Inter-Intra Cross-Modality Self-Supervised Video Representation Learning by Contrastive Clustering

Jiutong Wei, Guan Luo, Bing Li Weiming Hu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
weijiutong2018@ia.ac.cn, {gluo,bli,wmhu}@nlpr.ia.ac.cn

***Abstract*—This paper introduces an online self-supervised method that leverages inter- and intra-level variance for video representation learning. Most existing methods tend to focus on instance-level or inter-variance encoding but ignore the intra-variance existing in clips. The key observation to solving this problem is the underlying correlation between visual and audio, in which the distribution of flow patterns in feature space is diverse, but expresses complementary similar semantics. And in the semantic feature space, the horizontal dimension of the feature matrix could be regarded as cluster labels. These cluster labels should be consistent for different modalities of the same video clip. Based on this idea, we propose an end-to-end inter-intra cross-modality contrastive clustering scheme to simultaneously optimize the inter- and intra-level contrastive loss. Experiments show that our proposed approach is able to considerably outperform previous methods for self-supervised learning on HMDB51 and UCF101 when applied to video retrieval and action recognition tasks.**

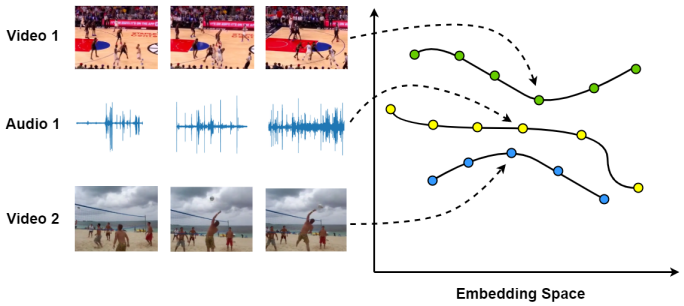

Fig. 1. The idea of our method is the "semantic confusion" problem. For example, similar audio may correspond to videos with very different spatial apparent informations. Based on strictly aligned optimization goal, the general audio-visual synchronization proxy task is prone to be inadequate in generalization and accuracy of self-supervised video representation learning.

I. Introduction

Recent progress in computer vision stems from a huge number of labeled videos as well as deep convolutional neural networks. Generally, a network pretrained with ImageNet consisting of one million high quality labeled images learns the general visual spatial features and has been used to initialize the network for multiple downstream tasks. However, the annotation of video data is labor-intensive and expensive, thus restricting supervised learning to relish a large quantity of free video resources on the Internet. Meanwhile, representations learned from labeled video data lack generality and robustness, *e.g.*, video features learned for action recognition do not well for video retrieval task [1], [2].

To tackle the aforementioned challenges, various attempts have also been made in self-supervised video representation learning [3], [4]. Particularly, video semantics rely on not only spatial feature, but also temporal variance, and directly employing 2D based methods for videos may not make good use of temporal information which is critical for video applications, i.e. different clips sampled from different time spans of a video exhibit different semantic meanings. For example, *jumping* and *throwing* are two different sub-actions although they are both sampled from a video classified as *playing basketball*. Recently, some efforts have been made to learn multimodal video representations in a self-supervised way, such as audio and optical flow. The main idea is predicting whether clips of different modalities are sampled from the same video [5]–[7]. However, previous approaches overlook the information exchange between modalities, and they are pervasively limited by the heterogeneous complexity of audio-visual/flow scenes, i.e., multiple sound sources, and fast-moving backgrounds. Although the data semantics of different modalities may be similar, using contrastive learning in a single shared feature space directly may cause confusion. For example, the visual information corresponding to the audio of a piece of music may be the performance of an orchestra in a concert hall, or a piece of background music in a movie.

In this paper, we delve into self-supervised video representation learning from the perspective of inter-intra cross-modality contrastive clustering. Our key idea are inspired by the observations that in the feature space, the specific features of video with different modes and the same fragment may be different, but the probability of cluster label attributes is consistent. As shown in Fig. 1, for a given video dataset, we use a deep network to learn both the visual and audio feature matrix whose rows and columns correspond to the instance and cluster representations, respectively. In other words, we treat the label as a special representation by projecting input instances into a subspace with a dimensionality of the cluster number. This label can be regarded as a semantic attribute of videos, which strengthens the underlying relationships between visual and audio at a higher semantic level.

Based on the above observations, we propose a Inter-Intra Contrastive Learning (IICL) framework to learn instance and cluster representations for multimodal video understanding.

