PRIVACY-PRESERVING ACTION RECOGNITION

Chengming Zou[1 3 *] *Ducheng Yuan*[2 *] *Long Lan*[4 *] *Haoang Chi*[4]

[1]Hubei Key Laboratory of Transportation Internet of Things, Wuhan University of Technology, Wuhan, 430070, China
[2] School of Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070, China
[3] Peng Cheng National Laboratory, Shenzhen, 518055, China
[4] Institute for Quantum Information & State Key Laboratory of High Performance Computing,College of Computer,
National University of Defense Technology, Changsha 410073, China

ABSTRACT

As the amount of data shared on the network increases, these data pose a threat to our privacy. This paper focuses on the privacy-preserving issues of action recognition for humans. Generally, the face is considered the most identifiable visual cue for a human. However, removing face information is not enough for many privacy-preserving scenes. Thus, we replace the human body with his poses and explore the pose presentation in the action recognition task. In privacy scenes, many human actions could not access in advance. To recognize these unseen actions, we study the zero-shot action recognition in the strict condition of privacy preservation. Specifically, we propose to use *unified actor score* (UAS) to enhance the action recognition accuracy. The experimental results show that UAS outperforms most of the state-of-the-art methods in standard datasets without sacrificing privacy.

Index Terms— privacy preserve, action recognition, zero-shot learning, generalized zero-shot learning, unified actor score

1. INTRODUCTION

In recent years, privacy concerns have been raised by the rapidly increasing number of visual data such as videos and images. Due to the amount of data shared on the network boosting, more sophisticated vision tasks can be performed. Therefore, privacy-preserving methods are at an even higher need nowadays.

To preserve the privacy of the people in Fig. 1(a), we can remove the face information, which is considered as the most identifiable visual cue for a human. However, as shown in Fig. 1(b), masking the face cannot completely make people unrecognized. Thus, [1] is developed to automatically detect people in images and hide them, while ensuring that the output image looks natural. As shown in Fig. 1(c), they mix the original images and projected images in which the people are hidden to generate output images.

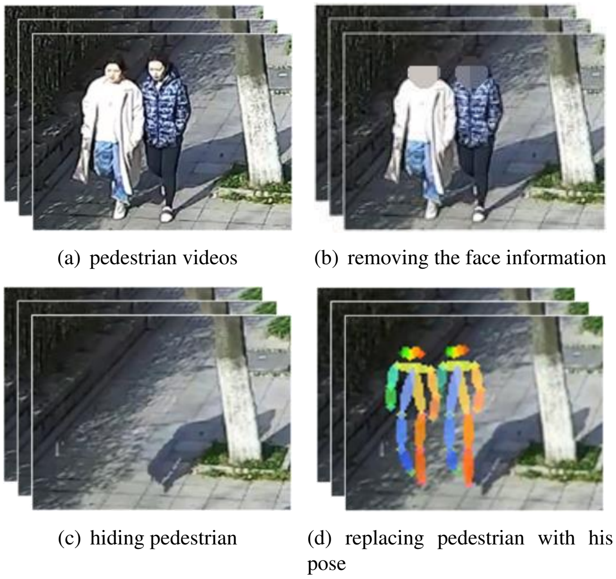

(a) pedestrian videos (b) removing the face information
(c) hiding pedestrian (d) replacing pedestrian with his pose

Fig. 1. Methods of preserving the privacy of the pedestrian.

However, such processing may weaken the recognition utility. By the term utility [2], we refer to certain system properties and intelligibility that represents the amount of useful information that can be extracted from visual data. Based on hiding the pedestrians, some methods replace the human body with his pose. As shown in Fig. 1(d), they get an acceptable trade-off between privacy protection and utility. In this situation, we explore a method to evaluate the utility of the pedestrian replacement method. For the evaluation of the pedestrian replacement method, we need to further analyze the relationship between utility and privacy protection. As detailed in Fig. 1(d), due to the occlusion of the visual angle and other problems, the pedestrian's hands are inserted in her purse and are not detected. However, it is still a good method to protect pedestrian privacy because the action can be recognized as walking. Therefore, we propose to use the action recognition accuracy to evaluate the pedestrian replacement method, which can better reflect the utility of the pedestrian

[*] Corresponding author.

