SELF-SUPERVISED VIDEO REPRESENTATION LEARNING WITH MOTION-CONTRASTIVE PERCEPTION

*Jinyu Liu[1], Ying Cheng[2], Yuejie Zhang[1], Rui-Wei Zhao[2], Rui Feng[1, 2, *]*

[1]School of Computer Science, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, China
[2]Academy for Engineering and Technology, Fudan University, China
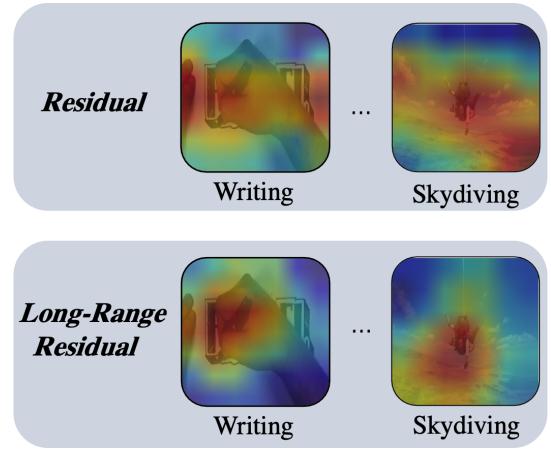{jinyuliu20, chengy18, yjzhang, rwzhao, fengrui}@fudan.edu.cn

ABSTRACT

Visual-only self-supervised learning has achieved significant improvement in video representation learning. Existing related methods encourage models to learn video representations by utilizing contrastive learning or designing specific pretext tasks. However, some models are likely to focus on the background, which is unimportant for learning video representations. To alleviate this problem, we propose a new view called *long-range residual frame* to obtain more motion-specific information. Based on this, we propose the **M**otion-**C**ontrastive **P**erception **Net**work (**MCPNet**), which consists of two branches, namely, **M**otion **I**nformation **P**erception (**MIP**) and **C**ontrastive **I**nstance **P**erception (**CIP**), to learn generic video representations by focusing on the changing areas in videos. Specifically, the MIP branch aims to learn fine-grained motion features, and the CIP branch performs contrastive learning to learn overall semantics information for each instance. Experiments on two benchmark datasets UCF-101 and HMDB-51 show that our method outperforms current state-of-the-art visual-only self-supervised approaches.

Index Terms— Self-Supervised Learning, Video Understanding, Contrastive Learning, Long-Range Residual Frame

1. INTRODUCTION

Human brains with a strong understanding ability can process visual information as seen by eyes and quickly infer its meaning. There are a lot of video understanding tasks, such as video segmentation and video event localization. The most significant difference between video and image understanding is that videos contain complex spatial-temporal contents, but pictures only have spatial information. Hence, how to learn effective video representation is essential yet challenging. However, large-scale datasets require laborious and expensive annotation such as Kinectis-400 [1], etc. Due to large-scale

*Corresponding author.

Fig. 1. **Comparison between models trained with residual frames and with long-range residual frames on Kinetics-100 [2].** The above activation maps are produced by the last convolutional layer of the S3D-G backbone. By utilizing long-range residual frames as one of the inputs in our proposed method, the learned representations can capture motion areas more accurately.

unlabelled data on the Internet, video self-supervised representation learning methods have attracted great attention.

Self-supervised video representation learning methods aim to learn helpful information through pretext tasks that leverage supervision in the data itself and significantly reduce the cost of collecting manual labels. In recent years, some efforts have been made in self-supervised video representation learning. The pretext tasks in these works can be divided into three categories: 1) spatial-focused learning, such as geometry guided learning [3]; 2) temporal-focused learning, such as frame sequencing [4, 5], clip orders prediction [6] and playback rate perception [7, 8, 9, 2]; 3) spatial-temporal learning, such as space-time cubic puzzles [10] and multi-task [11]. However, a problem with most of these pretext tasks is that the information contained in the RGB view is redundant. Some stationary and semantically irrelevant objects in

the background are likely to interfere with the model making judgments. Some researchers [12, 13] point out that poor video comprehension is primarily the result of background cheating. Therefore, some latest methods utilize extra views such as optical flow [14], residual frame [15], etc., to encourage models to focus on the changing areas rather than the irrelevant part of the background, which has achieved convincing improvement. However, the calculation of optical flow is generally expensive, and optical flow is sensitive to light changes. Thus residual frame that requires cheap computation is a more reasonable view, and we propose *long-range residual frame*, as shown in Fig. 1, to obtain more motion-specific information.