Specifically, IICL first learns the feature matrix of data pairs constructed through synchronization across multimodal features. After that, the instance/inter- and cluster/intra- level contrastive learning are conducted in the row and column axis of the multimodal feature matrix by gathering the positive pairs and scattering the negatives.

Our contributions can be summarized as follows:

- We propose a novel self-supervised learning framework namely Inter-Intra Contrastive Learning (IICL) for video representation learning, exploiting the complementary information from different modalities of the same data source;
- As a dual form of instance-level contrast learning, cluster-level contrast learning has sufficient ability to learn and express high-level semantic relations among multi-modal information. The proposed module could produce clustering favorite representations as proven in our experiments;
- The proposed framework is end-to-end trainable. Moreover, IICL could predict the cluster assignment for each new arriving data point in a timely manner without accessing the whole dataset, which allows the model to work in an online fashion. We demonstrate that the video representation learned by IICL can be transferred well to downstream tasks such as action recognition and video retrieval on UCF101, HMDB51.

II. Related Work

A. Single-modal Self-supervised Learning

There is a growing literature on self-supervised representation learning from videos [8]–[10]. We divide the existing self-supervised learning methods into two categories according to their sampling strategy, namely inter learning and intra learning. For intra learning, the constraints are in the sample itself. By using different forms of transformation, some low-level relations are broken down even though statistical or semantic information remains. Different proxy tasks are well-designed to help train the model. The most prevalent approaches include temporal order prediction [11], video colorization [12], spatiotemporal puzzling [13] and speed prediction [14]. These methods generally employ manually designed tasks to seek the spatio-temporal cues in video data, but the performance is limited. For inter learning, the distance in the feature space from the same instance should be close to each other while the distance between different instances should be far from each other. After contrastive losses [15] were proposed, contrastive learning has been proven to be an effective optimization objective in self-supervised learning. [7] propose to leverage the consistency between different modalities to enhance video representation. However, the video representations learned from these methods are mostly dominated by the background instead of the dynamic motions [16], which introduces strong background bias and impairs generalization ability in downstream applications. Therefore, we now propose IICL to balance inter-intra feature variances with multimodal.

B. Multi-modal Self-supervised Learning

Inspired by the human multimodal sensory system [17], [18], many approaches make good use of multimodal information, using one modality to promote the training of another modality. Recent approaches learn from unlabeled multimodal data for a specific target task, such as sound source localization [19] and audio-visual co-segmentation [20]. [21] carried out feature extraction from audio and images data with two independent variational autoencoders. The advantage of using multimodal input is restricted to improving the accuracy of clustering in the testing phase. However, in our experiment, all sensing data is available for use at testing phase in a single-modal pattern or combined with any other modal data in a multimodal pattern, depending on the inference environment, which will be more conducive to handling the insufficient capturing conditions in the real scene.

C. Clustering Videos

Benefit from the powerful feature representative capability of deep neural networks, deep clustering [22]–[24] has demonstrate promising performance on large-scale datasets. The most straightforward way of combining representation learning and clustering is to apply clustering algorithm after hidden feature extraction. XDC [25] performed k-means on the visual and audio features respectively to learn representations. However, they obtain separate clusters for multimodal data and treat those clusters as pseudo labels for supervised learning. Inspired by the idea of "label as representation" [26], [27], we propose to apply deep clustering to visual and audio to achieve intra-cluster contrastive learning. Compared with the above deep clustering methods, our aim at multimodally "labelling" an unlabelled video dataset, and our method works well for this task.

III. Methodology

As illustrated in Fig. 2, our method consists of four main components, namely, a multimodal encoder, an co-attention module, an instance-level contrastive module, and a cluster-level contrastive module. In brief, visual and audio encoders construct multimodal data pairs and extracts features from augmented samples, respectively. Co-attention module models the intra-modal interactions in audio and visual streams [5]. After that, we decouple the instance/inter- and cluster/intra-level contrastive learning into two independent subspaces to enhance representation learning at different levels.

A. Multimodal Encoder

In this paper, we consider visual and audio modalities from the training video clips. Let X be the set of N unlabeled video clips, and E_v and E_a be the visual and audio encoders, respectively. Let $F_v = \{f_v = E_v(x) \mid x \in X\}$ and $F_a = \{f_a = E_a(x) \mid x \in X\}$ be the set of visual and audio feature matrix extracted by the multimodal encoder, respectively.

