# Supplementary for "RSPNet: Relative Speed Perception for Unsupervised Video Representation Learning"

In the supplementary, we provide more experimental details and results on action recognition, video retrieval and RoI visualization.

## A More details on action recognition on *Something-V2*

We compare our RSPNet with random initialized models and supervised pre-trained models on *Something-V2* dataset. Following the settings in Lin *et al.* (Lin, Gan, and Han 2019), we train models for 50 epochs and set the initial learning rate to 0.01 (decays by 0.1 at epoch 20 and 40). For the supervised pre-trained models, the ResNet-18 and S3D-G are pre-trained on K-400 dataset, and C3D is pre-trained on Sport-1M dataset (Karpathy et al. 2014). Both K-400 and Sport-1M are large-scale datasets with manually annotated action labels, and thus the supervised pre-trained models are strong baselines for our unsupervised pre-trained RSPNet.

As shown in Table 3 in the paper, despite the absence of annotations, RSPNet consistently outperforms the random initialized models on all three models and even surpasses the supervised pre-trained models on ResNet-18 and C3D. These results reveal that the proposed RSP and A-VID pretext tasks help models to learn representative action features, which are beneficial to the downstream action recognition task.

## B More details and results on video retrieval

Following Xu et al. (2019), we conduct video retrieval experiments on the split 1 of UCF101 dataset and use the video clips in testing set to retrieve the clips in training set. We evenly sample 10 clips for each video and take the output of the last convolutional layer in spatial-temporal encoder as clip-level features. Max-pooling is performed over each clip-level features to squeeze the spatial dimensions. Then, we average 10 clip-level features to obtain a video-level feature vector, which is used for retrieval. Our RSPNet is pre-trained on K-400 dataset.

In Table 1, we show more experimental results with different values of $k$. Our RSPNet consistently outperforms all other methods under all $k$ values. We believe that the proposed pretext tasks, which seek to find out two clips with similar information, enforce models to learn discriminative video representation that is helpful for the video retrieval task. We also show qualitative results in Figure 1 (a), which demonstrate the effectiveness of RSPNet for retrieving videos.

| Method | Architecture | Top-$k$ | | | | |
|---|---|---|---|---|---|---|
| | | $k=1$ | $k=5$ | $k=10$ | $k=20$ | $k=50$ |
| OPN (Lee et al. 2017) | OPN | 19.9 | 28.7 | 34.0 | 40.6 | 51.6 |
| Büchler, Brattoli, and Ommer (2018) | CaffeNet | 25.7 | 36.2 | 42.2 | 49.2 | 59.5 |
| ClipOrder (Xu et al. 2019) | R3D | 14.1 | 30.3 | 40.0 | 51.1 | 66.5 |
| SpeedNet (Benaim et al. 2020) | S3D-G | 13.0 | 28.1 | 37.5 | 49.5 | 65.0 |
| VCP (Luo et al. 2020) | R(2+1)D | 19.9 | 33.7 | 42.0 | 50.5 | 64.4 |
| Pace (Wang, Jiao, and Liu 2020) | C3D | 31.9 | 49.7 | 59.2 | 68.9 | 80.2 |
| RSPNet (Ours) | C3D | 36.0 | 56.7 | 66.5 | 76.3 | 87.7 |
| | ResNet-18 | **41.1** | **59.4** | **68.4** | **77.8** | **88.7** |

Table 1: Video retrieval results on UCF101, measured by top-$k$ retrieval accuracy (%).

## C More details and results on RoI visualization

In this section, we provide more details on the calculation of RoI. We first give a brief introduction of the class activation mapping (CAM) technique (Zhou et al. 2016). Then, we introduce the calculation of RoI, which visualizes the salient regions contributing most to the similarity score for RSP and A-VID tasks.