This work was partially supported by National Grand R&D Plan (Grant No. 2020AAA0103501).

replacement method.

To find a better method to recognize actions in privacy scenes, we first analyze the difficulties in this task: 1) there are many unseen samples; 2) there are many people in the picture; 3) the images have different perspectives. To solve the above problems, we propose *unified actor score* (UAS), which is calculated by local object score, global object score, and actor score. For the unseen samples, we try to transfer action knowledge via a semantic embedding build from objects to classify unseen samples. For action localization, we combine the local object information and global object information to get bounding boxes. For different perspectives, the score is computed as the clustering results of human pose embedding [3], which calculates the similarity of two images from different perspectives and learns a compact view-invariant embedding space from 2D joint keypoints alone, without explicitly predicting 3D poses.

This paper has two main contributions: 1) we propose to use the action recognition accuracy to evaluate the pedestrian replacement method by analyzing the relationship between utility and privacy protection; 2) we propose the UAS to calculate the action recognition accuracy in privacy scenes.

2. RELATED WORK

2.1. Privacy evaluation

With the steadily increasing number of protection approaches as well as high variability of visual tasks and scenes, we need more evaluation methodologies for comparing the different approaches. [2] consider existing evaluation methods from two aspects, namely privacy, and utility. And [4] describes a comprehensive evaluation framework.

2.2. Zero-shot Action Recognition

In recent years, *zero-shot learning* (ZSL) and *generalized zero-shot learning* (GZSL) have gained considerable attention since they can deal with classifying images and videos with new categories previously unseen during training. In [5], *Zero-shot Action Recognition* (ZSAR) method is divided into two crucial steps. The first step is the visual embeddings, such as Convolutional 3D Network [6] and Inflated 3D Network [7]; the second step is semantic label embeddings, such as [8] and [9]. They are responsible for providing the features used to map the visual appearance to the semantic description of actions.

3. UNIFIED ACTOR SCORE (UAS)

We compute the unified actor score of a video V given an action class name. As detailed in Fig. 2, the score is calculated by the local object, global object, and actor pose.

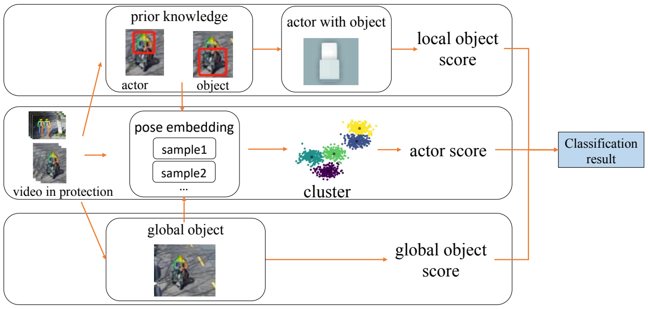


Fig. 2. Overview of UAS. We determine the video as the class with local object score, global object score, and actor score.

3.1. Local object score

We gather prior knowledge on actions, actors, objects, and their interactions to calculate local object scores. The local object score provides a spatial-aware embedding for each bounding box in each frame of a video.

Prior knowledge. In a set of video $V = \{(O, A, Z)\}$, $O = \{O_D, O_N\}$ denotes the objects with detectors O_D and names O_N, Z denotes the action class name, and $A = \{A_D, actor\}$ denotes the actor detector.

Given an action class name $z \in Z$, we aim to select a sparse subset of objects $O_Z \subset O$ that is relevant to the action. This selection relies on semantic textual representations as provided by word2vec [10]. The similarity between object O and the action class name z is given as:

$$w(O, Z) = \cos(e(O), e(z)), \tag{1}$$

where $e(\cdot)$ denotes the word2vec representation of the action class name.