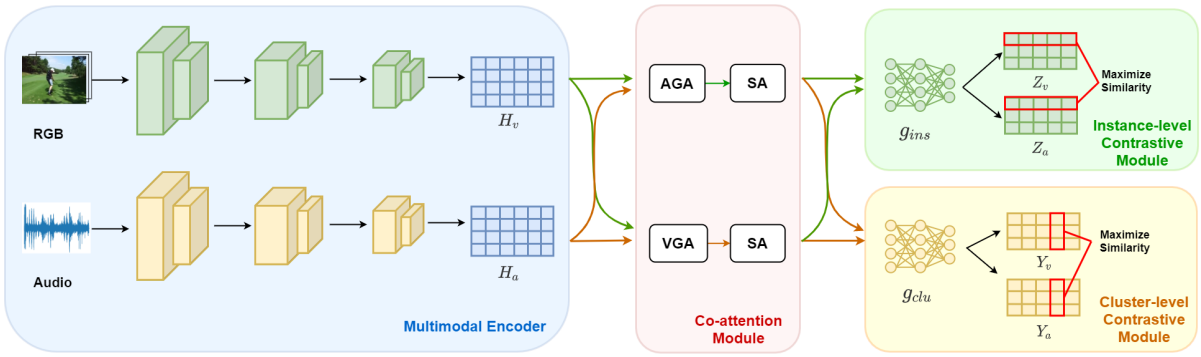

Fig. 2. A schematic illustration of our purposed Inter-Intra Contrastive Learning(IICL) technique. Note that our model is based on the two-stream architecture, which means IICL is generally applicable for other complementary views, *e.g.* optical flow or text.

B. Co-attention Module

Inspired by the great success of Self-Attention(SA), we use cross-modal attention transformer block to enhance the interactions between RGB and audio streams [5]. The usual SA block takes queries(Q), keys(K), and values(V) as inputs. The attended output is computed by

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where d denotes the dimension of Q, K and V. For visual stream, we use audio features F_a to guide the attention learning. In this case, we have $Q = F_a, K = V = F_v$. Therefore, the audio-guided attention learning process tends to focus on the values in the visual stream related to audio information. Similar to the visual stream, visual-guided attention has $Q = F_v, K = V = F_a$. This co-attention module delves into semantic interaction between different modal and outputs the attended visual features H_v and audio features H_a. Note that co-attention module can be stacked in multiple layers to refine the attention maps.

C. Inter-level Contrastive Module

Similar to learning with triplet loss, we employ a multimodal inter-level similarity constraint to ensure that a audio content is more similar to matched visual content than unmatched one, and vice versa. Our goal is to maximize the similarity between an anchor and a positive instance while minimizing the similarity between an anchor and a negative instance in the inter-level space. The constraint of the inter-level similarity is as follows:

$$sim(h_{v_i}, h_{a_i}) > sim(h_{v_i}, h_{a_j}) + \epsilon \tag{2}$$

where i means the i-th video, and ϵ indicates a margin constant. $sim(h_v, h_a)$ stands for the multimodal similarity features in the embedding space.

To filter the representation gaps of multimodal data, we stack a two fully-connected layer $g_{ins}(\cdot)$ to map the attended hidden feature matrix H to an instance-level subspace via $z = g_{ins}(h)$. By measuring similarity with cosine distance $sim(h_v, h_a) = (h_v \cdot h_a)/(||h_v||_1 \cdot ||h_a||_1)$, our goal can be achieved by optimizing a contrastive loss. For visual stream:

$$L_{v_{ins}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(z_{v_i}, z_{a_i})/\tau_{ins})}{\sum_{j=1}^{N}\exp(sim(z_{v_i}, z_{a_j})/\tau_{ins})} \tag{3}$$

where τ_{ins} is an instance-level temperature controlling the concentration of the feature embedding distribution. And the total inter-level contrastive loss is computed by

$$L_{ins} = L_{v_{ins}} + L_{a_{ins}} \tag{4}$$

D. Intra-level Contrastive Module

Considering the cluster consistency of visual and audio information in the subspace, we regard the cluster label as feature representation for contrastive learning. Note that after the hidden vector is projected into the subspace, the i-th element of the feature vector can be regarded as its probability of belonging to the i-th class, and it's similar to classification tasks.

