Probabilistic Representations for Video Contrastive Learning

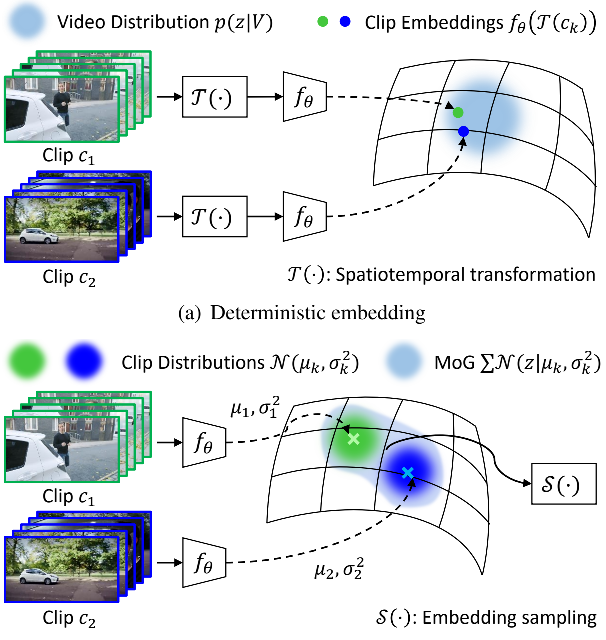
Jungin Park[1] Jiyoung Lee[2] Ig-Jae Kim[3] Kwanghoon Sohn[1]*
[1]Yonsei University [2]NAVER AI Lab [3]Korea Institute of Science and Technology (KIST)
{newrun, khsohn}@yonsei.ac.kr lee.j@navercorp.com

Abstract

*This paper presents **Pro**babilistic **Vi**deo **Co**ntrastive Learning, a self-supervised representation learning method that bridges contrastive learning with probabilistic representation. We hypothesize that the clips composing the video have different distributions in short-term duration, but can represent the complicated and sophisticated video distribution through combination in a common embedding space. Thus, the proposed method represents video clips as normal distributions and combines them into a Mixture of Gaussians to model the whole video distribution. By sampling embeddings from the whole video distribution, we can circumvent the careful sampling strategy or transformations to generate augmented views of the clips, unlike previous deterministic methods that have mainly focused on such sample generation strategies for contrastive learning. We further propose a stochastic contrastive loss to learn proper video distributions and handle the inherent uncertainty from the nature of the raw video. Experimental results verify that our probabilistic embedding stands as a state-of-the-art video representation learning for action recognition and video retrieval on the most popular benchmarks, including UCF101 and HMDB51.*

1. Introduction

Video is the vitality of the Internet, which means that understanding video content is essential for the most modern artificial intelligence (AI) agents. Alongside this, learning enriched spatiotemporal representations from *unlabeled videos* (*i.e*., self-supervised or unsupervised video representation learning) [59, 61, 62] has become a crucial research topic for the computer vision community. The interest in this topic is to learn deep features representing general visual contents, which has proven essential to improving performance on downstream tasks such as action recognition [10, 28, 82], action detection [86, 88], video retrieval [22, 80], and even multi-modal event recognition [3, 55]. However, self-supervised video representation learning has still remained challenging due to the inherent difficulty caused by the nature of the videos in comparison to static images.


Video Distribution $p(z|V)$
Clip Embeddings $f_\theta(\mathcal{T}(c_k))$
$\mathcal{T}(\cdot)$
f_θ
Clip c_1
Clip c_2
$\mathcal{T}(\cdot)$: Spatiotemporal transformation
Clip Distributions $\mathcal{N}(\mu_k, \sigma_k^2)$
MoG $\sum\mathcal{N}(z|\mu_k, \sigma_k^2)$
μ_1, σ_1^2
μ_2, σ_2^2
$\mathcal{S}(\cdot)$
$\mathcal{S}(\cdot)$: Embedding sampling

(a) Deterministic embedding

(b) Probabilistic embedding (**Ours**)

Figure 1. Contrary to (a) the deterministic point embedding methods, which estimate the subset of the video distribution, (b) the proposed method estimates the whole video distribution through the mixture of probabilistic embeddings.

Recent breakthroughs in self-supervised video representation learning have been developed with two different branches: (1) leveraging pretext tasks related to the coherence of videos and (2) using contrastive learning [26] for instance discrimination. Specifically, video coherence is empirically modeled through sub-properties of video contents associated with temporal ordering [21, 34, 45, 54, 83], optical flow [24], spatiotemporal statistics [39,51,75], and playback rate [76, 85]. Even though they have shown that spatiotemporal representations can be learned from unlabeled

*Corresponding author.

This research was supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA(NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289) and the Yonsei University Research Fund of 2021 (2021-22-0001).

videos, the learned representations inevitably contain task-specific information.

In contrast, instance discrimination methods [20, 47, 49, 59, 62, 68] have attempted to learn video representations by incorporating contrastive learning [26], which aims to discriminate different instances without using sub-properties of data [18, 81]. Concretely, to learn spatiotemporal representations from videos, existing works treat each video as an "instance" and embed video clips to deterministic points on the embedding space, as shown in Fig. 1(a). Based on contrastive objectives [15,63,77,79], positive point pairs are pulled together and negative point pairs are pushed away. The positive pairs are composed of clips from the same video [62] or different views (augmented versions) of the same clip [20, 59], and the negative pairs are composed of clips from different videos. To make such training pairs, several works have introduced carefully designed spatial and temporal transformations in the form of data augmentation, including a temporal mask [59] and temporally consistent spatial augmentation [62].

However, deterministic representations for video contrastive learning have critical limitations in three respects: First, representing the complicated and sophisticated video distribution as a set of deterministic points is insufficient to learn discriminative video representations. Unlike static images, videos are a collection of noisy temporal dynamics and contain a lot of redundant information, that makes the uncertainty of data high [52]. Therefore, an alternative to deterministic representations is required to describe overall video distribution. Second, improper sampling and transformation techniques to generate different views can cause performance fluctuation according to downstream tasks [89]. Moreover, improper temporal transformations (*e.g.* shuffling) that can harm the video contents weaken the effectiveness of contrastive learning [5]. Third, they often neglect common components that are likely to contain valid correspondences between semantically adjacent instances (*e.g.* same category, but different videos), leading to limited discrimination performance of learned representations, as demonstrated in [27].

