Spatial-temporal Concept based Explanation of 3D ConvNets

Ying Ji*
Nagoya University

Yu Wang*
Hitotsubashi University

Jien Kato†
Ritsumeikan University

Abstract

Convolutional neural networks (CNNs) have shown remarkable performance on various tasks. Despite its widespread adoption, the decision procedure of the network still lacks transparency and interpretability, making it difficult to enhance the performance further. Hence, there has been considerable interest in providing explanation and interpretability for CNNs over the last few years. Explainable artificial intelligence (XAI) investigates the relationship between input images or videos and output predictions. Recent studies have achieved outstanding success in explaining 2D image classification ConvNets. On the other hand, due to the high computation cost and complexity of video data, the explanation of 3D video recognition ConvNets is relatively less studied. And none of them are able to produce a high-level explanation. In this paper, we propose a STCE (Spatial-temporal Concept-based Explanation) framework for interpreting 3D ConvNets. In our approach: (1) videos are represented with high-level supervoxels, similar supervoxels are clustered as a concept, which is straightforward for human to understand; and (2) the interpreting framework calculates a score for each concept, which reflects its significance in the ConvNet decision procedure. Experiments on diverse 3D ConvNets demonstrate that our method can identify global concepts with different importance levels, allowing us to investigate the impact of the concepts on a target task, such as action recognition, in-depth. The source codes are publicly available at https://github.com/yingji425/STCE.

1. Introduction

With the rapid development of large-scale datasets and powerful computational devices, convolutional neural networks (CNNs) have been widely used in various computer vision tasks, such as image classification [16, 17, 37], semantic segmentation [24, 42], object detection [22, 27] and so on. Despite the fact that CNN models show competitive performance in these tasks, current neural networks are still regarded as black boxes. Due to the large number of parameters and high nonlinearity [25], the underlying prediction mechanism is opaque. This reduces the reliability of neural networks in high-stakes real-world applications such as autonomous driving and medical image analysis [18, 29]. In recent years, explainable artificial intelligence (XAI) has become a popular topic to help comprehend model predictions and increase the credibility of CNNs.

In general, the explanation methods can be divided into local methods and global methods. Local methods concentrate on understanding predictions on individual data instances, while global methods attempt to explain the overall logic of the target ConvNets at the class or dataset level. In this paper, we focus on the global explanation, which is crucial to comprehend the overall behavior of the black boxes.

There are already some methods that provide explanations for 2D image classification ConvNets [6, 14, 26, 28, 36], and most of them are local techniques. Zhou *et al.* [43] generated a Class Activation Map (CAM) using global average pooling for each image to highlight the discriminate regions that are used for the 2D ConvNet to predict class. Ribeiro *et al.* proposed Local Interpretable Model-agnostic Explanations (LIME) [28] to interpret the model by approximating the predictions in a local similarity neighborhood of a target image. However, these methods are not only limited to a single prediction, but they are also difficult for humans to comprehend. The highlighted regions are pixel-level, devoid of human-understandable semantic interpretation. More recently, interpretation with high-level concepts has attracted considerable attention. Kim *et al.* [19] introduced concept activation vectors (CAVs) which use the directional derivatives to quantify the importance of the network prediction to user-defined concepts. Based on [19], Ghorbani *et al.* [12] proposed ACE (Automatic Concept-based) to discover the relationship between image segments and image classification prediction.

Despite solid achievements in 2D image classification interpretation, only a few studies have attempted to interpret 3D action recognition ConvNets, primarily due to the huge computational cost and rich spatial-temporal content of video data. Existing 3D explanation methods are mainly extended from 2D local explanation methods. Stergiou *et*

*These authors contributed equally to this work.
†Jien Kato (jien@fc.ritsumei.ac.jp) is the corresponding author.

al. [33] proposed Saliency Tubes, which applied Grad-CAM [31] to 3D ConvNets. The activation maps of the 3D ConvNet's final convolutional layer are combined to produce heatmaps of input videos. Li *et al.* [21] adopt extremal perturbations (EP) [8] to the video case by adding a spatial-temporal smoothness constraint. However, these methods have two major drawbacks: (1) the discriminative 3D regions are based on a single frame and lack spatial-temporal consistency; and (2) the regions are pixel-level and lack high-level semantic information.

To address these issues, we extend 2D ACE [12] to 3D and propose a high-level global interpretation. For each class, videos are segmented into multiple spatial-temporal supervoxels. Similar supervoxels are grouped to form a meaningful concept. Our method can assign a score for each concept according to its contribution when network predicting. When interpreting the decision procedure of 3D action recognition ConvNets, instead of highlighting essential pixels for a single video, our method can answer two fundamental questions at the class level: *which objects or motions in the video are significant for a particular action recognition class* and *which object or motion is the most crucial clue in this class*.

Our main contributions can be summarized as follows:

1. We propose a novel Spatial-temporal Concept-based Explanation (STCE) for 3D ConvNets. The discriminative regions are spatial-temporal continuous and human-understandable. To the best of our knowledge, STCE is among the first to achieve action recognition interpretation based on high-level video supervoxels.
2. We validate our method using various 3D ConvNets on the Kinetics and KTH datasets. Both qualitative and quantitative results demonstrate that our method can explain the 3D action recognition ConvNets consistent with human cognition.