We denote $\mathbf{F} \in \mathbb{R}^{C\times H\times W\times T}$ as the features after the last convolutional layer with $C$ channels and $H, W, T$ spatial-temporal size. We perform global average pooling on $\mathbf{F}$ to obtain a feature vector $\mathbf{x} \in \mathbb{R}^{C\times 1}$. For a specific class, the prediction score $v$ is obtained by using a linear classifier with parameters $\mathbf{w} \in \mathbb{R}^{1\times C}$, *i.e.*, $v = \mathbf{w}\mathbf{x}$. The class activation map $\mathbf{M}_v \in \mathbb{R}^{H\times W\times T}$ for prediction score $v$ is defined as

$$\mathbf{M}_v = \mathbf{w}\mathbf{F}. \tag{1}$$

Such class activation maps demonstrate salient regions contributing most to prediction score $v$. Please refer to Zhou et al. (2016) for more details.

As for RSPNet, we calculate the similarity $s$ between video clip features $\mathbf{x}_i$ and $\mathbf{x}_j$ using cosine distance, *i.e.*, $s = (\mathbf{W}_j\mathbf{x}_j)^\top(\mathbf{W}_i\mathbf{x}_i) = ((\mathbf{W}_j\mathbf{x}_j)^\top\mathbf{W}_i)\mathbf{x}_i$, where $\mathbf{W}_i \in \mathbb{R}^{128\times C}$ and $\mathbf{W}_j \in \mathbb{R}^{128\times C}$ are the parameters for the projection