To overcome these limitations while maximizing the advantages of contrastive learning, we propose probabilistic representations for video contrastive learning, named **ProViCo**, in which video clips are represented as random variables in a stochastic embedding space. As shown in Fig. 1(b), clips sampled from a video are represented as distinct normal distributions and the distribution of the whole video is approximated by a Mixture of Gaussians (MoG) of clip distributions. We construct the positive and negative pairs based on the probabilistic distance between embeddings sampled from each video distribution. Moreover, we propose an uncertainty-based stochastic contrastive loss that incorporates uncertainty (*i.e.*, the inherent noise of videos) into the soft contrastive loss [57]. By leveraging uncertainty, we can reduce the effect of noisy samples or improper training pairs on self-supervised representation learning, and can make useful applications such as estimation of difficulty or chance of failure on test data.

To sum up, our contributions are as follows: (1) We propose a novel ProViCo to effectively represent the probabilistic video embedding space. To the best of our knowledge, this is the first attempt to leverage probabilistic embeddings for self-supervised video representation learning. (2) We introduce the probabilistic distance-based positive mining to exploit semantic relations between videos and present the stochastic contrastive loss to weaken the adverse impact of unreliable instances. (3) We demonstrate the effectiveness of the proposed probabilistic approach through the uncertainty estimation and extensive experiments on downstream tasks, including action recognition and video retrieval.

2. Related Work

Self-supervised video representation learning. Early works on self-supervised video representation learning have been studied with pretext tasks to exploit the spatiotemporal cues, including prediction of motion and appearance [75], spatiotemporal transformations [34, 39, 54, 83], frames [51], and playback rate [76, 85]. Recently, contrastive learning methods with instance discrimination tasks have been proposed for video representations [20, 59, 62]. The popular self-supervised visual representation learning frameworks [9, 13, 25, 30] have been transformed to empower the temporal robustness of the encoder for video representations, improving momentum contrastive learning [20,59]. Further, motion estimation [47], temporal relations [32], multi-level feature optimization [61], and meta-learning framework [49] have been incorporated into contrastive learning. While several works have employed additional signals (*e.g.* audio [19, 42, 52, 58, 60, 78] and optical flow [24, 27, 84]) to improve the performance, we focus on RGB-only self-supervised video representation learning without growth of training costs following [20, 59, 61].

Probabilistic representation. Learning representations in a stochastic embedding space has first been proposed for word embeddings [72]. Thanks to the high robustness of handling the inherent hierarchies of the language of probabilistic embeddings, they have been extensively explored in natural language processing [48, 56]. The probabilistic embeddings for vision tasks have been introduced for face recognition [11, 65], speaker diarization [66], human pose estimation [69], and prototype embeddings for few-shot recognition [64]. In recent years, hedged instance embedding (HIB) [57] has been proposed to learn probabilistic embeddings based on the variational information bottleneck (VIB) principle [1, 2]. The soft contrastive loss is formulated as a probabilistic alternative to contrastive loss to handle the one-to-many mapping. With the HIB objectives, probabilistic cross-modal embeddings [16] have been studied to learn joint embeddings between images and captions for one-to-many image-text retrieval. In contrast with [16, 57] trained in a supervised manner using image-text/label pairs, we learn probabilistic representations by only self-supervision without any labels. To this end, we propose a novel stochastic contrastive loss that is suited for self-supervised learning to optimize the model according to data uncertainty.

Uncertainty in Computer Vision. As a method for improving the interpretability and the robustness to input data of deep neural networks, uncertainty have been extensively studied for a long time [8, 23, 37]. In general, uncertainty is categorized into two types by different sources: (1) epistemic uncertainty (*i.e.*, model uncertainty) which indicates uncertainty in the model parameters and (2) aleatoric uncertainty (*i.e.*, data uncertainty) that originated from the inherent noise of data. Epistemic uncertainty can be reduced by providing enough training data [8, 36], whereas aleatoric uncertainty cannot be eliminated with additional training data [37]. In computer vision, uncertainty has been explored for various tasks such as semantic segmentation [36, 37], object detection [14, 43], person re-identification [87], and face recognition [11, 38, 65]. Although some works [11, 37, 65] have considered the aleatoric uncertainty, they cannot be directly applied for self-supervised learning due to the absence of label information. Specific to deep video understanding, uncertainty has been used for video instance segmentation [53], weakly-supervised temporal action localization [46], and video future frame prediction [12]. They have mainly focused on the predictive uncertainty estimated from the output of the model. In this work, we explore the aleatoric uncertainty of videos for self-supervised video contrastive learning.

3. Method

3.1. Background and Motivation

Contrastive learning is a promising framework to learn video representations in a self-supervised manner. Given a fixed-length clip q from a video $\mathcal{V}$, let $\{k_+\}$ be a set of positives sampled from the same video as q (or augmented versions of q) and $\{k_-\}$ is a set of negatives sampled from other instances in a batch. The goal of contrastive learning is to maximize the similarities between q and $\{k_+\}$ and minimize similarities between q and $\{k_-\}$ through contrastive loss, such as the InfoNCE [71]:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\sum_{k\in\{k_+\}} \exp(\text{sim}(q,k)/\tau)}{\sum_{k\in\{k_+,k_-\}} \exp(\text{sim}(q,k)/\tau)}, \quad (1)$$

where τ is a scaling temperature parameter and $\text{sim}(\cdot,\cdot)$ is a similarity function. The contrastive loss improves instance discrimination power by formulating relative distance between instances in the dynamic dictionaries [30] instead of matching an input to a fixed target. However, as investigated in [27], current approaches have focused on instance discrimination by treating each data sample as a “class”, which makes the model neglect the semantic relations between different videos. We ascribe this to the risk of unstable matching arising from the lack of tools to measure the confidence of proximity between unlabeled instances during training.