Similar to the inter-level contrastive module, we employ another two fully-connected layer $g_{clu}(\cdot)$ to map the hidden feature matrix into an C-dimensional intra-level space via $y = g_{clu}(h)$, where C equals the number of clusters. For cluster-level contrastive learning, we calculate the cluster feature matrix of visual and audio stream separately, $Y_v, Y_a \in \mathcal{R}^{N\times C}$. Let $\tilde{y}_i$ be the i-the column of Y and $\tilde{y}_i$ can be regarded as a feature vector of the current batch data in the i-th cluster.

Intuitionally, given a video clip, its visual and audio stream clustering should be similar. We assume that each sample belongs to only one cluster, the rows of Y tend to be one-hot, which means that all columns should differ from each other. Based on these considerations, Eq.3 is adopted to achieve cluster-level contrastive learning. Again, we use cosine distance to measure the cluster-level similarity. For visual stream:

$$L_{v_{clu}} = -\frac{1}{C}\sum_{i=1}^{C}\log\frac{\exp(sim(\tilde{y}_{v_i}, \tilde{y}_{a_i})/\tau_{clu})}{\sum_{j=1}^{C}\exp(sim(\tilde{y}_{v_i}, \tilde{y}_{a_j})/\tau_{clu})} \tag{5}$$

where τ_{clu} is the cluster-level temperature parameter to control the softness. To avoid the trivial solution that most instances are assigned to the same cluster [28], we add an entropy of cluster assignment probabilities within a mini-batch under each modal $H(Y) = -\sum_{i=1}^{C}[P(\tilde{y}_{v_i})\log P(\tilde{y}_{v_i}) + P(\tilde{y}_{a_i})\log P(\tilde{y}_{a_i})]$, where $P(\tilde{y}_i) = \sum_{j=1}^{N} Y_{ji}/||Y||_1$. And the total intra-level contrastive loss is computed by

$$L_{clu} = L_{v_{clu}} + L_{a_{clu}} - H(Y) \tag{6}$$

E. Optimization

The optimization of IICL proceeds in two stages: initialization and co-training.

Initialization. First, the multimodal encoder are only trained with instance-level contrastive loss. Specifically, the visual and audio encodes are pretrained by optimizing L_{ins} respectively.

Co-training. Two level losses are optimized at the same time and the final objective function is the combination of the instance- and cluster- level contrastive loss:

$$L = L_{ins} + \lambda L_{clu} \tag{7}$$

where λ is a weight parameter applied to balance the clustering contrastive loss across the training process.

IV. Experiments

A. Datasets

The UCF101 [29], HMDB51 [30], Kinetics400 [31], AudioSet [32] and IG65M [33] datasets were used because of the widespread evaluation in video self-supervised learning methods. Note that we filter out around 7K videos in Kinetics that have no audio.

B. Implementation Details

We choose the R(2+1)D [34] and ResNet [35] architecture as E_v and E_a, respectively. We use a 32-frame RGB clip as E_v's input at 30 fps. The resized video spatial resolution is 128×128. E_a's input is a 2D $Q \times P$ spectrogram image extracted from the audio data, where Q is the number of MEL filters and P is the number of audio frames. We sample 2 seconds and use $Q = 40$ MEL filters and $P = 100$ audio frames. And the depth of co-attention module is set to 1.

For the inter-level contrastive module, the dimensionality of the row space is 1024, and the instance-level temperature parameter τ_{ins} is fixed to 0.5 in all experiments. As for the intra-level contrastive module, the dimensionality of the cluster space is naturally set to the number of clusters, and the intra-level temperature parameter τ_{cls} is set to 0.5 for all datasets.

At the *initialization* stage, we train both visual and audio encoders with instance-level contrastive loss for 300 epochs, and the batch size is 256. At the *co-training* stage, the weight factor $\lambda = 0.5$ is adopted to simultaneously optimize the two level contrastive module and the backbone network for another 300 epochs. For optimization, we use Adam with 5×10^{-4} learning rate and 10^{-6} weight decay.

C. Downstream Tasks for Evaluation

1) Video Retrieval: In this task, the features extracted by the encoder is directly used for nearest-neighbor (NN) retrieval. We use the testing set videos to query the k-NNs from the training set and report recall at k (R@k). If the top k nearest neighbors contain at least one video of the same class, a correct retrieval is counted.

2) Video Action Recognition: In this task, we experiment on two settings: (1) fc-only: we fix the pretrained encoder and train a single linear classifier with cross-entropy loss, (2) finetune: we finetune the whole pretrained encoder on the downstream task.