2. Related work

In this section, we introduce existing attribution explanation literature for 2D image classification ConvNets and 3D action recognition ConvNets.

2.1. Interpretation for 2D ConvNets

Given an input image and a trained 2D ConvNet, the objective of the attribution method is to quantify the contribution of each element in the input. On the basis of which attribute the explanation model evaluates, there are mainly two types of techniques: input and concept attribution. The input attribution explains the ConvNet prediction outcomes in terms of the significance of the input image pixels. Concept attribution, on the other hand, identifies the contribution of human-understandable concepts to the predicted class of an image.

Input attribution The input attribution method is the most commonly used in recent literature. Activation-based methods, such as CAM [43], Grad-CAM [31], Grad-CAM++ [5], and Score-CAM [38], generate weights by utilizing the activations or gradients from intermediate layers of the neural network, then project back the feature maps to the input size in order to produce a heatmap. Perturbation-based methods [7–9, 26, 44] focus on perturbing the input image pixels using occlusion, mask, or generative algorithms. The importance of each pixel is quantified according to the output changes. Since the semantic meanings of pixels are diverse and highly dependent upon one another, explanation methods based on input attribution may result in contradictory explanations for different data instances in the same class [19].

Concept attribution To address this issue, recent research employs human-friendly concepts to interpret 2D ConvNet predictions. The concepts are generated from training data or user-interested data. In [19], every concept is represented by a concept activation vector (CAV). The importance of the concept is evaluated based on the changes in target images toward the direction of the concept. Ghorbani *et al.* [12] defined the concept as superpixel segmentation extracted from input images in order to compute CAVs without human supervision. Based on [19], Goyal *et al.* [13] utilized a conditional VAE model to measure the causal effect of different concepts. Ge *et al.* [11] discussed the structural relationships between concepts with a GNN-based graph reasoning network, so that both visual and structural clues can be used for explanation.

2.2. Interpretation for 3D ConvNets

The goal of interpretation for 3D ConvNets is to investigate the essential regions in both spatial and temporal dimensions of video data. Only a few methods visualize the prediction process of 3D ConvNets. Several methods understand videos using 2D local input attribution techniques initially designed for images by introducing a temporal domain. Srinivasan *et al.* [32] utilized the Layer-wise Relevance Propagation (LRP) [2] to interpret the action recognition based on handcrafted features and Fisher vector. Hartley *et al.* [15] improved the 2D Superpixels Weighted by Average Gradient (SWAG) [14] to the video version by averaging and smoothing a saliency map at the superpixel level. Li *et al.* [21] introduced a smoothness loss function to smooth the perturbation results in both spatial and temporal dimensions.

However, these methods are only able to provide coarse video regions that lack exact semantic meaning. To our knowledge, no research has yet been proposed on the concept attribution for 3D action recognition ConvNets. Hence, the fundamental idea of this paper is to provide a concept-based high-level interpretation for video understanding.

