.3	77.8	69.7	89.1

Table 4. **Semi-Supervised Learning** on ImageNet. We show top-1 and top-5 accuracy results on ImageNet validation set when fine-tuned on 1% or 10% labeled data.

a frame from middle of the video results in better performance. We use temporally consistent spatial augmentations by random crop, Gaussian blur and color jittering. Further, we also perform random temporal cropping with size 32 frames from the same video as positives. We train Vi^{2}CLR for 300 epochs with a batch-size of 64 on each GPU. We use 8 V100 32GB GPUs for our model training. For linear classification experiments, we train the image and video encoders for 100 epochs on both setups: (1) frozen weights , and (2) full fine-tuning. We use Adam optimizer, with a weight decay of 0.0001 and an initial learning rate of 0.01, which is reduced by a factor of 10 every 100 epochs. For all experiments, the temperature parameter is set to 0.08.

4.3. Impact of Clustering on Vi^{2}CLR

Integral to our core method, we adopt the recently proposed FINCH algorithm [58] to obtain weak labels from clustering. FINCH belongs to the family of hierarchical clustering algorithms and automatically discovers meaningful partitions without requiring hyper-parameters such as the number of clusters K. In contrast, existing self-supervised clustering based methods such as [44, 9, 10] specify the number of clusters manually. Additionally, FINCH provides clusters with very high purity at early partitions; and it is a fast and scalable algorithm with computational complexity of $\mathcal{O}(Nlog(N))$. Following the suggestions given in [63, 59], we use clusters from the second partition, to mine cluster assignments as this partition increases diversity, without compromising on the quality of the labels.

In Table 1, we show the results for linear classification task with frozen weights and where only a single linear layer is trained with cross-entropy loss on the ImageNet dataset. As an alternative to FINCH, we also perform experiments with K-means [4] as a baseline to obtain clusters. Note that K-means requires a prior knowledge such as number of clusters. For a fair empirical comparison, we use the FINCH estimated clusters K as input to K-means.We can observe that FINCH not only automatically discovers meaningful partitions of the data, but also achieves higher performance as compared to K-means. We observed the FINCH partition provides 10K-15K clusters within our training setup.

4.4. Image Classification

Linear classification. We evaluate the learned representation of a ResNet-50, a 2D ConvNet encoder trained with Vi^{2}CLR. For this evaluation, we follow the same setup from [12, 44] and perform linear classification task with frozen weights and only a single linear layer is trained with cross-entropy loss on ImageNet and Pascal VOC2007 datasets. In Table 2, we report the results. Vi^{2}CLR achieves the highest single-crop top-1 accuracy among all self-supervised methods that use a ResNet-50 model with no more than 300 pre-training epochs.

Few shot classification. For this evaluation, we use the learned representation of our 2D ConvNet encoder trained with Vi^{2}CLR for object classification with few training samples per-category. Following the same setup from [23], we train linear SVMs using fixed representations on PASCAL VOC2007 [20]. We vary the number k of samples per-class and report the results average over 5 runs. In Table 3, we show the results. One can observe

Method	UCF				HMDB			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Jigsaw [49]	19.7	28.5	33.5	40.0	-	-	-	-
OPN [43]	19.9	28.7	34.0	40.6	-	-	-	-
Buchler [8]	25.7	36.2	42.2	49.2	-	-	-	-
VCOP [80]	14.1	30.3	40.4	51.1	7.6	22.9	34.4	48.8
MemDPC [27]	20.2	40.4	52.4	64.7	7.7	25.7	40.6	57.7
CoCLR-RGB [28]	53.3	69.4	76.6	82.0	23.2	43.2	53.5	65.5
Vi^{2}CLR	55.4	70.9	78.3	83.6	24.6	45.1	54.9	67.6

Table 5. **Video Retrieval.** Comparison of Vi^{2}CLR with the state-of-the-art on nearest-neighbour video retrieval on UCF101 and HMDB51. Given query test clips, our goal is to find training clips that are from the same class using Recall at k (R@k) metric.

Method	ImageNet	VOC'07	UCF101	HMDB51
Vi^{2}CLR-$\mathcal{L}_{\text{ImgNCE}}$	71.2	86.1	-	-
Vi^{2}CLR-$\mathcal{L}_{\text{VideoNCE}}$	-	-	86.6	51.2
Vi^{2}CLR-$\mathcal{L}_{\text{CenterNCE}}$	72.7	88	87.8	53.4
Vi^{2}CLR-Full	74.6	89.4	89.1	55.7

Table 6. Performance comparison of using different contrastive objective functions to train Vi^{2}CLR

Vi^{2}CLR outperforms MoCo [29], PCL [44], SimCLR [12] by a great margin.

Semi-supervised classification. For this evaluation, we use the learned representation of our 2D ConvNet encoder trained with Vi^{2}CLR for image classification with fine-tuning the entire encoder and a linear classification layer on a randomly selected subset (1% or 10%) of ImageNet training data (with few labels). We follow the same setup from [46]. In Table 4, we show top-1 and top-5 accuracy results on ImageNet validation set. Our method substantially outperforms the previous state-of-the-art on both self-supervised and semi-supervised learning methods.