In this paper, we propose a novel **M**otion-**C**ontrastive **P**erception **Net**work (**MCPNet**), which consists of two branches, namely, **M**otion **I**nformation **P**erception (**MIP**) and **C**ontrastive **I**nstance **P**erception (**CIP**). For the MIP branch, which aims to learn fine-grained motion features, we sample an RGB clip and a long-range residual clip from the same video, requiring the model to distinguish whether the playback speeds of these two clips are the same. For the CIP branch, which is designed to learn overall semantics information for each instance, we encourage the representations of an RGB clip and a long-range residual clip sampled from the same video with different playback speeds to be close enough in feature space. Two branches are trained jointly during pretraining. Experimental results on two datasets show that the learned features perform well on two downstream tasks, i.e., action recognition and video retrieval.

To summarize, the contributions of this paper are as follows:

- We propose a simple-yet-effective view called long-range residual frame for self-supervised video representation learning, which contains more motion-specific information.

- We propose a novel Motion-Contrastive Perception Network (MCPNet), consisting of a MIP branch and a CIP branch, encouraging the model to focus more on the moving objects and less on the static and irrelevant objects in the background.

- Experiments show that our model can achieve state-of-the-art results for two downstream tasks of action recognition and video retrieval on both two benchmark datasets UCF-101 and HMDB-51.

2. RELATED WORK

Contrastive Learning. Contrastive learning has achieved a lot of success in self-supervised learning [8, 15, 14, 2, 16], which does not pay too much attention to pixel details but can focus on abstract semantic information. Chen et al. [17] proposed a simple framework without requiring specialized architectures or memory bank, which enables the contrastive

tasks to learn useful representations. Tian et al. [18] presented a contrastive multi-view coding approach for video representation learning, which used different views of input videos to maximize the instance-level distinction. He et al. [19] proposed MoCo, which could build a large and consistent dictionary with a queue that could enqueue and dequeue learned embeddings and a moving-averaged encoder that maintained consistency. Their work facilitates contrastive unsupervised learning.

Self-Supervised Visual Representation Learning. Early visual self-supervised representation learning methods mainly focused on images, which designed pretext tasks such as colorization [20], jigsaw puzzles [21], image rotations [22], relative position [23], etc. Later some self-supervised approaches for videos emerged, such as space-time cubic puzzles [10], frame order prediction [4, 5], clip order prediction [6], multi-task [11], etc. Recently, playback speed perception [7, 8, 9, 2] has attracted a lot of attention, which requires the model have a deep understanding of video contents to accomplish tasks. Some latest works used additional views such as optical flow [14], residual frame [15], etc. However, many pretext tasks ignore the background cheating problem in RGB views. Some approaches use additional views to improve the performance of models but do not have a carefully designed pretext task. Our method makes full use of long-range residual frames, relative speed perception, and contrastive learning to learn generic video representations.
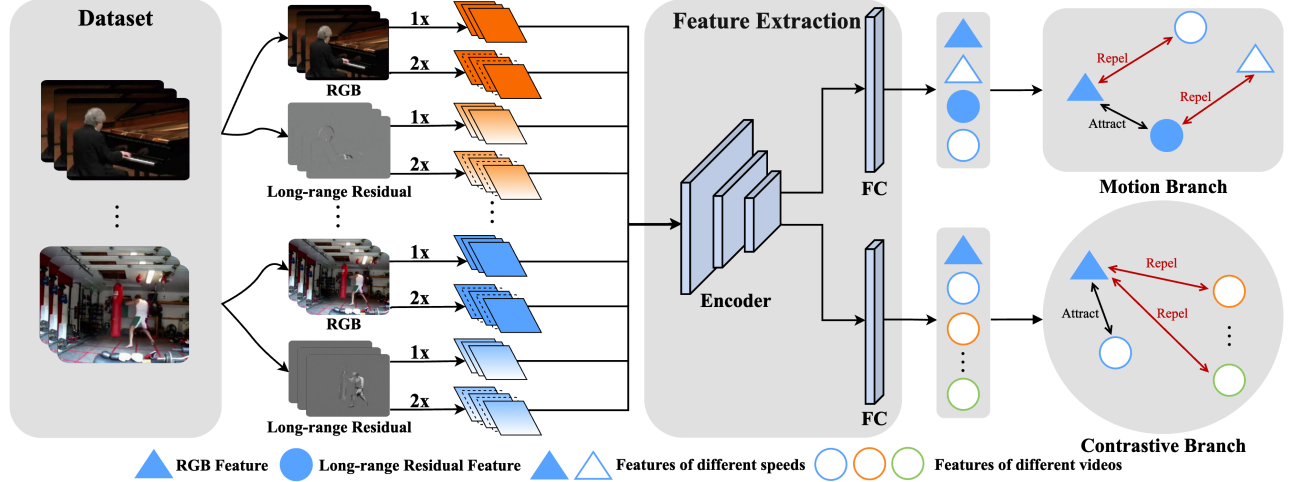
3. METHOD