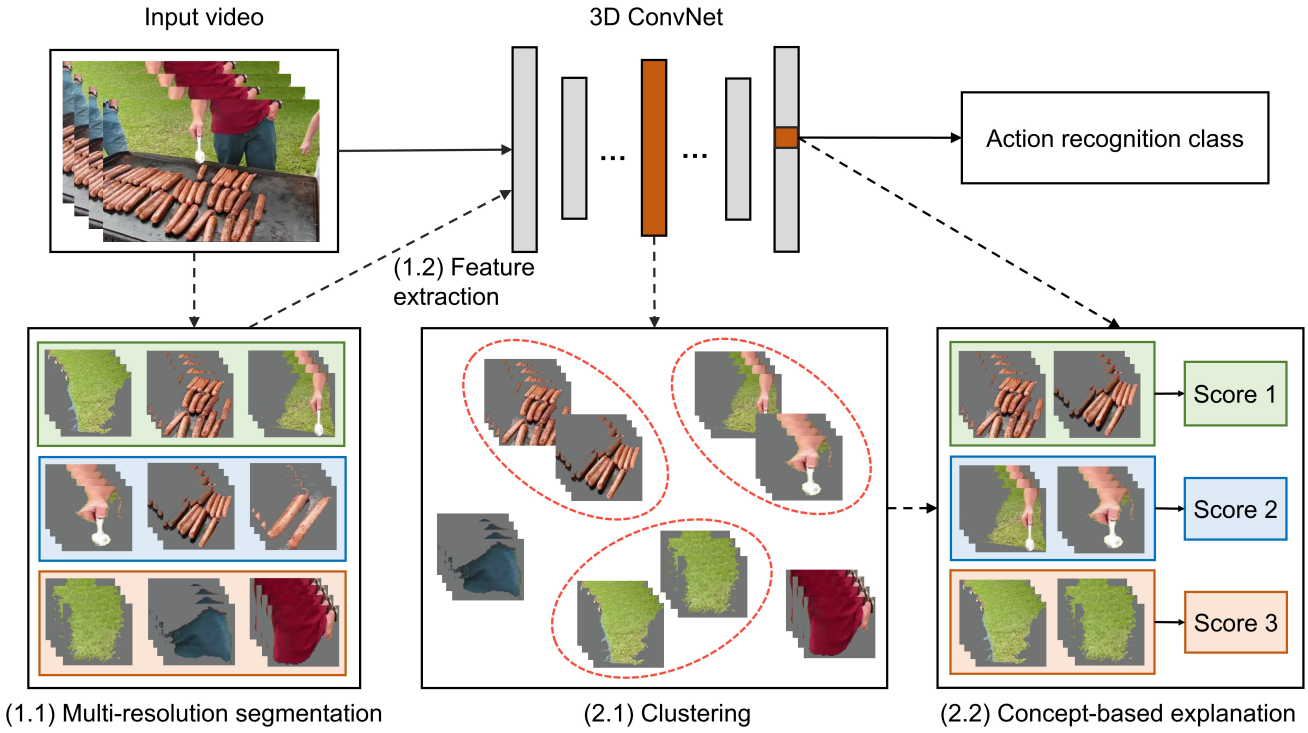


Figure 1. **Illustration of the proposed Spatial-temporal Concept-based Explanation (STCE) framework**. The input is videos from the same class. The video shown here is from "cooking sausages" class in the Kinetics-700 dataset. The procedure consists of two steps: (1) raw videos are first segmented into multi-resolution spatial-temporal volumes. The green, blue, and orange color shown in "multi-resolution segmentation" step indicates that videos are segmented into 15, 50, and 80 supervoxels, respectively. A 3D ConvNet trained on the dataset is then used to extract the feature vector of each supervoxel; (2) the supervoxels are grouped into different clusters. Each cluster is a meaningful concept, such as "hand" or "sausages" or "grass". Such high-level concepts are friendly for humans to comprehend. STCE then calculates the importance score for each concept. The higher score represents that the concept is more important for the ConvNet.

3. Proposed method

Overview In this section, we introduce the details of the proposed Spatial-temporal Concept-based Explanation (STCE) method. The pipeline is shown in Figure 1. Given a video classification dataset and a 3D ConvNet that has been trained using the dataset, we interpret the network by investigating the most important spatial-temporal volumes from the training videos. Videos are first segmented into supervoxels. Similar supervoxels with each class are then clustered and lead to a set of spatial-temporal concepts. STCE finally evaluates the importance score of each concept with respect to the class it belongs. Within the prediction-making procedure, the network pays more attention to the concepts with high scores.

3.1. Supervoxel representation

Let $V = \{(v_n, y)\}_{n=1}^{N}$ be an action recognition dataset which contains N videos. v_n is the nth video and $y \in (1, Y)$ is the label. Each video is first segmented into supervoxels. In contrast to previous research [40, 41], which simply divided videos into segments with equal time intervals, we use a 3D SLIC [1] to divide videos due to its superior performance in video segmentation [39]. In this case, videos are segmented into meaningful spatial-temporal volumes, such as a wheel of a moving car or a swinging arm. Since a video contains information ranging from fine-grained still texture to coarse-grained continuous action motion, each video is segmented 3 times with different levels of resolution to preserve the hierarchical information. For each video v_n, $[s_n^{small}, s_n^{middle}, s_n^{large}]$ contains different size of segments. To avoid calculational cost for redundant supervoxels, we calculate the similarity between every two supervoxels. When the Jaccard index score [10] between two supervoxels is larger than a threshold (we used 0.5 in the experiments), these two segments are recognized as similar pairs. Duplicate segments will be removed, and only the most distinguishable supervoxels will remain.

We train a 3D ConvNet from scratch on V and use it as a feature extractor. Each supervoxel is resized to the standard input size of the network. The empty regions in each frame are filled with average image value, as depicted in grey in

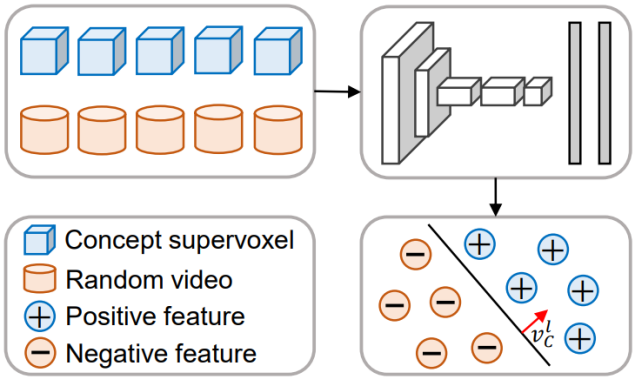


Figure 2. Pipeline to generate a concept activation vector. The inputs are concept supervoxels and the same number of random videos. The direction of the red arrow is orthogonal to the decision boundary of the classifier. The vector v_c^l is used to represent the concept.

Figure 1. The feature vectors are extracted from the top layer l for each supervoxel.