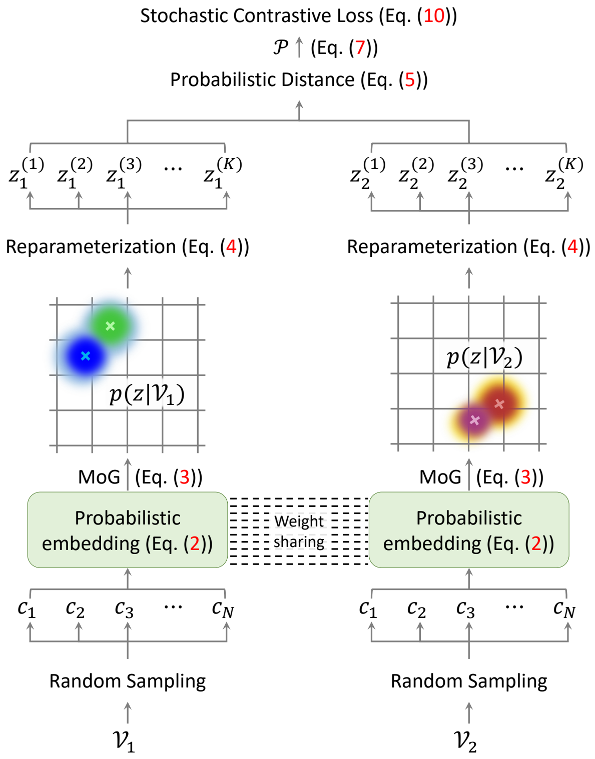

Stochastic Contrastive Loss (Eq. (10))
$\mathcal{P}$ ↑ (Eq. (7))
Probabilistic Distance (Eq. (5))
$z_1^{(1)}$ $z_1^{(2)}$ $z_1^{(3)}$ ⋯ $z_1^{(K)}$
$z_2^{(1)}$ $z_2^{(2)}$ $z_2^{(3)}$ ⋯ $z_2^{(K)}$
Reparameterization (Eq. (4))
Reparameterization (Eq. (4))
$p(z|\mathcal{V}_1)$
$p(z|\mathcal{V}_2)$
MoG (Eq. (3))
MoG (Eq. (3))
Probabilistic embedding (Eq. (2))
Weight sharing
Probabilistic embedding (Eq. (2))
c_1 c_2 c_3 ⋯ c_N
c_1 c_2 c_3 ⋯ c_N
Random Sampling
Random Sampling
$\mathcal{V}_1$
$\mathcal{V}_2$

Figure 2. **ProViCo** estimates the video distribution $p(z|\mathcal{V})$ as a Mixture of N Gaussians with probabilistic clip embeddings. We construct the positive and negative pairs based on the probabilistic distance between two video distributions. The model learns probabilistic embedding network parameters and minimizes uncertainties of input videos through the stochastic contrastive loss.

As a feasible solution, we propose the Probabilistic Representations for Video Contrastive Learning (ProViCo), in which videos are represented as probability distributions in a stochastic embedding space. In our framework, the uncertainty of videos is used as a key tool to measure the confidence of proximity between video distributions. Next we first describe the probabilistic embedding for video clips and extend it to the whole representation learning using a proposed stochastic contrastive loss. The overall procedure of ProViCo is illustrated in Fig. 2.

3.2. Probabilistic Video Embedding

Given a video $\mathcal{V}$, let $\{c_1, ..., c_N\}$ be a set of clips sampled from $\mathcal{V}$, and $v_{c_n} = f_\theta(c_n)$ represents the output of the backbone network (*i.e.*, encoder) parameterized by θ. We formulate a probability distribution $p(z|c_n)$ for a clip c_n as a normal distribution with a mean vector and a diagonal covariance matrix in a stochastic embedding space $\mathbb{R}^D$:

$$p(z|c_n) \sim \mathcal{N}(g_\mu(v_{c_n}), \text{diag}(g_\sigma(v_{c_n}))), \tag{2}$$

where g_μ is a fully-connected (FC) layer followed by LayerNorm [4] and ℓ_2 normalization, and g_σ is a separate FC layer without any normalization, following [16]. With N clip distributions, we represent the whole video distribution $p(z|\mathcal{V})$ as a Mixture of Gaussians [57] such that

$$p(z|\mathcal{V}) \sim \sum_{n=1}^{N} \mathcal{N}(z; g_\mu(v_{c_n}), \text{diag}(g_\sigma(v_{c_n}))). \tag{3}$$

From $p(z|\mathcal{V})$, we sample K embeddings $\{z^{(1)}, ..., z^{(K)}\} \overset{\text{iid}}{\sim} p(z|\mathcal{V})$, that represent "self-augmented" versions of the video representation. Specifically, we use reparameterization trick [41] for stable training, such that

$$z^{(k)} = \sigma(\mathcal{V}) \cdot \epsilon^{(k)} + \mu(\mathcal{V}), \tag{4}$$

where $\mu(\mathcal{V})$, $\sigma(\mathcal{V})$ are the mean and the standard deviation of $p(z|\mathcal{V})$, and $\{\epsilon^{(1)}, ..., \epsilon^{(K)}\}$ are sampled iid from the D-dimensional unit Gaussian distribution.

3.3. Mining Positive and Negative Pairs

Contrary to the deterministic representation methods [20, 59], we construct the positive and negative pairs based on the probabilistic distance that contains the uncertainty of embedded distributions. Specifically, given a embedding pair $z_i^{(k)} \sim p(z|\mathcal{V}_i)$ and $z_j^{(k')} \sim p(z|\mathcal{V}_j)$ sampled from i-th and j-th video distributions in a batch, we define the distance between two embeddings as Bhattacharyya distance [7]:

$$\begin{aligned} \text{dist}(z_i^{(k)}, z_j^{(k')}) =& \frac{1}{4}(\log(\frac{1}{4}(\frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} + 2)) \\ &+ \lambda \cdot \frac{(z_i^{(k)} - z_j^{(k')})^\top (z_i^{(k)} - z_j^{(k')})}{\sigma_i^2 + \sigma_j^2}), \end{aligned} \tag{5}$$

where σ_i^2, σ_j^2 are variances for the i-th and j-th video distributions that represents the uncertainty of each video, and λ is a scaling factor according to the dimension of the embedding space. The distance between two video distributions can be factorized via Monte-Carlo estimation:

$$\text{dist}(\mathcal{V}_i, \mathcal{V}_j) \approx \frac{1}{K^2} \sum_{k}^{K} \sum_{k'}^{K} \text{dist}(z_i^{(k)}, z_j^{(k')}). \tag{6}$$