4.5. Video Retrieval and Classification

Video retrieval. For this evaluation, we use the learned representation of an S3D, a 3D ConvNet encoder trained with Vi^{2}CLR on video retrieval downstream task. For the retrieval task, we evaluate the extracted representation directly for nearest-neighbour (NN) retrieval without any further training. We apply the setup from [47] and test if the query (test set) clip and its nearest neighbors in the gallery set (train set) belong to the same class. The performance is measured using Recall at k (R@k). In Table 5, we show the results. We show that a better representation learned using Vi^{2}CLR helps obtain effective video retrieval.

Video Classification. For video classification evaluation, we use the learned representation of our 3D ConvNet encoder trained with Vi^{2}CLR for video classification downstream task. We consider two setups, (a) the entire encoder weights are frozen and only a single linear classification layer is trained, and (b) the entire encoder and a linear classification layer are fine-tuned. The classification layer is trained with cross-entropy loss on target datasets, UCF101, HMDB51 and Kinetics-400. In Table 7, we show the comparison of our method to state-of-the-art methods in light of the recent self-supervised action classification method on these datasets. We can observer that the trained S3D model by our Vi^{2}CLR surpasses and achieves superior results than methods which use other modality of data like optical flow [28], multi-modal information. There are some results from ELO [55] or XDC [1] and they use datasets like Youtube8M and IG65M which are 100-150 times larger than Kinetics dataset that we used.

4.6. Ablation Study

We have studied the impact of each objective function for video and image streams to show the effectiveness of the whole training set up of Vi^{2}CLR. Since our proposed method is trained with three different objective functions, we present the performance of each separately on the down-stream tasks. Therefore, we had to train models in three different setups; joint training of video and image with $\mathcal{L}_{\text{CenterNCE}}$ only; training image stream with $\mathcal{L}_{\text{ImgNCE}}$; and video stream with $\mathcal{L}_{\text{VideoNCE}}$ independently. For all three setups, we utilize clustering and instance cluster assignments and additionally for the later two setups, we further perform sampling using the positive and negative pairs for the loss calculation. For joint training, we perform clustering on joint 2D/3D embedding, while when training the video or image stream encoder we performed clustering on each streams embedding as described in Section 3.2. For the single-stream training setup, the corresponding datasets used for down-stream tasks were ImageNet and Pascal VOC2007 for image stream; and UCF101 and HMDB51 for video stream training.

In Table 6, we compare full Vi^{2}CLR training total loss against each objective function on corresponding down-stream tasks. It can be observed in all cases, the full Vi^{2}CLR achieves superior results compared to training with

Method	Training Dataset	ConvNet Arch.	Input Res.	Weight Frozen	UCF101	HMDB51	Kinetics-400
S3D [79]	Kinetics-400	S3D	224	Supervised	96.8	75.9	74.7
DynamoNet [15]	Kinetics-400	STCNet	112	Supervised	97.8	76.8	77.9
R(2+1)D [15]	Kinetics-400	3D ResNet-50	224	Supervised	96.8	74.5	74.3
CBT [68]	Kinetics-600	S3D	112	✓	54.0	29.5	-
MemDPC [27]	Kinetics-400	R-2D3D	224	✓	54.1	30.5	-
MIL-NCE [45]	HTM	S3D	224-	✓	82.7	53.1	-
XDC [1]	IG65M	R(2+1)D	224	✓	85.3	56.0	-
ELO [55]	Youtube8M	R(2+1)D	224	✓	-	64.5	-
CoCLR [28]	UCF	S3D	128	✓	70.2	39.1	-
CoCLR [28]	Kinetics-400	S3D	128	✓	74.5	46.1	-
Vi^{2}CLR	UCF	S3D	128	✓	70.8	39.6	-
Vi^{2}CLR	Kinetics-400	S3D	128	✓	75.4	47.3	63.4
OPN [43]	UCF	VGG	227	✗	59.6	23.8	-
3D-RotNet [35]	Kinetics-400	R3D	112	✗	62.9	33.7	-
ST-Puzzle [37]	Kinetics-400	R3D	224	✗	63.9	33.7	-
VCOP [80]	UCF	R(2+1)D	112	✗	72.4	30.9	-
DPC [26]	Kinetics-400	R-2D3D	128	✗	75.7	35.7	-
CBT [68]	Kinetics-400	S3D	112	✗	79.5	44.6	-
DynamoNet [15]	Kinetics-400	STCNet	112	✗	88.1	59.9	-
SpeedNet [6]	Kinetics-400	S3D-G	224	✗	81.1	48.8	-
MemDPC [27]	Kinetics-400	R-2D3D	224	✗	86.1	54.5	-
AVTS [38]	Kinetics-400	I3D	224	✗	83.7	53.0	-
XDC [1]	Kinetics-400	R(2+1)D	224	✗	84.2	47.1	-
XDC [1]	IG65M	R(2+1)D	224	✗	94.2	67.4	-
GDT [53]	Kinetics-400	R(2+1)D	112	✗	89.3	60.0	-
MIL-NCE [45]	HTM	S3D	224	✗	91.3	61.0	-
ELO [55]	Youtube8M	R(2+1)D	224	✗	93.8	67.4	-
CVRL [56]	Kinetics-400	R3D-50	224	✗	92.2	66.7	70.4
CoCLR [28]	UCF	S3D	128	✗	81.4	52.1	-
CoCLR [28]	Kinetics-400	S3D	128	✗	87.9	54.6	-
Vi^{2}CLR	UCF	S3D	128	✗	82.8	52.9	-
Vi^{2}CLR	Kinetics-400	S3D	128	✗	89.1	55.7	71.2