3) Audio Event Recognition: To assess the audio representation, we train a linear classifier on the frozen audio encoder for the audio event classification dataset DCASE [36] as previous work [37] to provide a fair comparison.

D. Ablation Study

1) The Number of Clusters: This section demonstrates the effectiveness of changing the hyperparameter C in contrastive clustering. IICL is pretrained on AudioSet and we monitor the top-1 accuracy of action classification and retrieval performance on UCF101 split 1.

We explore the effects of changing the hyperparameter C in clutser-level contrastive module, using $C = 100, 200, 400,$ and 800. The results shown in Table I demonstrate that our clutser-level contrastive learning is indeed effective for the downstream tasks. We hypothesize that it is because the number of clusters directly affects the semantic complexity of feature representation. We set $C = 400$ for the following experiments.

TABLE I
THE EFFECT OF THE NUMBER OF CONTRASTIVE CLUSTERS C ON IICL PERFORMANCE.

C	Classification Top1		Retrieval
	fc-only	finetune	R@1
100	72.6	93.1	57.6
200	72.8	93.5	58.1
400	**73.5**	**94.3**	59.3
800	73.2	94.1	**59.5**

2) Inter- and Intra-level Contrastive Module: To verify the effectiveness of the inter- and intra-level contrastive module, we evaluate the clustering performance of features and conduct ablation studies on UCF101 and HMDB51 by removing one of the two module. Note that we perform k-means in the instance space instead when the cluster-level contrastive module is removed. Normalized Mutual Information (NMI) and Accuracy (ACC) are empolyed to evaluate the clustering results. Higher values of these metrics indicate better clustering performance. The results are shown in Table II.

According to the experimental results, we can find that inter- and intra-level contrastive module complement each other to improve the overall performance of the whole model.

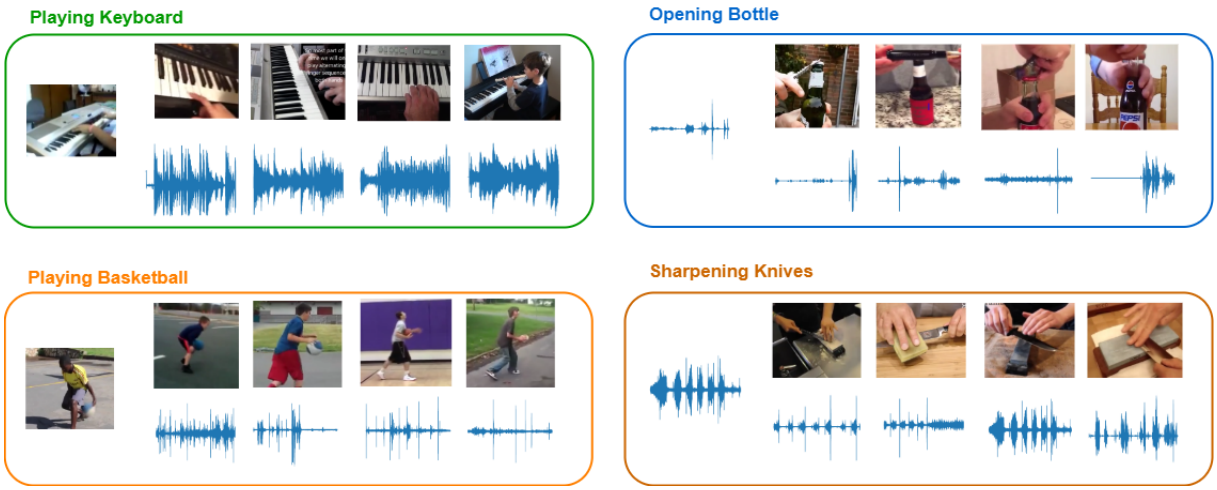

Fig. 3. The qualitative retrieval results of IICL. In each box, the most left data are the anchor data samples, and the data on the right is the nearest neighbor instances of the same and different modes retrieved. (The upper and lower lines of visual and audio information do not correspond)

TABLE II
THE EFFECT OF THE INTER- AND INTRA-LEVEL CONTRASTIVE MODULE. IICL IS PRETRAINED ON KINETICS400.