The positive pairs $\mathcal{P}$ are defined as all video pairs closer than threshold distance τ, such that

$$\mathcal{P} = \{(\mathcal{V}_i, \mathcal{V}_j) \mid \text{dist}(\mathcal{V}_i, \mathcal{V}_j) < \tau \text{ or } i = j\}. \tag{7}$$

Since we regard each embedding $z_i^{(k)}$ as a self-augmented version of $\mathcal{V}_i$, we also set a pair $(\mathcal{V}_i, \mathcal{V}_i)$ as a positive. The negative pairs are then the complement of the positive pairs (*i.e.*, $\bar{\mathcal{P}}$). By defining positive and negative pairs based on the probabilistic distance, we can construct the confident sample pairs considering the uncertainty of videos.

3.4. Stochastic Contrastive Loss

As an alternative to conventional contrastive objectives such as (1), we introduce a stochastic contrastive loss to discriminate positive and negative pairs, and minimize the uncertainty of videos, simultaneously. The stochastic contrastive loss incorporates the uncertainty of each video into the soft contrastive loss [57] that transformed contrastive loss for probabilistic embeddings. For a pair of videos $(\mathcal{V}_i, \mathcal{V}_j)$, the soft contrastive loss is formulated by

$$\mathcal{L}_{\text{soft}}(\mathcal{V}_i, \mathcal{V}_j) = \begin{cases} -\log p(m|\mathcal{V}_i, \mathcal{V}_j) & \text{if } (\mathcal{V}_i, \mathcal{V}_j) \in \mathcal{P} \\ -\log(1 - p(m|\mathcal{V}_i, \mathcal{V}_j)) & \text{otherwise} \end{cases}, \tag{8}$$

where $p(m|\mathcal{V}_i, \mathcal{V}_j)$ is the match probability [16,57] with the sigmoid function $s(\cdot)$ and learnable scalars (a, b):

$$p(m|\mathcal{V}_i, \mathcal{V}_j) = \frac{1}{K^2} \sum_{k}^{K} \sum_{k'}^{K} s(-a||z_i^{(k)} - z_j^{(k')}||_2 + b). \tag{9}$$

Finally, the stochastic contrastive loss between $\mathcal{V}_i$ and $\mathcal{V}_j$ is defined as:

$$\mathcal{L}_{\text{stoc}}(\mathcal{V}_i, \mathcal{V}_j) = \frac{1}{4\sigma_i^2\sigma_j^2} \mathcal{L}_{\text{soft}}(\mathcal{V}_i, \mathcal{V}_j) + \frac{1}{2}(\log \sigma_i^2 + \log \sigma_j^2), \tag{10}$$

where the first term is for the instance discrimination between the probabilistic embeddings obtained with the model and the seconde term is a regularization term to prevent the model from predicting infinite uncertainty for all videos. Two terms complement each other to control the contribution of unreliable pairs and uncertainties. More concretely, the probabilistic distance in (5) is decreased for the video pair with substantial uncertainties, such that improper positive pairs can be constructed. However, high uncertainty (*i.e.*, large $\sigma_i^2\sigma_j^2$) attenuates the contribution of the first term in the stochastic contrastive loss, penalizing unreliable pairs, and exaggerates the second term that reduces uncertainties. For pairs with low uncertainty, the model will focus on instance discrimination, as the first term have the larger contribution. These properties of the stochastic contrastive loss make the model robust to noisy videos.

3.5. Total Objectives

We employ the additional KL regularization term between the video distribution and the unit Gaussian prior $\mathcal{N}(0, I)$ to prevent the predicted variance from collapse to zero, following [16]:

$$\begin{aligned} \mathcal{L}_{\text{KL}}(\mathcal{V}_i, \mathcal{V}_j) = & \text{KL}(p(z_i|\mathcal{V}_i)||\mathcal{N}(0, I)) \\ & + \text{KL}(p(z_j|\mathcal{V}_j)||\mathcal{N}(0, I)), \end{aligned} \tag{11}$$

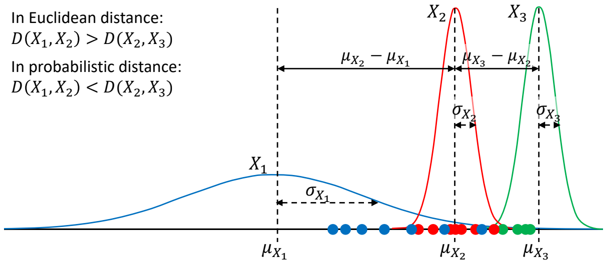


Figure 3. **1-dimensional toy example for the distance between probability distributions.** The Euclidean distance between the mean values without considering variances cannot represent the probabilistic similarity between probability distributions.

Therefore, the overall objective for ProViCo is a weighted sum of all loss functions defined as:

$$\mathcal{L}_{\text{ProViCo}} = \mathcal{L}_{\text{stoc}} + \beta \cdot \mathcal{L}_{\text{KL}}, \tag{12}$$

where $\mathcal{L}_{\text{stoc}}$ and $\mathcal{L}_{\text{KL}}$ are the sum of each term for all pairs in a batch, and β controls the trade-off between two terms.

3.6. Rethinking Objectives with Distance

In our framework, the model learns to achieve two main objectives: instance discrimination in a stochastic embedding space and uncertainty minimization for input videos. For robust instance discrimination learning, we argue that the semantic relations beyond the instances are to be examined for mining positive and negative pairs. Although the conventional distance metric (*e.g.* Euclidean distance or cosine similarity) can measure the similarity between instances using the mean of distributions, it is inadequate to represent the probabilistic similarity due to the variance of each distribution. For example, as illustrated in Fig. 3, the higher similarity between probability distributions is not always guaranteed by the smaller Euclidean distance. From this observation, we determine the positive and negative pairs based on the probability distance between video distributions, as described in Sec. 3.3. On the other side, directly optimizing the probability distance to discriminate semantic instances may lead to unexpected results. Namely, minimizing the probability distance between positive pairs without any constraints can lead to an increase in the uncertainty of videos, decreasing the discrimination power of learned representations. To address this issue, our stochastic contrastive loss is designed to optimize the indirect probabilistic distance incorporating uncertainty and soft contrastive loss, which utilizes Euclidean distance. With such carefully designed training objectives, the model accomplish both discrimination of semantic instances and uncertainty minimization.