3.2. Concept-based explanation

After extracting the deep features, supervoxels of class y are categorized into distinct concepts. By calculating the Euclidean distance between every pair of supervoxels, similar supervoxels are grouped as a single concept. To preserve the distinctiveness between different clusters, we retain only a small number of segments (40 in our experiment) that are close to the center of each concept. The remaining segments are discarded. Videos in class y can be represented as C concepts, where $C = [c_1, c_2, ..., c_C]$, and each concept c contains 40 supervoxels. s_c^y denotes all the segments belonging to the cth concept.

To determine which concept the network pays more attention to when making the prediction, we continue to evaluate the importance rating for each concept. To this end, we first calculate a concept activation vector (CAV) [19] to characterize the concept. The pipeline to generate the vector v_c^l is illustrated in Figure 2. s_c^y are put into the trained ConvNet as positive samples, while a group of random videos from irrelevant datasets is used as negative samples. Using the 3D ConvNet, features are extracted from both concept supervoxels and random videos. Then a linear classifier is learned to separate the positive and negative samples. The vector v_c^l that is orthogonal to the decision boundary is used to represent the cth concept.

In order to figure out the impact of the concept c given to a video v_n from class y, we follow the idea from [19] to calculate the gradient of logit with respect to the activations of v_n in layer l. Thus the importance score of a particular concept can be computed as $I_{c,y,l}(v_n)$:

$$I_{c,y,l}(v_n) = \lim_{\epsilon \to 0} \frac{p_{l,y}(f_l(v_n) + \epsilon v_c^l) - p_{l,y}(f_l(v_n))}{\epsilon} = \nabla p_{l,y}(f_l(v_n)) \cdot v_c^l \quad (1)$$

where $f_l(v_n)$ is the feature vector of the input video, $p_{l,y}$ is the logit for the video v_n from class y, and v_c^l is the concept vector.

When $I_{c,y,l}(v_n)$ is greater than zero, it indicates that this concept positively affects the ConvNet's prediction for video v_n. If $I_{c,y,l}(v_n)$ is less than zero, the concept has a negative impact.

For one class with K input videos, we compute the directional derivatives for each video. The total importance score for one concept is defined as:

$$S_{c,y,l} = \frac{|v_n \in V : I_{c,y,l}(v_n) > 0|}{K} \in [0, 1] \quad (2)$$

For each concept c, the score $S_{c,y,l}$ computes the proportion of input videos that are positively influenced by the concept. And the higher S indicates the most concerning part for a 3D ConvNet to recognize the video. By sorting the scores, we can at last determine the importance rank of each concept for class y. Unlike previous research, which assessed the importance score of each pixel, our method interprets the ConvNet using concepts with videos from the entire class.

4. Experiment

In this section, we present empirical evaluations of our proposed STCE interpretation method for the 3D ConvNet. Section 4.1 describes the dataset and system set-up. In Section 4.2, the evaluation metric for the experiments is introduced. Section 4.3 presents the quantitative results of adding and removing concepts. Section 4.4 interprets the ConvNet by visualizing the concept frames compared to raw videos. Finally, Section 4.5 discusses the influence of different parameters.

4.1. Implementation details

Datasets We evaluate our method on two popular datasets: Kinetics-700 human action recognition dataset [3] and KTH Action dataset [30].

The Kinetics dataset contains 700 action classes. Our STCE interprets the performance of ConvNet at the class level. Thus, we randomly select 10 classes from the raw dataset to conduct the interpretability experiment. As training data, a total of 6,846 videos are utilized, while as test data, 480 videos are utilized. The video clips have variable high resolutions.

The current KTH dataset includes six types of human actions: walking, jogging, running, boxing, hand waving, and

hand clapping. In total, the dataset contains $2,391$ video sequences. Each video has a low resolution of 160×120. We follow the experiment setup of [23], 80% of the dataset (1528 videos) are used as training data, and the remaining 20% (863 videos) are used as validation.

3D ConvNet Experiments are conducted on three standard 3D ConvNet architectures: C3D [34], R3D-18 [35] and Inflated 3D (I3D) network [4]. Each of the networks is trained from scratch. Following [34], video frames in the Kinetics dataset are resized into 128×171 pixels. Due to the low resolution, videos in the KTH dataset are resized to 120×120. Random horizontal flipping and random cropping are used in data augmentation. The training video frames are randomly cropped to the standard input size of 112×112, while the test video frames are center cropped. In the training stage, 16 continuous frames are randomly chosen as input. In the test stage, the middle 16 continuous frames are fed into the network. All the ConvNets are optimized using stochastic gradient descent (SGD) with momentum set to 0.9. The total number of iterations is 150 epochs. The batch size is 64. The learning rate starts from 0.01 for the first 50 epochs and decreases by a factor of 10 for every 50 epoch. The accuracy derived from the end-to-end ConvNet is the baseline in our experiments.