Dataset	Loss	NMI	ACC
HMDB51	$L_{ins} + L_{clu}$	**0.523**	**0.583**
	L_{ins}	0.495	0.561
	L_{clu}	0.402	0.457
UCF101	$L_{ins} + L_{clu}$	**0.748**	**0.771**
	L_{ins}	0.735	0.767
	L_{clu}	0.712	0.728

3) Cross-modal Retrieval: This experiment evaluate the cross-modality representation ability of cluster-level contrastive module. Specifically, we utilize the cosine value on Kinetics400 to calculate the instance-level retrieval similarity between different modalities, and employ the average of all returned accuracy (mAP) to evaluate our method. This metric measures the ranking information and accuracy jointly. Specifically, we summary the mAP results of two comparison methods for cross-modal retrieval tasks: retrieving audio using visual queries (Visual2Audio) and retrieving video using audio queries (Audio2Visual). The results are shown in Table III, more fully proves the effectiveness of cluster-level contrastive module.

TABLE III
THE PERFORMANCE OF CROSS-MODAL RETRIEVAL TASK. IICL IS PRETRAINED ON KINETICS400.

Dataset	Loss	Visual2Audio	Audio2Visual
Kinetics400	w L_{clu}	**0.745**	**0.759**
	w/o L_{clu}	0.728	0.742

Fig. 3 shows examples of the qualitative results of our method. Considering the label of each cluster is unknown, the ground truth are shown as the reference. In each box, the most left data are the anchor data samples, and the data on the right is the nearest neighbor instances of the same and different modes retrieved. From the figure, we can observe that our method successfully retrieves instances of similar semantics from a single modality data, even though the training process is conducted in a multimodal manner.

E. Comparison to State-of-the-Art

Given one of our best learning setups from ablations, we extend training time and compare our feature representations to the state-of-the-art in multimodal downstream benchmarks

For video retrieval task, which evaluates the quality of features extracted by the pre-trained instance-level contrastive module. To make a fair comparison, all models are pretrained on UCF101. Testing set are utilized to query the top k nearest samples based on their corresponding visual features. We consider k equals to 1, 5, 10, 20, specifically. As shown in Table IV, IICL significantly beats all other self-supervised methods.

TABLE IV
COMPARISON WITH OTHERS ON NEAREST-NEIGHBOUR VIDEO RETRIEVAL ON UCF101.

Method	UCF			
	R@1	R@5	R@10	R@20
Buchler [1]	25.7	36.2	42.2	49.2
VCOP [38]	14.1	30.3	40.4	51.1
CoCLR [7]	55.9	70.8	76.9	82.5
Huang [39]	41.7	57.4	66.9	76.1
MCN [40]	53.8	70.2	78.3	83.4
IICL(ours)	**56.8**	**73.3**	**82.1**	**85.3**

For action recognition task, we finetune the pretrained encoder for UCF101 and HMDB51 video classification, and compare against state-of-the-art self-supervised methods in table V. IICL performs well on both datasets. When comparing the models that are only trained on the RGB stream, *e.g.* ST-Puzzle and SpeedNet, the proposed method significantly outperforms all previous approaches. When pretraining on the Kinetics datasets, IICL achieves state-of-the-art performance

among all listed self-supervised methods. When pretraining on the AudioSet dataset, we also have good results, similar to GDT [37].

TABLE V
COMPARISON ON VIDEO ACTION RECOGNITION.

Method	Dataset	Architecture	UCF	HMDB
ST-Puzzle [13]	Kinetics400	R3D	63.9	33.7
SpeedNet [14]	Kinetics400	S3D-G	81.1	48.8
CoCLR [7]	Kinetics400	S3D	90.6	62.9
XDC [25]	Kinetics400	R(2+1)D-18	86.8	52.6
AVTS [41]	Kinetics400	MC3-18	85.8	56.9
CPD [42]	Kinetics400	3D-Resnet50	88.7	57.7
AVID [43]	Kinetics400	R(2+1)D-18	87.5	60.8
GDT [37]	Kinetics400	R(2+1)D-18	89.3	60.0
IICL(ours)	Kinetics400	R(2+1)D-18	**90.8**	**61.5**
AVTS [41]	AudioSet	MC3-18	89.0	61.6
XDC [25]	AudioSet	R(2+1)D-18	93.0	63.7
XDC [25]	IG65M	R(2+1)D-18	95.5	68.9
GDT [37]	AudioSet	R(2+1)D-18	92.5	66.1
GDT [37]	IG65M	R(2+1)D-18	95.2	**72.8**
ELO [44]	Youtube2M	R(2+1)D-50x3	93.8	67.4
IICL(ours)	AudioSet	R(2+1)D-18	92.8	63.9
IICL(ours)	IG65M	R(2+1)D-18	**95.8**	72.5

For audio classification task, we also achieve state-of-the-art performance among all listed self-supervised methods on DCASE (see Tabel VI).