4. Experiments

4.1. Datasets

Kinetics-400 [35] (K400) dataset contains ∼240k training videos of 400 human action categories. The test set consists of ∼38k videos, with about 100 videos for each class. We use the whole training videos to obtain initial parameters by pretraining the network using the proposed method. We measure action recognition performance on the test set using linear evaluation.

Model	Backbone (# params)	Acc. (%)
Supervised	R3D-50 (31.8M)	74.7
Supervised	R3D-18 (20.2M)	68.9
Supervised	R(2+1)D (15.4M)	71.7
SeCo [84]	ResNet-50 (23.0M)	61.9
CVRL [62]	R3D-50 (31.8M)	66.1
CVRL [62]	R3D-101 (51.4M)	67.6
ρBYOL [20]	R3D-50 (31.8M)	68.3
ρMoCo [20]	R(2+1)D (15.4M)	57.2
CORP [32]	R3D-50 (31.8M)	66.3
ProViCo (Ours)	R3D-18 (20.2M)	62.8
ProViCo (Ours)	R(2+1)D (15.4M)	65.7

Table 1. Linear evaluation for action recognition on the Kinetics-400 [35] dataset corresponding to the backbone networks. The models are pretrained on Kinetics-400. We shaded the considerably different experimental settings in terms of the backbone network, frame resolution, and batch size.

UCF101 [67] dataset consists of ∼13k videos of 101 action categories. To compare previous works [59, 61], we perform pretraining on train split 1 and evaluate action recognition on test split 1 using two evaluation protocols (described in Sec. 4.2). The video retrieval performance is evaluated on test split 1 using nearest-neighbors.

HMDB51 [44] is a relatively small dataset, containing ∼7k videos of 51 action categories. Among the three splits, we use train split 1 for finetuning and measure action recognition and video retrieval performance on test split 1.

4.2. Experimental Setup

Pretraining. For pretraining, we randomly sample two 16-frame clip with the temporal stride of 1 from each video. All frames in each clip are fixed to a size of 112 × 112 by random cropping. The backbone networks are trained for 200 epochs with a mini-batch size of 96. By using a half-period cosine learning rate scheduler [50], we warm-up the learning rate in the first 20 epochs from an initial learning rate of 10^{-4} with Adam optimizer [40]. We set the scaling factor of the probabilistic distance λ in (5) to $1/4D$ according to the embedding space size D and threshold distance τ in (7) to 0.15. The KL-divergence hyperparameter β is set to 10^{-4} and the number of embeddings K is set to 10 throughout the experiments.

Backbone networks. To provide a fair comparison, we employ two popular 3D networks as backbone networks: R3D-18 [28, 29] and R(2+1)D-18 [70]. These architectures have model parameters of 20.2M and 15.4M, respectively, which are much lighter compared to models such as R3D-50 or R3D-101. For action recognition, we evaluate the performance using both backbone networks. For video retrieval

Type	Model	Backbone (# params)	Input size	Batch size	Pretrain data	Finetune	UCF	HMDB
(ii)	MFO [61]	R3D-18 (20.2M)	112 × 112	256	K400	✗	63.2	33.4
(ii)	CATE [68]	R3D-50 (31.7M)	224 × 224	1024	K400	✗	84.3	53.6
(iii)	CORP [32]	R3D-50 (31.7M)	224 × 224	512	K400	✗	90.2	58.7
(ii)	**ProViCo (Ours)**	R3D-18 (20.2M)	112 × 112	96	K400	✗	82.9	52.2
(ii)	**ProViCo (Ours)**	R(2+1)D (15.4M)	112 × 112	96	K400	✗	84.1	53.5
(i)	DSM [73]	I3D (25.0M)	224 × 224	128	K400	✓	74.8	52.5
(ii)	CVRL [62]	R3D-50 (31.7M)	224 × 224	1024	K400	✓	92.2	66.7
(ii)	VideoMoCo [59]	R(2+1)D (15.4M)	112 × 112	128	K400	✓	78.7	49.2
(ii)	MFO [61]	R3D-18 (20.2M)	112 × 112	256	K400	✓	79.1	47.6
(ii)	CATE [68]	R3D-50 (31.7M)	128 × 128	1024	K400	✓	88.4	61.9
(ii)	MCN† [49]	R3D-18 (20.2M)	128 × 128	80	K400	✓	89.7	59.3
(i)	TCLR [17]	R(2+1)D (15.4M)	112 × 112	40	K400	✓	84.3	54.2
(iii)	TEC [33]	R(2+1)D (15.4M)	112 × 112	192	K400	✓	87.1	59.8
(ii)	**ProViCo (Ours)**	R3D-18 (20.2M)	112 × 112	96	K400	✓	85.6	58.4
(ii)	**ProViCo (Ours)**	R(2+1)D (15.4M)	112 × 112	96	K400	✓	87.2	59.4
(i)	VCP [51]	C3D (34.8M)	112 × 112	-	UCF101	✓	68.5	32.5
(i)	PRP [85]	R(2+1)D (15.4M)	112 × 112	-	UCF101	✓	72.1	35.0
(i)	RTT [34]	R(2+1)D (15.4M)	112 × 112	256	UCF101	✓	81.6	46.4
(i)	DSM [73]	C3D (34.8M)	112 × 112	128	UCF101	✓	70.3	40.5
(ii)	MFO [61]	R3D-18 (20.2M)	112 × 112	256	UCF101	✓	76.2	41.1
(ii)	MCN† [49]	R3D-18 (20.2M)	128 × 128	80	UCF/HMDB	✓	85.4	54.8
(iii)	CORP [32]	R3D-50 (31.7M)	224 × 224	512	UCF/HMDB	✓	93.5	68.0
(ii)	TCLR [17]	R(2+1)D (15.4M)	112 × 112	40	UCF101	✓	82.8	53.6
(iii)	TEC [33]	R(2+1)D (15.4M)	112 × 112	192	UCF101	✓	85.2	56.9
(ii)	**ProViCo (Ours)**	R3D-18 (20.2M)	112 × 112	96	UCF101	✓	83.7	57.1
(ii)	**ProViCo (Ours)**	R(2+1)D (15.4M)	112 × 112	96	UCF101	✓	86.1	58.0
(ii)	**ProViCo (Ours)**	R3D-50 (31.7M)	224 × 224	512	UCF101	✓	94.6	68.2