3.1. Long-Range Residual Frames

Multi-view inputs have been proved to be efficient for instance-based video contrastive learning. Optical flow [15, 14] contains rich motion information but is computationally intensive. The residual frame was first employed by Tao et al. [24] in video representation learning, which could save frame difference with much lower computational volume compared to optical flow. Tao et al. [15] demonstrated that stacked residual frames were also very effective in self-supervised video representation learning. The calculation of the residual frame is shown as below:

$$Res_{i\sim j} = |Frame_{i\sim j} - Frame_{i+1\sim j+1}|, \tag{1}$$

where Res represents residual frame, $Frame$ represents RGB frame, $i \sim j$ is the index interval of the sampled frames.

The common aim of the residual frame and optical flow is to encourage the model to focus on the changing areas of videos rather than the stationary parts. We observe that the variation between adjacent frames is relatively small, while the variation between two frames farther away in the time dimension is huge. Based on this observation, we propose the *long-range residual frame*, which is generated by selecting two frames with a suitable interval in the time dimension to

Fig. 2. An overview of our Motion-Contrastive Perception Network. Our method consists of two branches: Motion Information Perception (MIP) and Contrastive Instance Perception (CIP). MIP branch learns fine-grained motion features by distinguishing the speed difference between two different view clips. CIP branch performs contrastive learning by predicting whether two different view clips come from the same video to learn general semantics information for each instance.

make a difference. Compared to the residual frame, the long-range residual frame contains more frame difference information with the same amount of computation but does not add too much interference information. The process to get long-range residual frames can be formulated as:

$$LongRes_{i\sim j} = |Frame_{i\sim j} - Frame_{i+t\sim j+t}|, \quad (2)$$

With stacked long-range residual frames, the movement preserved covers greater information both in spatial and temporal dimensions compared to stacked residual frames. Therefore, models can extract more specific motion features by focusing on the movements in videos.

3.2. Motion-Contrastive Perception Network

We propose the Motion-Contrastive Perception Network, which consists of two branches: motion information perception and contrastive instance perception, as shown in Fig. 2.
Motion Information Perception. MIP branch aims to capture fine-grained motion features by distinguishing the speed difference between two different view clips. Let $V = \{v_i\}_{i=1}^N$ be a video set containing N videos. Given a video v_i, we sample an RGB clip r_i and two long-range residual clips l_i^1, l_i^2 with playback speeds s_r, s_{l1} and s_{l2}, respectively, where $s_r = s_{l1} \neq s_{l2}$. It needs to be mentioned that an accelerated RGB clip is obtained by taking frames at intervals (e.g., for a 2x playback speed RGB clip, sampling interval is set as 2 frames), and the same playback speed long-range residual clip can be generated by inputting the accelerated RGB clip into Eq. (2) for calculation. We feed the clips into the video encoder $e(\cdot; \theta)$ followed by a projection head $g_m(\cdot; \theta_m)$ to obtain the features f_i^r, f_i^{l1} and f_i^{l2}. We use triplet loss [25] as the loss function of MIP, which can be formulated as:

$$\mathcal{L}_{mip} = \max(\gamma - (sim(f_i^r, f_i^{l1}) - sim(f_i^r, f_i^{l2})), 0), \quad (3)$$

where $\gamma > 0$ is a certain margin and set to 2.0, $sim(,)$ is a dot product function to measure the similarity between two features. The similarity of a positive pair $\{f_i^r, f_i^{l1}\}$ should be larger than a negative pair $\{f_i^r, f_i^{l2}\}$ by a margin γ.
Contrastive Instance Perception. Contrastive learning aims to distinguish different instances from feature space to gain abstract semantics information. In the CIP branch, we ensure that the speeds of RGB view and long-range residual view are always different, thus encouraging the model to pay attention to the overall semantic information for each instance. Two different views of the same video v_i, *e.g.*, $\{r_i, l_i\}$, are treated as positive, while the views from different videos, *e.g.*, $\{r_i, l_j\}$ $(i \neq j)$, are regarded as negative. Specifically, we sample an RGB clip r_i from v_i and N long-range residual clips $\{l_n\}_{n=1}^N$ from V. Then we feed each clip into the encoder $e(\cdot; \theta)$ followed by a projection head $g_c(\cdot; \theta_c)$ to obtain the corresponding features f_i^r and $F^l = \{f_1^l, ..., f_i^l, ..., f_n^l\}$. The feature set F^l consists of one positive sample f_i^l and $n - 1$ negative samples. We use InfoNCE loss [26] as our CIP loss, which can be formulated as:

$$\mathcal{L}_{cip} = -log \frac{exp(sim(f_i^r, f_i^l)/\tau)}{\sum_{j=1}^n exp(sim(f_i^r, f_j^l)/\tau)}, \quad (4)$$

where n is the number of negative samples, f_i^r and f_i^l are extracted features of two different views from ith video. τ is a temperature hyper-parameter.
Optimization. MIP loss and CIP loss from the two branches are combined to get the final loss, which is defined as:

$$\mathcal{L}_{final} = \alpha \cdot \mathcal{L}_{mip} + (1 - \alpha) \cdot \mathcal{L}_{cip}, \quad (5)$$

where $\alpha = 0.5$ is a fixed hyper-parameter to control the importance of each term. MIP branch and CIP branch are pre-trained jointly, and losses $\mathcal{L}$, $\mathcal{L}_{mip}$ and $\mathcal{L}_{cip}$ are used to optimize model parameters θ, θ_m and θ_c with gradient descent.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets. In our experiments, we consider four video datasets, namely, Kinetics-400 [1], Kinetics-100 [2], UCF-101 [27], and HMDB-51 [28]. The Kinetics-400 dataset contains 400 human action categories and provides approximately 240k training video clips and 20k validation video clips. For the self-supervised pre-training, we use the training split of the Kinetics-400 by discarding all of the labels. Each sample is temporally trimmed to be approximately 10 seconds. In ablation studies, to reduce training costs, we use the Kinetics-100 dataset, which consists of 100 classes with the least disk size of videos from Kinetics-400.

For the downstream tasks, UCF-101 and HMDB-51 are used to evaluate the effectiveness of our method. UCF-101 consists of 13,320 videos from 101 human action classes, and HMDB-51 contains about 7K video clips of 51 different human motion classes.

Backbones. We explore three different backbone networks as the video encoder in ablation studies, *i.e.*, R3D-18 [29], R(2+1)D [30], and S3D-G [31]. For action recognition task, the results of R3D-18 and S3D-G are reported. For video retrieval task, the result of R3D-18 is reported.

Self-supervised Pre-training. We sample 16 frames with 112×112 spatial size for each clip unless specified otherwise. The possible playback speed s for clips is set to 1x and 2x. We also use random cropping with resizing, horizontal flips, and color jittering for augmentation. The SGD algorithm is used to optimize our model. We set the initial learning rate to 0.1, scaled linearly with the batch size b, i.e., the learning rate is set to 0.1×b/32. The batch sizes of S3D-G and R3D-18 are 28 and 256, respectively. We train our models using 4 NVIDIA Quadro RTX 6000 for 200 epochs.