STCE configuration After training a 3D ConvNet, the next step is to interpret the prediction procedure. We randomly select 200 videos from the training set to generate concepts per class. We employ three different resolution levels to segment videos. Each video is segmented into 15, 50, and 80 supervoxels separately. Similar supervoxels within a single video are eliminated. The number of clusters for each class is set to 25. Each cluster is a concept. We only retain 40 supervoxels in each concept. The activation for each supervoxel is extracted from the top layer l. For C3D, the features from the last fully connected layer (fc7) are extracted. The global average pooling layer is used to extract features for R3D and I3D. Furthermore, we also generate 50 groups of random videos from the HMDB database [20]. The random videos are used to differentiate the concept voxels and calculate concept activation vectors, as described in Figure 2. All experiments are implemented in TensorFlow framework with two 24G NVIDIA RTX 3090 GPUs.

4.2. Evaluation overview

This section introduces the evaluation procedure for STCE. We validate the concepts calculated in Section 3.2 on the test data. After calculating the importance score $I_{c,y,l}$ with training data, the cth concept for class y can be represented as (r_c^y, f_c^y), where r_c^y represents the importance rank of the concept, and f_c^y is the feature vector of the clustering center that has the same dimension as the supervoxel's activation. To quantitatively evaluate the influence of each concept, we compute the recognition accuracy by adding and removing video concepts one by one from the test video.

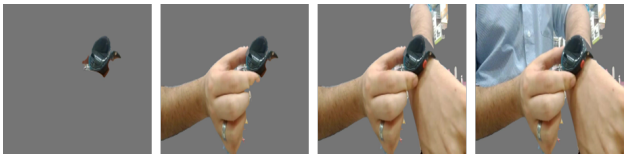


Figure 3. Example of adding concepts from a blank video. The sample frame is extracted from the "checking watch" class. In each step, supervoxels belonging to a specific concept are added to the existing video. For example, the first video represents adding supervoxels belonging to the "watch" concept. The second video represents adding supervoxels that belong to the "left hand" concept to the first video.

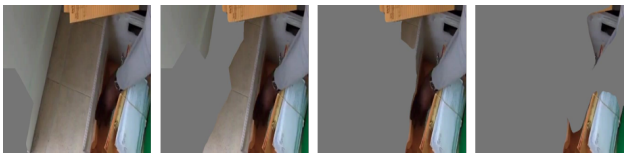


Figure 4. Example of removing concepts from test video. The sample frame is extracted from "delivering mail" class. In each step, all supervoxels from one concept are removed from the raw video.

For each test video t_x, the video is also segmented into P supervoxels. The pth segment s_p^x can be represented as t_x masked with a mask m_p^x:

$$s_p^x = m_p^x \odot t_x \tag{3}$$

As demonstrated by Equation 4, each supervoxel is assigned to the closest concept c by calculating the distance between it and each clustering center.

$$c = \arg\min_c D(f_p^x, f_c^y) \tag{4}$$

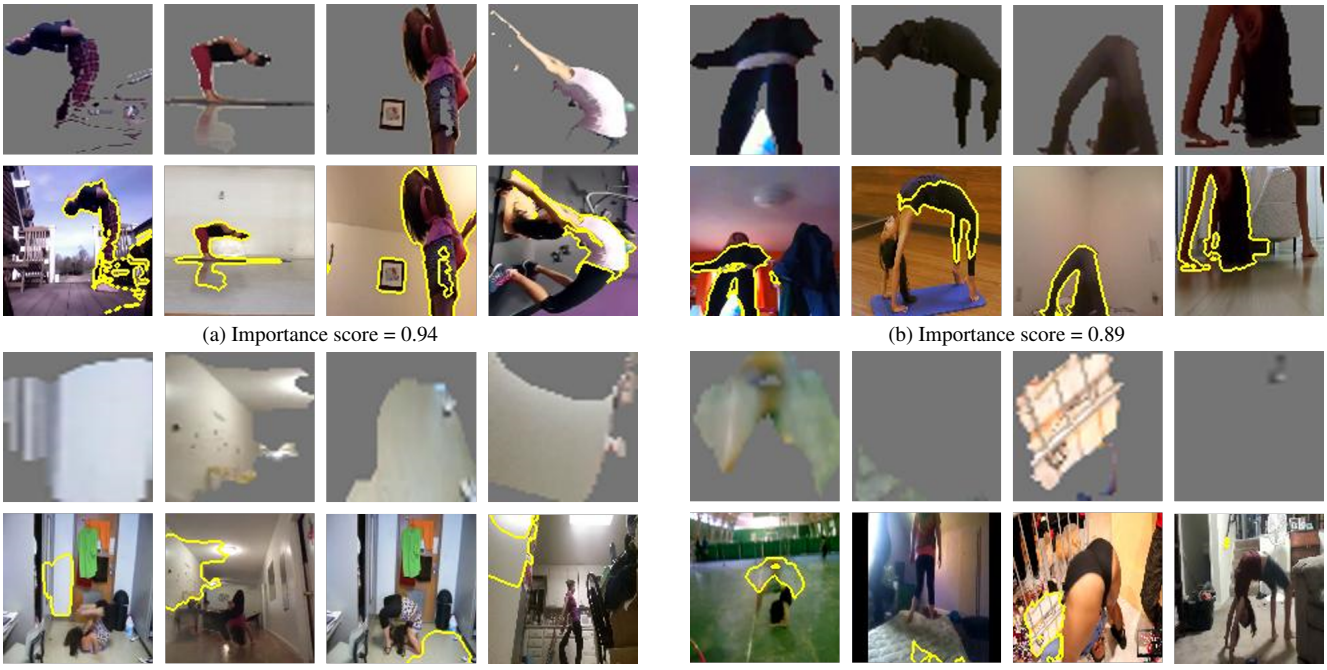
where f_p^x is the feature vector of s_p^x.