Table 2. Action recognition performance on UCF101 [67] and HMDB51 [44] dataset corresponding to the pretrained dataset and backbone networks. **Finetune** ✓ means the whole networks are finetuned end-to-end, while ✗ means the backbone network is fixed and the linear classifier is updated only. We shaded the considerably different experimental settings in terms of the backbone network, frame resolution, and batch size. † denotes that additional residual views are used [49].

and ablation studies, we report only the results of R3D-18.

Evaluation protocols. We evaluate the proposed method for action recognition and video retrieval tasks. Following previous works [20, 62], we adopt two evaluation protocols to verify the learned video representations: (i) *Linear evaluation* provides a straightforward evaluation for learned representations by fixing all the parameters in the backbone network and finetuning only the fully-connected (FC) layers. (ii) *Finetuning* updates parameters in both the pretrained backbone and the additional FC layers. For action recognition, we pretrain the backbone network on the Kinetics-400 [35] and UCF101 [67] datasets, respectively. We report the top-1 accuracy evaluated on the Kinetics-400 and UCF101 datasets using two evaluation protocols. In addition, we evaluate the performance on the HMDB51 [44] dataset using finetuning. For video retrieval, we measure top-1, 5, 10, and 20 accuracies using nearest-neighbors without additional training and compare with state-of-the-art methods on UCF101 and HMDB51 datasets.

4.3. Action Recognition

We first compare the action recognition performance of ProViCo with state-of-the art methods. In our framework, the action of the video is predicted by averaging the output probabilities of the classifier for all embeddings sampled in (4). We observe that the performance is significantly affected by the architecture of backbone networks, the video frame resolution, and the batch size used during training. Since we set the minimum level of these components, we mainly compare the results with similar conditions.

Linear evaluation on K400. We report the linear evaluation results on Kinetics-400 [35] in Tab. 1. The first three rows represent the results of supervised learning for each backbone network, showing the significant performance gap according to the network. The comparison between our method and ρMoCo [20] shows that our probabilistic approach outperforms the deterministic approach by a large margin (8.5% performance gain with the same backbone networks). Compared to methods (CVRL [62], ρBYOL [20] and CORP [32]) using about twice as many network parameters as R(2+1)D, our method attains competitive performance, further reducing the gap between self-supervised and supervised learning.

Linear evaluation on UCF101 and HMDB51. We provide the linear evaluation results on UCF101 [67] and

Method	Backbone (# params)	Pretrain	UCF101				HMDB51			
			R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
SpeedNet [6]	S3D-G (9.1M)	K400	13.0	28.1	37.5	49.5	-	-	-	-
MFO [61]	R3D-18 (20.2M)	K400	41.5	60.6	71.2	80.1	20.7	40.8	55.2	68.3
CATE [68]	R3D-50 (31.7M)	K400	54.9	68.3	75.1	82.3	33.0	56.8	69.4	82.1
TEC [33]	R3D-18 (20.2M)	K400	66.9	**83.1**	88.8	93.3	36.4	**64.1**	74.1	83.8
ProViCo (Ours)	R3D-18 (20.2M)	K400	**67.6**	81.4	**90.1**	**94.7**	**40.1**	60.6	**75.2**	**85.2**
VCP [51]	R3D-18 (20.2M)	UCF101	18.6	33.6	42.5	53.5	7.6	24.4	36.3	53.6
Pace [76]	R3D-18 (20.2M)	UCF101	23.8	38.1	46.4	56.6	9.6	26.9	41.1	56.1
PRP [85]	R3D-18 (20.2M)	UCF101	22.8	38.5	46.7	55.2	8.2	25.8	38.5	53.3
DSM [73]	I3D (25.0M)	UCF101	17.4	35.2	45.3	57.8	7.6	23.3	36.5	52.5
STS [74]	R3D-18 (20.2M)	UCF101	38.3	59.9	68.9	77.2	18.0	37.2	50.7	64.8
MFO [61]	R3D-18 (20.2M)	UCF101	39.6	57.6	69.2	78.0	18.8	39.2	51.0	63.7
MCN [49]†	R3D-18 (20.2M)	UCF101	53.8	70.2	78.3	83.4	24.1	46.8	59.7	74.2
TCLR [17]	R3D-18 (20.2M)	UCF101	56.2	72.2	79.0	85.3	22.8	45.4	57.8	73.1
TEC [33]	R3D-18 (20.2M)	UCF101	62.5	**78.4**	84.1	88.8	32.0	**60.8**	72.2	81.7
ProViCo (Ours)	R3D-18 (20.2M)	UCF101	**63.8**	75.1	**84.8**	**89.2**	**35.9**	55.2	**74.3**	**81.8**

Table 3. Performance comparisons for video retrieval task evaluated on UCF101 [67] and HMDB51 [44] datasets. We report top-1, 5, 10, 20 accuracies. † denotes that additional residual views are used [49].

HMDB51 [44] datasets in the first block of Tab. 2. The models are pretrained on Kinetics-400 dataset and evaluated on each dataset using linear evaluation. The results show that our method outperforms MFO [61] with about ×2.5 smaller batch size, showing 19.7% and 18.8% performance improvements on UCF101 and HMDB51 datasets, respectively. In addition, our method provides comparable performance to CATE [68] and CORP [32], even though they used ×2 more parameters, ×2 larger frame resolution, and ×5 and ×10 larger batch size, respectively.