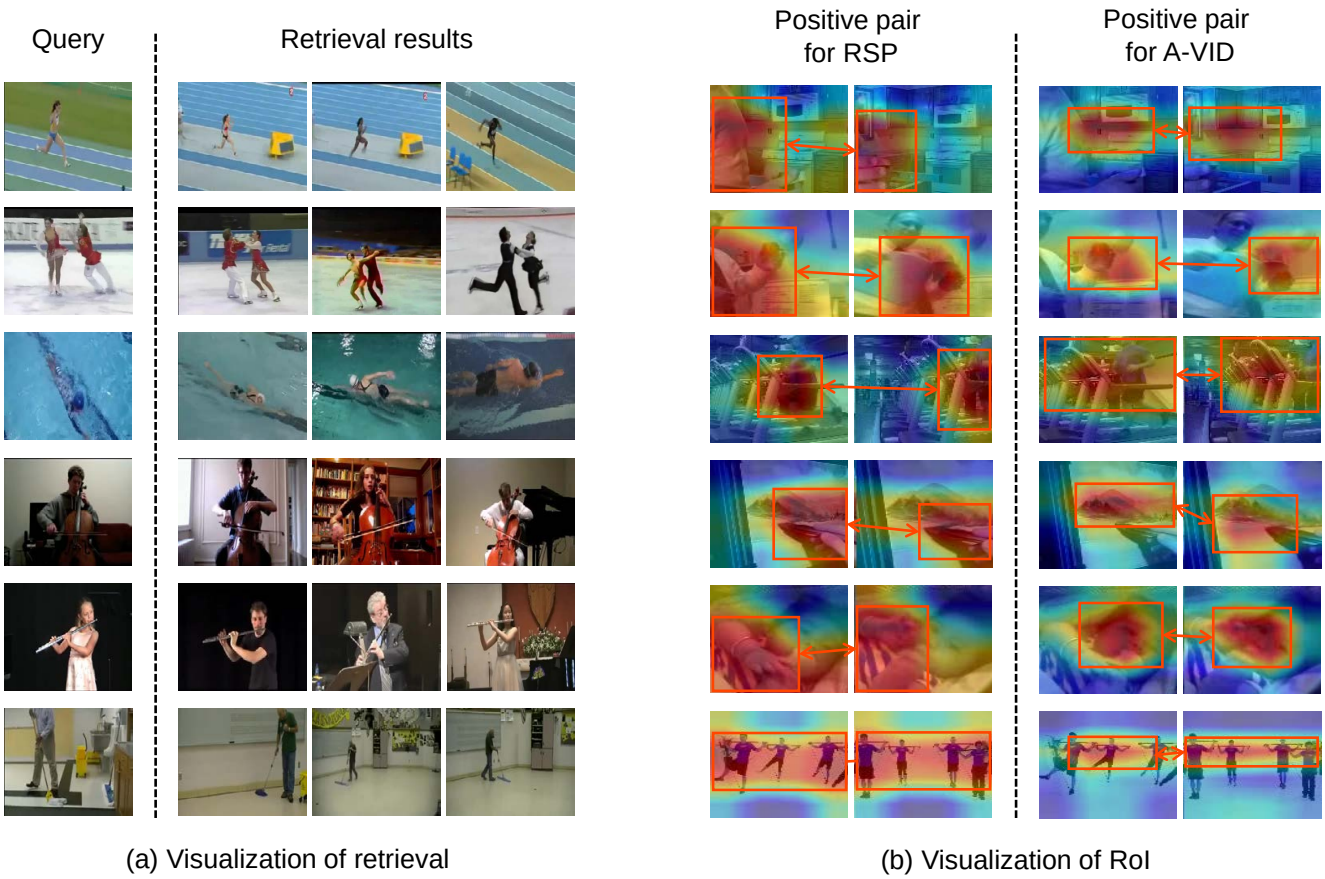


Figure 1: Visualization results of (a) video retrieval and (b) RoI.

head $g_m$ (or $g_a$). $\mathbf{W}_i$ and $\mathbf{W}_j$ can be the same (Chen et al. 2020) or different (Wu et al. 2018; He et al. 2020). Clip features $\mathbf{x}_i$ are average pooled from features $\mathbf{F}_i \in \mathbb{R}^{C\times H\times W\times T}$. In analogy with Equation (1), the similarity activation maps $\mathbf{M}_s \in \mathbb{R}^{H\times W\times T}$ of clip $\mathbf{c}_i$ for similarity score $s$ can be defined as

$$\mathbf{M}_s = ((\mathbf{W}_j\mathbf{x}_j)^\top \mathbf{W}_i)\mathbf{F}_i. \tag{2}$$

Such similarity activation map indicates the salient regions of clip $\mathbf{c}_i$ that are used by models to figure out whether the two clips are positive pair. We can also obtain activation maps of clip $\mathbf{c}_j$ in a similar manner.

Although both RSP and A-VID pretext tasks are based on the same features $\mathbf{F}_i$ (and $\mathbf{F}_j$), we use two independent projection heads $g_m$ and $g_a$ to map $\mathbf{F}_i$ (and $\mathbf{F}_j$) to different 128-D embedding spaces, as shown in Figure 2 in the paper. Thus, the parameters of linear layers for two pretext tasks are different. Consequently, the activation maps can be different and models can focus on learning different clues for completing each specific pretext task. The visualization results in Figure 1 (b) show that models tend to learn discriminative motion features for RSP task and appearance features for A-VID task.

## References


Benaim, S.; Ephrat, A.; Lang, O.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; Irani, M.; and Dekel, T. 2020. SpeedNet: Learning the Speediness in Videos. In *CVPR*.

Büchler, U.; Brattoli, B.; and Ommer, B. 2018. Improving Spatiotemporal Self-supervision by Deep Reinforcement Learning. In *ECCV*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations .

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Li, F. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*.

Lee, H.; Huang, J.; Singh, M.; and Yang, M. 2017. Unsupervised Representation Learning by Sorting Sequences. In *ICCV*.

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *ICCV*.

Luo, D.; Liu, C.; Zhou, Y.; Yang, D.; Ma, C.; Ye, Q.; and Wang, W. 2020. Video Cloze Procedure for Self-Supervised Spatio-Temporal Learning. In *AAAI*.

Wang, J.; Jiao, J.; and Liu, Y. 2020. Self-supervised Video Representation Learning by Pace Prediction. *arXiv* abs/2008.05861.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *CVPR*.

Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; and Zhuang, Y. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *CVPR*.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*.