Assume that we have a blank video volume of the same size as the test video t_x. When supervoxels from different concepts are added to the blank video, the visible regions of the video can be generated as a spatial-temporal volume:

$$R_x^q = \sum_{j=1}^{q} M_j^x \odot t_x \tag{5}$$

where M_j^x is the sum of supervoxel masks that belongs to concept j. q is the number of concepts that will be set in the following experiments.

As shown in Figure 3, we add supervoxels to a blank video one by one. The intermediate examples are R_x^q with different values of q. When adding all the supervoxel segments from t_x, the blank video will be the same as the test video t_x.



(a) Importance score = 0.94 (b) Importance score = 0.89

(c) Importance score = 0.26 (d) Importance score = 0.25

Figure 5. Visualization of 4 concepts from the "bending back" class using C3D network. The first row of each subfigure is highlighted supervoxels frames. The second row is video frames from raw videos. Figures 5a and 5b are the most two important concepts for ConvNet prediction. The importance scores are 0.94 and 0.89. Figures 5d and 5c are two concepts with the least significance. The importance scores are 0.25 and 0.26.

Table 1. The recognition accuracies of adding concepts using the Kinetics dataset. The baseline is the end-to-end accuracy (%) by 3D ConvNets.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	22.29	34.79	43.33	50.83	**58.54**	79.58
	Random	21.88	33.33	39.79	49.17	55.83	
	Least	23.33	30.63	37.29	45.83	52.29	
R3D	Top	11.67	23.96	32.92	39.38	**46.67**	75.62
	Random	10.63	21.25	32.50	37.71	41.25	
	Least	9.79	16.04	26.04	33.13	41.46	
I3D	Top	23.33	37.92	46.88	54.38	**61.88**	85.63
	Random	25.83	37.50	46.04	52.71	56.67	
	Least	25.42	37.50	47.29	51.46	55.83	

Table 2. The recognition accuracies of removing concepts in Kinetics Dataset. The accuracy decreases the most when the most significant concepts are removed.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	74.38	60.21	55.42	47.50	**43.54**	79.58
	Random	74.38	64.38	59.38	51.88	45.00	
	Least	75.21	66.04	60.21	53.75	47.71	
R3D	Top	69.79	66.25	50.83	39.58	**24.38**	75.62
	Random	72.29	64.38	51.04	39.79	28.13	
	Least	73.33	64.38	51.67	42.29	28.13	
I3D	Top	74.58	65.63	58.96	49.79	**41.88**	85.63
	Random	78.33	70.83	60.42	53.75	43.96	
	Least	80.21	71.25	65.00	57.92	46.25	

In contrast, when supervoxels are removed from raw video t_x, the visible regions are represented as $(1 - M_j^x) \odot t_x$. Figure 4 demonstrates the procedure of removing different concepts.

4.3. Quantitative analysis

In our experiments, q in Equation 5 is set to 5, which indicates at most 5 different concepts will be removed from the raw video. For each test video, when adding and removing the concept, we feed the spatial-temporal volume R_x^q into the ConvNet and make a prediction. We then compare the action recognition accuracy with the baseline.

Table 1 represents the experimental results of adding concepts using the Kinetics dataset. For each model, the first row represents the accuracy of adding concepts with the highest scores. The second raw is adding concepts with random scores. The third row is adding concepts with the lowest scores. It can be seen that adding the most important concepts can lead to higher recognition accuracy for the ConvNets, whereas the concepts with the lowest importance score can offer very little information. In addition, we observe that after adding 5 important concepts, the accuracy

Table 3. Recognition accuracy of adding concepts on the KTH dataset with Standard setting.

Model	Concepts	1	2	3	4	5	baseline
C3D	Top	21.21	23.52	29.43	39.17	**46.23**	91.31
	Random	19.35	23.52	25.26	28.27	32.91	
	Least	17.27	18.77	20.97	25.26	31.87	

Table 4. Recognition accuracy of removing concepts on the KTH dataset with Standard setting.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	89.92	84.94	81.11	76.83	**70.45**	91.31
	Random	90.50	89.80	86.67	82.04	74.74	
	Least	90.50	89.80	89.46	86.44	81.23	

Table 5. Recognition accuracy of adding concepts on the KTH dataset with Small setting.