Finetuning on UCF101 and HMDB51. In the second and third blocks of Tab. 2, we report the comparison results for finetuning evaluation protocol. While the methods in the second blocks are pretrained on the Kinetics-400, the methods in the third blocks are pretrained on the UCF101 which is more smaller dataset than Kinetics-400. We analyze the results by dividing previous approaches into three aspects: (i) pretext tasks without contrastive learning [34,51,73,85]; (ii) contrastive learning [17, 49, 59, 61, 62, 68]; (iii) contrastive learning with pretext tasks using additional classifier branches [32, 33]. First of all, the results show that our method significantly outperforms methods of pretext tasks [34, 51, 73, 85] without contrastive learning on both pretraining datasets, regardless of the backbone network, the frame resolution, and the batch size. In comparison with contrastive learning approaches [17,49,59,61,62,68], our method achieves significant improvements, except for [32, 62, 68], which use a much larger batch size and a considerably deeper architecture with higher computational requirements. While MCN [49] used additional residual views with RGB view, our method shows competitive performances and even outperforms when the network pretrained on UCF101. The results of TEC [33] pretrained on Kinetics-400 show that the deterministic approach achieves a similar performance to our probabilistic approach by combining contrastive learning and pretext tasks that use additional parameters, as mentioned in [30]. However, our method demonstrates robustness to the small number of training videos in the results pretrained on UCF101, showing 0.9% and 1.4% performance degradation on each dataset, while 1.9% and 2.9% degradation in [33]. With the deeper architecture (*i.e.* R3D-50) and a larger batch size, our method achieves state-of-the-art performance, outperforming CORP [32] by 1.1% and 0.2% on UCF101 and HMDB51, respectively.

4.4. Video Retrieval

Tab. 3 presents the video retrieval performance on UCF101 and HMDB51 datasets according to the pretraining dataset. For the video retrieval, we compute nearest-neighbors using the match probability in (9) between two videos with Monte-Carlo estimation, unlike prior works, which used cosine similarity. We also provide the results using cosine similarity in the supplementary material. Our model achieves significantly improved top-1 accuracy performance on overall experimental results regardless of the network architecture [68] and input data [49]. The results also show the advantage of learning probabilistic video representations over utilizing contrastive learning to learn deterministic video representations, as in previous works [17,49,61,68]. We ascribe these results to our probabilistic approach, which learns probabilistic distributions of data and utilizes hard positive pairs so that learned representations are more suitable for matching tasks.

4.5. Ablation Study and Analysis

To further validate and fully investigate the components of our method, we conduct ablation experiments for action recognition according to the value of the KL-divergence hyperparameter β and the number of sampled embeddings

KL-divergence hyperparameter ($K = 10$, $N = 2$)				
Parameter β	10^{-5}	10^{-4}	10^{-3}	10^{-2}
Acc. (%)	-	**83.7**	83.4	81.6
Number of sampled embeddings ($\beta = 10^{-4}$, $N = 2$)				
Parameter K	5	7	10	12
Acc. (%)	81.2	83.1	**83.7**	-

Table 4. Ablation studies for KL-divergence hyperparameter β and the number of sampled embeddings K.

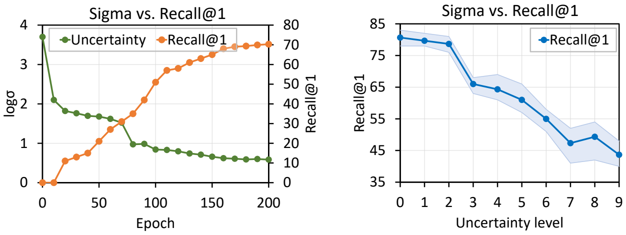


Figure 4. (**Left**) **Uncertainty versus performance** during training. (**Right**) **Performance versus uncertainty level** for test set.

K. In addition, we provide the uncertainty analysis to verify the impact of the uncertainty on representation learning. For all experiments in this section, we use R3D-18 as the backbone network and batch size is fixed to 96. Note that the backbone network is pretrained and evaluated on UCF101 [67] using finetuning for action recognition and nearest-neighbors for video retrieval.

KL-divergence hyperparameter. We report the action recognition performance in accuracy according to the value of the hyperparameter β in (12) to explore the effect of the KL-divergence regularization. As shown in the first block of Tab. 4, the maximum performance is obtained with $\beta = 10^{-4}$. Formally, an increase in β yields that the variance of the embedding is learned close to the unit variance, reducing the discriminability of distributions. On the contrary, when β is too small (10^{-5}), the stochastic contrastive loss diverges as the variance approaches zero.

Number of sampled embeddings. In the second block of Tab. 4, we report the performance according to the number of sampled embeddings K in (4). At testing time, we use the same number of embeddings and average the output of the classifier to predict the class of the video. Since large numbers of embeddings reflect the entire distribution of the video via Monte-Carlo estimation, the performance increased as K increases. However, a larger number of K leads to more computational requirements. We choose $K = 10$ in consideration of computational costs.

Uncertainty and performance. To analyze the correlation between the uncertainty and the discriminability of learned representations, we measure the inherent uncertainty of videos and report the video retrieval performance on UCF101 according to the training step. For every 10 epochs, we estimate the average uncertainty for all videos in the training set and compute the average of top-1 accuracy. As shown in the left side of Fig. 4, the model learns to minimize uncertainty of videos, thereby improving the retrieval performance. Furthermore, we analyze the correlation between the performance and the uncertainty level by evaluating the retrieval performance on three test splits of UCF101. We divide the uncertainty measured for all videos into 10 uniform bins according to the uncertainty level and compute the average top-1 performance in each of the bins. As presented in the right side of Fig. 4, results show the negative correlation between the uncertainty and the performance, indicating the performance drops as the uncertainty increases. The additional uncertainty analyses with visualization are presented in the supplementary material.

5. Conclusion

We have introduced ProViCo that learns video representations in a self-supervised manner by estimating the distribution and the uncertainty of videos in a stochastic embedding space. The probabilistic framework provides a discriminative sample embedding without any spatiotemporal transformations, while not impairing the nature of the video (by artificial transformations). We constructed the positive and negative pairs based on the probabilistic distance to hold more semantically related candidates for robust contrastive learning without class annotations. The proposed stochastic contrastive loss enables not only the learning of video representations from reliable sample pairs by attenuating the impact of uncertain samples but also minimization of uncertainty from the inherent nature of the raw video. Extensive experiments showed that the probabilistic embedding can be a powerful alternative to the deterministic counterparts, achieving state-of-the-art performance.