Fine-tuning. We initialize the models with the weights from the pre-trained MCPNet, and added two fully-connected layers with randomly initialized weights for classification. We fine-tune the entire model on UCF-101 and HMDB-51 with a learning rate of 0.005 for the action recognition task.

Evaluations. To evaluate the generalizability and transferability of our proposed method, we apply the pre-trained model to action recognition and video retrieval tasks. For action recognition, the top-1 accuracies on UCF-101 and HMDB-51 are reported. For video retrieval, top-1, top-5, top-10, top-20, and top-50 accuracies are compared with existing approaches.

4.2. Ablation Studies

We conduct ablation experiments on the influence of t, the effectiveness of individual branch, and the effectiveness of long-range residual frames, respectively. All models are pre-trained with 200 epochs on the Kinetics-100 dataset, except for the w/o pre-training setting.

Influence of t. As depicted in Eq. (2), t is the suitable interval between two frames for long-range residual frame generation. We conduct experiments to explore the influence of this hyper-parameter. Table 1 shows the results of 5 settings of t with S3D-G backbone. The setting of $t = 1$ means that the residual frames are used in our method. It can be observed that the best result is obtained when $t = 4$, so we will set $t = 4$ for all subsequent experiments.

Table 1. Results of different t settings on action recognition.

Settings	UCF-101(%)	HMDB-51(%)
$t = 1$	70.2	40.3
$t = 2$	70.9	40.8
$t = 3$	71.6	41.2
$t = 4$	**72.3**	**41.7**
$t = 5$	71.1	41.4

Effectiveness of Individual Branch. To figure out the contributions of each branch to the final performance, we conduct ablation studies on six ablated models with the S3D-G backbone built upon our base model. The results of action recognition on UCF-101 are shown in Table 2. Compared to training from scratch, pre-training with only the MIP branch can significantly improve the performance from 45.30% to 68.44% on the UCF-101 dataset. Meanwhile, when combing CIP and MIP, we also investigate the speed sampling configuration for the CIP branch. The RGB clips and the long-range residual clips used in the CIP branch can be sampled with the same speeds, random speeds, or different speeds. When the speeds of RGB clips and long-range residual clips are always different, combining CIP and MIP further improves the performance from 68.44% to 72.35%, indicating the effectiveness of cooperative work of the proposed two branches.

Table 2. Ablation study for different components of MCPNet on action recognition.

Method	Configuration	UCF-101(%)
w/o pre-training	-	45.30
w/ CIP only	different speed	64.87
w/ MIP only	-	68.44
MIP + CIP	random speed	70.46
MIP + CIP	same speed	70.74
MIP + CIP (Ours)	different speed	**72.35**

Table 3. Comparisons between residual view and long-range residual view of different backbones for action recognition accuracy (%) on the UCF-101 dataset.

Backbone	Random	Residual	LongRes
R3D-18	42.4	67.8	**69.1**
R(2+1)D	56.0	78.2	**79.0**
S3D-G	45.3	70.2	**72.3**

Table 4. Comparisons with state-of-the-art methods for action recognition accuracy (%) on UCF-101 and HMDB-51 datasets.

Methods	Pre-train Dataset	Backbone	Resolution	UCF-101	HMDB-51
Shuffle&Learn [4]	UCF-101	CaffeNet	224	50.2	18.1
CMC [18]	UCF-101	CaffeNet	224	59.1	26.7
VCP [32]	UCF-101	R(2+1)D	112	66.3	32.2
ClipOrder [6]	UCF-101	R(2+1)D	112	72.4	30.9
PRP [7]	UCF-101	R(2+1)D	112	72.1	35.0
IIC [15]	UCF-101	R3D-18	112	74.4	38.3
3D-RotNet [33]	Kinetics-400	R3D-18	112	62.9	33.7
ST-Puzzle [10]	Kinetics-400	R3D-18	224	63.9	33.7
DPC [34]	Kinetics-400	R3D-18	128	68.2	34.5
SpeedNet [9]	Kinetics-400	S3D-G	224	81.1	48.8
Pace [8]	Kinetics-400	S3D-G	224	87.1	52.6
CoCLR [14]	Kinetics-400	S3D-G	128	87.9	54.6
RSPNet [2]	Kinetics-400	S3D-G	224	89.9	59.6
ASCNet [16]	Kinetics-400	S3D-G	224	90.8	60.5
MCPNet (Ours)	Kinetics-400	R3D-18	112	82.2	52.5
MCPNet (Ours)	Kinetics-400	S3D-G	224	**91.5**	**62.6**