Model	Concepts	1	2	3	4	5	baseline
C3D	Top	35.46	54.35	66.63	69.52	**73.81**	91.31
	Random	27.69	43.92	59.68	67.44	71.84	
	Least	22.94	34.07	48.44	65.82	68.37	

Table 6. Recognition accuracy of removing concepts on the KTH dataset with Small setting.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	89.69	86.10	78.10	66.86	**59.68**	91.31
	Random	90.50	88.88	78.68	73.93	69.18	
	Least	91.43	89.92	85.75	78.91	71.73	

can exceed 70% of the baseline for C3D and I3D, and 60% of the baseline for R3D.

Table 2 demonstrates the influence of removing concepts. It is evident that removing the essential concepts will result in a reduction in accuracy. Especially for R3D, the accuracy is only 30% of the baseline after only removing five concepts. These experimental results indicate that our STCE is capable of revealing which concept the ConvNets focus on and how much role it plays during the prediction.

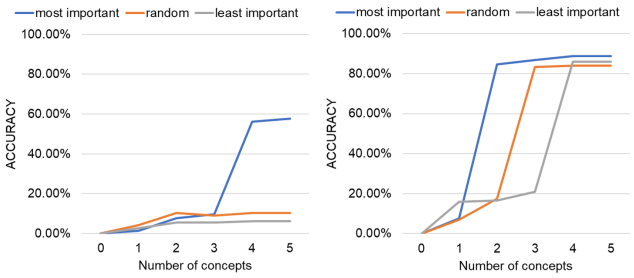
4.4. Qualitative analysis

In order to qualitatively evaluate our proposed model, we visualize video frames of the detected concepts in Figure 5. In particular, we illustrate the most and the least significant concept examples from the "bending back" class in Kinetics dataset. Figure 5a and 5b present the supervoxel frames belongs to the top 2 important concepts. The highest importance score is close to 1, indicating that this concept positively influenced nearly all of the test videos in this class. The first row of each figure shows the highlighted regions, while the second row displays the corresponding raw video frames. It is evident that the dominant actions are body parts and bending actions for predicting the "bending back" class.

Similarly, we also visualize two groups of supervoxels from the least important concepts in Figure 5c and 5d. In contrast, these highlighted regions are primarily located in the background and lack significance. The visualization results interpret what the 3D ConvNet focuses on when recognizing actions. It is obvious that the concepts are intuitive and consistent with human understanding. The remarkable consistency of both quantitative and qualitative results confirms that our proposed STCE is effective for interpreting 3D ConvNets.

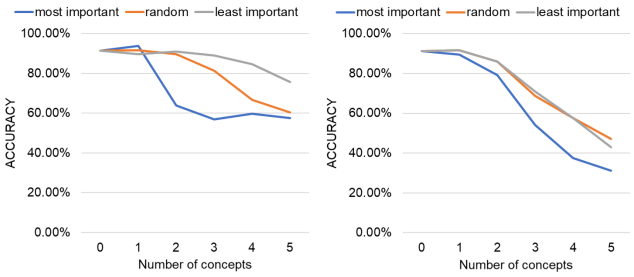
4.5. Discussion

In this section, the influence of various parameter settings is examined. We mainly explore the number of concepts and supervoxels, through comparative experiments on the KTH dataset. In particular, we establish two types of parameter settings for extracting important concepts.



(a) KTH standard setting (b) KTH small setting

Figure 6. The performance of adding concepts using standard and small settings in the "jogging" class from the KTH dataset. The horizontal axis is the number of concepts. The vertical axis is the recognition accuracy. The left Figure 6a is the accuracy with the standard setting. The right Figure 6b indicates the results with the small setting.



(a) KTH standard setting (b) KTH small setting

Figure 7. The recognition accuracy of removing different concepts using standard and small settings in the "jogging" class from the KTH dataset.

Standard setting This setting is the same as experiments on the Kinetics dataset in Section 4.3. Each video is divided

(a) Raw video frames

(b) Concept supervoxel frames excrated using the standard setting

(c) Concept supervoxel frames extracted using the small setting

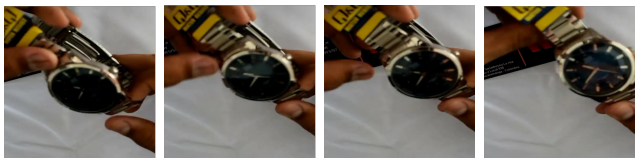
Figure 8. The concept frames from the "boxing" class in the KTH dataset with standard and small settings.

into 15, 50, and 80 segments separately. The number of concept clusters is set to 25 .

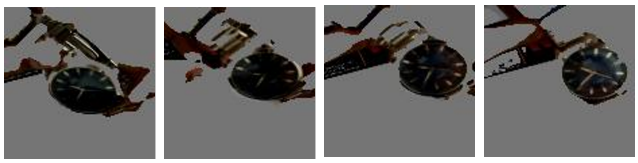
Small setting We also conduct STCE with small parameters because the KTH dataset has a relatively low resolution. In this instance, each video is segmented into 15, 30 and 60 segments, respectively. All the supervoxels in the same class are clustered into 15 concepts.