6. Broader Impact

Self-supervised video representation learning is an appealing topic in computer vision with many downstream applications. A successful representation learning framework (such as that presented in this work) takes a significant step toward realizing these applications by alleviating the huge financial and environmental costs that would otherwise be necessary. To promote relative research, we discuss that there are possible directions for future work. ProViCo represents the probabilistic video distributions without a temporal encoding. Beyond the simple contrastive learning, pretext tasks to enforce the temporal encoding [32, 33] may further improve the capacity of learned representations. In addition, while the improvement of ProViCo is consistently competitive on the Kinetics-400, UCF101, and HMDB51 benchmarks, all videos in these benchmark datasets are “trimmed videos.” It would be interesting to consider “untrimmed videos” that randomly sampled clips can not be directly utilized for training due to severe background clutter, and their resultant high aleatoric uncertainty. We hope that our uncertainty-based approach will be a valuable foundation for video representation learning on large-scale untrimmed video datasets such as ActivityNet [31].

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Research*, 19, 2018. 2

[2] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *Int. Conf. Learn. Represent.*, 2017. 2

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 1

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[5] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020. 2

[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 *CVPR*, 2020. 7

[7] Anil Kumar Bhattacharyya. On a measure of divergence between two multinomial populations. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943. 4

[8] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015. 3

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NIPS*, 2020. 2

[10] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1

[11] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, 2020. 2, 3

[12] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. A hierarchical variational neural uncertainty model for stochastic video prediction. In *ICCV*, 2021. 3

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[14] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *ICCV*, 2019. 3

[15] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 2

[16] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 3, 4

[17] Ishan Dave, Rohit Gupta, Mamshad Mayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *arXiv preprint arXiv:2101.07974*, 2021. 6, 7

[18] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(9):1734–1747, 2015. 2

[19] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, 2019. 2

[20] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, 2021. 2, 4, 5, 6

[21] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017. 1

[22] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 3

[24] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J. Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, 2018. 1, 2

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NIPS*, 2020. 2

[26] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 1, 2

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NIPS*, 2020. 2, 3

[28] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *ICCVW*, 2017. 1, 5

[29] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 5

[30] Kaining He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3, 7

[31] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 8

[32] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *ICCV*, 2021. 2, 5, 6, 7, 8

[33] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *ICCV*, 2021. 6, 7, 8

[34] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *ECCV*, 2020. 1, 2, 6, 7

[35] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5, 6

[36] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *BMVC*, 2015. 3

[37] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017. 3

[38] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *CVPR*, 2019. 3

[39] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019. 1, 2

[40] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[41] Durk P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *NIPS*, 2015. 4

[42] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NIPS*, 2018. 2

[43] Florian Kraus and Klaus Dietmayer. Uncertainty estimation in one-stage object detection. In *ITSC*, 2019. 3

[44] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 5, 6, 7

[45] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and MingHsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 1

[46] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*, 2021. 3

[47] Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *ICCV*, 2021. 2

[48] Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. In *ICLR*, 2019. 2

[49] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *ICCV*, 2021. 2, 6, 7

[50] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 5

[51] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *AAAI*, 2020. 1, 2, 6, 7

[52] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *ICLR*, 2021. 2

[53] Kira Maag, Matthias Rottmann, Serin Varghese, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Improving video instance segmentation by light-weight temporal uncertainty estimates. In *IJCNN*, 2021. 3

[54] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016. 1, 2

[55] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012. 1

[56] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*, 2014. 2

[57] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *ICLR*, 2019. 2, 3, 4

[58] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2

[59] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: contrastive video representation learning with temporally adversarial examples. In *CVPR*, 2021. 1, 2, 4, 5, 6, 7

[60] Mandela Patrick, Yuki M. Asano, Bernie Huang, Ishan Misra, Florian Metze, Joao Henriques, and Andrea Vedaldi. Space-time crop & attend: Improving cross-modal video representation learning. In *ICCV*, 2021. 2

[61] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *ICCV*, 2021. 1, 2, 5, 6, 7

[62] Rui Qian, Tianjian Meng, Boqing Gong, and Ming-Hsuan Yang. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 1, 2, 5, 6, 7

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[64] Tyler Scott, Karl Ridgeway, and Michael Mozer. Stochastic prototype embeddings. In *ICMLW*, 2019. 2

[65] Yichun Shi, Anil K Jain, and Nathan D Kalka. Probabilistic face embeddings. In *ICCV*, 2019. 2, 3

[66] Anna Silnova, Niko Brummer, Johan Rohdin, Themos Stafylakis, and Lukas Burget. Probabilistic embeddings for speaker diarization. In *Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020. 2

[67] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from from videos in the wild. *arXiv preprint arXiv:1212.0402*. 5, 6, 7, 8

[68] Chen Sun, Arsha Nagrani, Yonglong Tian, and Cordelia Schmid. Composable augmentation encoding for video representation learning. In *ICCV*, 2021. 2, 6, 7

[69] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *ECCV*, 2020. 2

[70] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 5

[71] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[72] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *ICLR*, 2015. 2

[73] Jinpeng Wang, Yuting Gao, Ke Li, Jianguo Hu, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *AAAI*, 2021. 6, 7

[74] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Wei Liu, and Yun hui Liu. Self-supervised video representation learning by uncovering spatio-temporal statistics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 7

[75] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019. 1, 2

[76] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, 2020. 1, 2, 7

[77] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 2

[78] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 2

[79] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005. 2

[80] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *CVPR*, 2021. 1

[81] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[82] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 1

[83] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019. 1, 2

[84] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*, 2021. 2, 5

[85] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, 2020. 1, 2, 6, 7

[86] Serena Yeung, Olga Russakovsky, Greg Mori, and Li FeiFei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016. 1

[87] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *ICCV*, 2019. 3

[88] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *CVPR*, 2019. 1

[89] Nanxuan Zhao, Zhirong Wu, Rynson W.H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *ICLR*, 2021. 2