Table 7. **Video Classification.** Comparison of self-supervised methods. All of the methods (except Weight Frozen: Supervised) has been trained with a self-supervised method and then fine-tuned on UCF101, HMDB51, and Kinetics-400. The ✓ means the encoder weights are frozen and a classification layer is trained, and ✗ means the entire encoder and a classification layer are fine-tuned.

a single objective function. We believe the process of joint training of video/image encoders, and cluster learning optimizes the contrastive objectives which lead to learning a robust and effective visual representation.

5. Conclusion

Recently self-supervised learning efforts like contrastive learning techniques have shown substantial progress in comparison to supervised pipelines. The community has witnessed the great impact of the self-supervised works on transfer learning as well. In this work, We have presented a joint self-supervised contrastive visual representation learning for videos and images. The method, Video/Image for Visual Contrastive Learning of Representation (Vi^{2}CLR) offers a complementary learning process with dynamic and static visual clues to learn semantic clusters and find similar instances in the representation space. Vi^{2}CLR broadly shows how different visual understanding ConvNets for downstream tasks like video action recognition, video retrieval, image and object classification can be benefited from robust and discriminative feature representation and a pre-training stage. The extensive evaluations on the proposed method prove state-of-the-art performances in the field of self-supervised learning for videos and images.

References

[1] Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016.

[6] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020.

[7] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. *arXiv preprint arXiv:1704.05310*, 2017.

[8] Uta Buchler, Biagio Brattoli, and Bjorn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–786, 2018.

[9] 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.

[10] 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.

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

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

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

[14] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. Spatio-temporal channel correlation networks for action classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–299, 2018.

[15] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6192–6201, 2019.

[16] 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.

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

[18] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

[19] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–929, 2020.

[20] 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*, 2010.

[21] 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.

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[23] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6391–6400, 2019.

[24] 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.

[25] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020.

[28] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[29] 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*, pages 9729–9738, 2020.

[30] 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.

[31] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua

Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[32] Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2016.

[33] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2733–2742, 2018.

[34] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. *arXiv preprint arXiv:1807.06653*, 2(3):8, 2018.

[35] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018.

[36] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[37] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019.

[38] 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.

[39] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[40] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020.

[41] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875*, 2019.

[42] Benjamin Laxton, Jongwoo Lim, and David Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[43] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[44] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[45] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[46] 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*, pages 6707–6717, 2020.

[47] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[48] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018.

[49] 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.

[50] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017.

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

[52] 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.

[53] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020.

[54] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014.

[55] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020.

[56] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[58] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2019.

[59] Vivek Sharma, Naila Murray, Diane Larlus, Saquib Sarfraz, Rainer Stiefelhagen, and Gabriela Csurka. Unsupervised meta-domain adaptation for fashion retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1348–1357, 2021.

[60] Vivek Sharma, Saquib Sarfraz, and Rainer Stiefelhagen. A simple and effective technique for face clustering in tv series. In *CVPR workshop on Brave New Motion Representations*, 2017.

[61] Vivek Sharma, Makarand Tapaswi, M Saquib Sarfraz, and Rainer Stiefelhagen. Self-supervised learning of face representations for video face clustering. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[62] Vivek Sharma, Makarand Tapaswi, M Saquib Sarfraz, and Rainer Stiefelhagen. Video face clustering with self-supervised representation learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):145–157, 2019.

[63] Vivek Sharma, Makarand Tapaswi, M Saquib Sarfraz, and Rainer Stiefelhagen. Clustering based contrastive learning for improving face representations. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2020.

[64] Vivek Sharma, Makarand Tapaswi, and Rainer Stiefelhagen. Deep multimodal feature encoding for video ordering. *arXiv preprint arXiv:2004.02205*, 2020.

[65] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[66] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[67] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 625–634, 2020.

[68] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

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

[70] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8287–8296, 2019.

[71] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016.

[72] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018.

[73] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. *arXiv preprint arXiv:2008.05861*, 2020.

[74] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[75] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[76] 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.

[77] 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*, pages 3733–3742, 2018.

[78] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[79] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.

[80] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[81] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 6210–6219, 2019.

[82] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[83] 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.

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

[85] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015.