Table 3 and Table 4 illustrate the accuracy of action recognition with the standard setting, while Table 5 and Table 6 show the accuracy with small setting. It can be seen that both settings are consistent with the tendency demonstrated in Section 4.3. However, we also observe that, despite the fact that adding concepts will undoubtedly improve accuracy, the accuracy can only reach 50% of the baseline in the standard setting. The phenomenon is the same when concepts are removed from the test video. On the other hand, using small parameters and clusters can improve the effectiveness of concepts more then standard setting, which can reach 80% of the baseline. We take the "jogging" class as an example and display the statistical chart in Figure 6 and Figure 7. From the statistical chart, we can conclude that for low-resolution datasets, the ConvNets obtain more information from large-scale concepts.

To more intuitively visualize the difference between these two settings, Figure 8 shows the concept results with both settings extracted from the "boxing" class from the same raw video. Figure 8b represents the supervoxel frames from the most important concepts with standard setting. Figure 8c is also the most essential concept but uses a small setting. Due to the low resolution and large blank background, it is evident that most of the essential regions for the



(a) Raw video frames



(b) Concept supervoxel frames using standard setting

Figure 9. The concept frames from the "checking watch" class in the Kinetics dataset with standard settings.

KTH dataset are located on human body parts. This means that using a standard setting will result in quite a number of backgrounds, which can not improve the recognition accuracy. When using a small setting, the clustered concepts are easier to recognize. For comparison, we also visualize the concept from the "checking watch" class in the Kinetics dataset in Figure 9, the high-resolution datasets contain abundant information such as watch bands, hands, desks, and watches. Even small concepts are sufficient to provide enough information.

5. Conclusion

In this paper, we proposed a Spatial-temporal Concept-based Explanation (STCE) framework for interpreting 3D ConvNet. In contrast to the prior pixel-level strategy, which focuses on a single instance, our research is the first attempt to offer a human-understandable high-level explanation. In our method, videos from an entire class are segmented and clustered into concepts. Each concept comprises similar meaningful supervoxels, such as arms or watches. We then compute the importance scores for each concept. Extensive experiments on three different 3D ConvNets demonstrate the efficiency of STCE. Later, we visualize the detected concepts according to the scores, here we discover that the most and the least essential concepts are consistent with human perception. Finally, we investigate the choice of various parameters for the low-resolution dataset. The number of concepts and clusters does not affect the tendency reported in the experiments. We believe our method successfully discloses the prediction mechanism under the 3D ConvNet. However, because the concepts are calculated from the class level, our method will be time-consuming for large datasets. In the future, we will concentrate on reducing time costs and enhancing ConvNet performance with important concepts.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 2

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 4

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 2

[6] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 1

[7] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017. 2

[8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019. 2

[9] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 2

[10] Feng Ge, Song Wang, and Tiecheng Liu. New benchmark for image segmentation evaluation. *Journal of Electronic Imaging*, 16(3):033011, 2007. 3

[11] Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2195–2204, 2021. 2

[12] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*, 2019. 1, 2

[13] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 2

[14] Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. Swag: Superpixels weighted by average gradients for explanations of cnns. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 423–432, 2021. 1, 2

[15] Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. Swag-v: explanations for video using superpixels weighted by average gradients. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 604–613, 2022. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 1

[18] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 440–450, 2022. 1

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1, 2, 4

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 5

[21] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1120–1129, 2021. 2

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[23] Yiqing Liu, Tao Zhang, and Zhen Li. 3dcnn-based real-time driver fatigue behavior detection in urban rail transit. *IEEE Access*, 7:144648–144662, 2019. 5

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[25] Jan Macdonald, Mathieu E. Besançon, and Sebastian Pokutta. Interpretable neural networks with frank-wolfe: Sparse relevance maps and relevance orderings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 14699–14716. PMLR, 2022. 1

[26] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1, 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1

[29] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[30] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004. 4

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[32] Vignesh Srinivasan, Sebastian Lapuschkin, Cornelius Hellge, Klaus-Robert Müller, and Wojciech Samek. Interpretable human action recognition in compressed domain. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1692–1696. IEEE, 2017. 2

[33] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp, and Ronald Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1830–1834. IEEE, 2019. 2

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 5

[35] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5

[36] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019. 1

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 1

[38] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 2

[39] Murong Wang, Xiabi Liu, Yixuan Gao, Xiao Ma, and Nouman Q Soomro. Superpixel segmentation: A benchmark. *Signal Processing: Image Communication*, 56:28–39, 2017. 3

[40] Yilin Wang, Suhang Wang, Jiliang Tang, Neil O'Hare, Yi Chang, and Baoxin Li. Hierarchical attention network for action recognition in videos. *arXiv preprint arXiv:1607.06416*, 2016. 3

[41] Hao Yang, Chunfeng Yuan, Li Zhang, Yunda Sun, Weiming Hu, and Stephen J Maybank. Sta-cnn: Convolutional spatial-temporal attention learning for action recognition. *IEEE Transactions on Image Processing*, 29:5783–5793, 2020. 3

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 1

[43] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2

[44] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017. 2