Effectiveness of Long-Range Residual Frames. We conduct ablation studies on both residual view and long-range residual view. The results of the action recognition task with different video encoders on UCF-101 are illustrated in Table 3. Compared to the residual frames, the model with long-range residual frames consistently improve 1.3%, 0.8%, and 2.1% on R3D-18, R(2+1)D, and S3D-G, respectively. The comparison results demonstrate the effectiveness of the long-range residual frames.

4.3. Evaluation on Action Recognition Task

We compare the results of fine-tuning all parameters with other state-of-the-art methods. Specifically, we pre-train our models on Kinetics-400 and then fine-tune the pre-trained models on UCF-101 and HMDB-51. For S3D-G, we use video frames with 224 × 224 spatial size as input for pre-training and fine-tuning to exploit the proposed approach's potential further. Considering that long-range residual frames and RGB frames are both RGB information in our experiments, we only include the RGB-only results of CoCLR for a fair comparison. From Table 4, we can observe that our models with both S3D-G and R3D-18 backbones outperform other state-of-the-art self-supervised approaches.

4.4. Evaluation on Video Retrieval Task

To further verify the effectiveness of our MCPNet, we also evaluate it on video retrieval task, which can better reflect the semantic-level learning capability. RGB views of original video clips are considered for video retrieval. We directly extract features from the pre-trained model with R3D-18 backbone for video retrieval without fine-tuning. Given the visual feature of a video from the test set as query, video retrieval task aims to return k nearest videos from the training set. When the class of retrieval video is the same as that of the query video, this retrieval result is considered correct. The top-1, top-5, top-10, top-20, and top-50 retrieval accuracies have been shown in Table 5. Compared to other state-of-art methods, our method achieves competitive performance on the UCF-101 dataset. We observe that the top-1 accuracy of CoCLR is better than ours. However, CoCLR uses the optical flow that is computationally complex with extremely accurate motion information as input and adopts dual models. At the same time, we only adopt one single model and use the long-range residual frame with negligible computational cost.

Table 5. Comparisons with previous methods for video retrieval task on the UCF-101 dataset.

Method	Top-1	Top-5	Top-10	Top-20	Top-50
SpeedNet [9]	13.0	28.1	37.5	49.5	65.0
ClipOrder [6]	14.1	30.3	40.0	51.1	66.5
Jigsaw [21]	19.7	28.5	33.5	40.0	40.9
OPN [5]	19.9	28.7	34.0	40.6	51.6
Buchler [35]	25.7	36.2	42.2	49.2	59.5
VCP [32]	18.6	33.6	42.5	53.5	68.1
CMC [18]	26.4	37.7	45.1	53.2	66.3
Pace [8]	31.9	49.7	59.2	68.9	80.2
IIC [15]	42.4	60.9	69.2	77.1	86.5
RSPNet [2]	41.1	59.4	68.4	77.8	88.7
CoCLR [14]	**53.3**	69.4	76.6	82.0	-
MCPNet (Ours)	48.5	**71.6**	**78.8**	**86.0**	**92.8**

To qualitatively assess the capabilities of our model, we have given some examples in Fig. 3. The left is the query video from the UCF-101 testing set, and the right shows the top-3 nearest neighbors from the UCF-101 training set. It can be seen that our method can accurately retrieve videos of the same category.
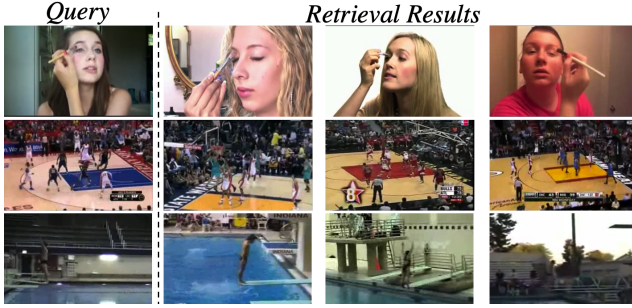
Query   *Retrieval Results*

Fig. 3. Qualitative examples of the video retrieval task.