TABLE VI
COMPARISON ON AUDIO EVENT RECOGNITION.

Method	DCASE
Ensemble [45]	78
SoundNet [46]	88
AVTS [41]	94
XDC [25]	95
IICL(ours)	**96.8**

V. CONCLUSION

Grounded on the idea that the semantic information about different modalities is complementary but not exactly the same and the idea of "label as representation", we proposed the Inter-Intra Contrastive Learning (IICL) method which dually conducts contrastive learning at the instance- and cluster- level under a unified framework for multimodal video representation learning. The proposed cross-modality cluster contrastive module shows promising clustering performance in clustering. In the future, we plan to extend it to other modalities and tasks where other features can be interpreted as provide complementary information.

ACKNOWLEDGMENT

This work was supported by the China State Railway Group CO., Ltd (N2021S010).

REFERENCES

[1] U. Buchler, B. Brattoli, and B. Ommer, "Improving spatiotemporal self-supervision by deep reinforcement learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 770–786.

[2] D. Luo, C. Liu, Y. Zhou, D. Yang, C. Ma, Q. Ye, and W. Wang, "Video cloze procedure for self-supervised spatio-temporal learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 701–11 708.

[3] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6964–6974.

[4] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "Videomoco: Contrastive video representation learning with temporally adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 205–11 214.

[5] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3884–3892.

[6] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[7] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *arXiv preprint arXiv:2010.09709*, 2020.

[8] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1801–1810.

[9] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3636–3645.

[10] C. Gan, B. Gong, K. Liu, H. Su, and L. J. Guibas, "Geometry guided convolutional neural networks for self-supervised video representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5589–5597.

[11] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *European Conference on Computer Vision*. Springer, 2016, pp. 527–544.

[12] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 391–408.

[13] D. Kim, D. Cho, and I. S. Kweon, "Self-supervised video representation learning with space-time cubic puzzles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8545–8552.

[14] S. Benaim, A. Ephrat, O. Lang, I. Mosseri, W. T. Freeman, M. Rubinstein, M. Irani, and T. Dekel, "Speednet: Learning the speediness in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9922–9931.

[15] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[16] J. Wang, Y. Gao, K. Li, Y. Lin, A. J. Ma, H. Cheng, P. Peng, F. Huang, R. Ji, and X. Sun, "Removing the background by adding the background: Towards background robust self-supervised video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 804–11 813.

[17] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[18] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[19] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[20] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2357–2361.

[21] A. Kurobe, Y. Nakajima, H. Saito, and K. Kitani, "Audio-visual self-supervised terrain type discovery for mobile platforms," *arXiv preprint arXiv:2010.06318*, 2020.

[22] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*. PMLR, 2016, pp. 478–487.

[23] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *International conference on neural information processing*. Springer, 2017, pp. 373–382.

[24] X. Li, R. Zhang, Q. Wang, and H. Zhang, "Autoencoder constrained clustering with adaptive neighbors," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 443–449, 2020.

[25] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *arXiv preprint arXiv:1911.12667*, 2019.

[26] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020.

[27] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *2021 AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[28] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama, "Learning discrete representations via information maximizing self-augmented training," in *International conference on machine learning*. PMLR, 2017, pp. 1558–1567.

[29] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[30] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[31] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[32] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[33] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 046–12 055.

[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[37] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "Multi-modal self-supervision from generalized data transformations," *arXiv preprint arXiv:2003.04298*, 2020.

[38] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 334–10 343.

[39] L. Huang, Y. Liu, B. Wang, P. Pan, Y. Xu, and R. Jin, "Self-supervised video representation learning by context and motion decoupling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 886–13 895.

[40] Y. Lin, X. Guo, and Y. Lu, "Self-supervised video representation learning with meta-contrastive network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8239–8249.

[41] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *arXiv preprint arXiv:1807.00230*, 2018.

[42] T. Li and L. Wang, "Learning spatiotemporal features via video and text pair discrimination," *arXiv preprint arXiv:2001.05691*, 2020.

[43] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 475–12 486.

[44] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 133–142.

[45] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[46] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, pp. 892–900, 2016.