5. CONCLUSION

In this paper, we propose a novel **M**otion-**C**ontrastive **P**erception **Net**work (**MCPNet**), consisting of two branches. One is the Motion Information Perception (MIP), which learns fine-grained motion features by distinguishing the speed difference between two different view clips, and the other is the Contrastive Instance Perception (CIP) to learn overall semantics information for each instance by distinguishing whether two different view clips come from the same video. By utilizing the proposed stacked long-range residual frames produced by original RGB frames, that is only using information from the RGB frames, our method outperforms state-of-the-art visual-only self-supervised methods on action recognition and video retrieval tasks.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Joao et al., "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[2] Chen et al., "RSPNet: Relative speed perception for unsupervised video representation learning," in *AAAI*, 2021.

[3] Gan et al., "Geometry guided convolutional neural networks for self-supervised video representation learning," in *CVPR*, 2018.

[4] Misra et al., "Shuffle and learn: unsupervised learning using temporal order verification," in *ECCV*, 2016.

[5] Lee et al., "Unsupervised representation learning by sorting sequences," in *ICCV*, 2017.

[6] Xu et al., "Self-supervised spatiotemporal learning via video clip order prediction," in *CVPR*, 2019.

[7] Yao et al., "Video playback rate perception for self-supervised spatio-temporal representation learning," in *CVPR*, 2020.

[8] Sauder et al., "Self-supervised video representation learning by pace prediction," in *ECCV*, 2020.

[9] Benaim et al., "Speednet: Learning the speediness in videos," in *CVPR*, 2020.

[10] Kim et al., "Self-supervised video representation learning with space-time cubic puzzles," in *AAAI*, 2019.

[11] Doersch et al., "Multi-task self-supervised visual learning," in *ICCV*, 2017.

[12] Wang et al., "Removing the background by adding the background: Towards background robust self-supervised video representation learning," in *CVPR*, 2021.

[13] Huang et al., "Self-supervised video representation learning by context and motion decoupling," in *CVPR*, 2021.

[14] Han et al., "Self-supervised co-training for video representation learning," in *NeurIPS*, 2020.

[15] Tao et al., "Self-supervised video representation learning using inter-intra contrastive framework," in *ACM MM*, 2020.

[16] Huang et al., "ASCNet: Self-supervised video representation learning with appearance-speed consistency," in *ICCV*, 2021.

[17] Chen et al., "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[18] Tian et al., "Contrastive multiview coding," in *ECCV*, 2020.

[19] He et al., "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[20] Gustav et al., "Learning representations for automatic colorization," in *ECCV*, 2016.

[21] Noroozi et al., "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.

[22] Gidaris et al., "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[23] Doersch et al., "Unsupervised visual representation learning by context prediction," in *ICCV*, 2015.

[24] Tao et al., "Rethinking motion representation: Residual frames with 3d convnets for better action recognition," *arXiv preprint arXiv:2001.05661*, 2020.

[25] Schroff et al., "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[26] Gutmann et al., "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS*, 2010.

[27] Soomro et al., "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[28] Kuehne et al., "HMDB: A large video database for human motion recognition," in *ICCV*, 2011.

[29] Hara et al., "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *CVPR*, 2018.

[30] Tran et al., "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.

[31] Xie et al., "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018.

[32] Luo et al., "Video cloze procedure for self-supervised spatio-temporal learning," in *AAAI*, 2020.

[33] Jing et al., "Self-supervised spatiotemporal feature learning via video rotation prediction," *arXiv preprint arXiv:1811.11387*, 2018.

[34] Han et al., "Video representation learning by dense predictive coding," in *ICCVW*, 2019.

[35] Büchler et al., "Improving spatiotemporal self-supervision by deep reinforcement learning," in *ECCV*, 2018.