Scoring actor boxes with object interaction. We exploit our prior knowledge to compute a score for the detected bounding boxes in all frames of each test video $v \in V$. Given a bounding box b in the frame F of video v, we define a score function for box b given an action class Z as:

$$s(b, F, Z) = p(A_D|b) + \sum_{O \in O_Z} r(o, b, F, Z), \tag{2}$$

where $p(A_D|b)$ is the probability of an actor being presented in the bounding box b as specified by the detector A_D. The $p(A_D|b)$ is computed by $p(A_D|b) = N(A_D, b)/N(b)$, where $N(A_D, b)$ denotes the number of bounding boxes that contain actors and $N(b)$ denotes the number of bounding boxes in the frame. The function r expresses the object presence and the relation with the actor, and it is defined as:

$$r(o, b, F, Z) = w(o, Z) \times (\max_{f \in F_n} p(oD|f) \times m(o, A, b, f)), \tag{3}$$

$$m(o, A, b, f) = 1 - JSD_2(d(A, o)||d(b, f)), \tag{4}$$

where $JSD_2(\cdot||\cdot) \in [0, 1]$ denotes the Jensen-Shannon divergence with base 2 logarithm [11]. The $p(oD|f)$ is computed by $p(oD|f) = N(oD, f)/N(f)$, where $N(oD, f)$ denotes the number of bounding boxes which contain objects and $N(f)$ denotes the number of bounding boxes in the frame. The more similar distributions yield the smaller JSD_2 value, hence we need to maximize $JSD_2(d(A, o)||d(b, f))$.

Scoring local objects. The score function of Eq. (2) provides a spatial-aware embedding score for each bounding box in each frame of a video. We link those boxes over time that by themselves have a high score from our spatial-aware embedding and have a high overlap amongst each other. Once we have a tube from the optimization, we remove all boxes from that tube and compute the next tube from the remaining boxes.

Let T denote the discovered action tube in a video. The corresponding score is given as:

$$t_{emb}(T, Z) = \frac{1}{|T|} \sum_{t \in T} s(t_b, t_F, Z), \tag{5}$$

where t_b and t_F denote a boundinsg box and its corresponding frame in tube T.

3.2. Global object score

We get the global object scores for classification, which provide classifier scores over a whole video.

To distinguish tubes from different videos in a collection, contextual awareness in the form of relevant global object classifiers is also a viable information source. Let $G = \{GC, GN\}$ denote the set of global objects with corresponding classifiers and names. Given an action class name Z, we select the top relevant objects $GZ \subset G$ again using textual embedding. The score of a video V is then computed as a linear combination of the word2vec similarity and classifier probabilities over the top relevant objects:

$$t_{global}(V, Z) = \sum_{g \in G_z} \varpi(g, Z) \cdot p(g|V), \tag{6}$$

where $p(g|V)$ denotes the probability of global object g being in video V and is computed by $p(g|V) = N(g, V)/N(V)$.

3.3. Actor score

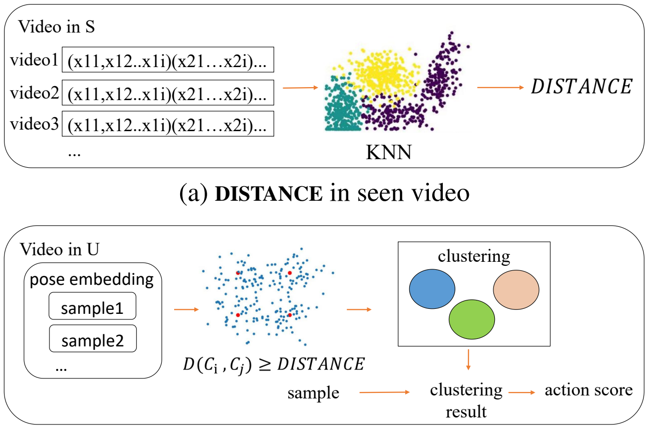
The actor score focuses on human poses. However, the video may contain the actions of multiple people, thus we need to locate the action by combing the tube score from local object information with the video score from the global objects. For localization, we combine the tube score from our spatial-aware embedding with the video score from the global objects into a score for each individual tube T as:

$$t(T, V, Z) = t_{emb}(T, Z) + t_{global}(V, Z), \tag{7}$$

Then, we select the top-scoring tubes per video and rank the tubes overall videos based on their scores for localization. In the privacy scenes, we use the bounding box obtained in the Eq. (7) for action localization and get the human pose in the process of human replacement.

A different perspective is the biggest challenge of privacy scenes. We use the pr-vipe [3] model to calculate the similarity of two images from different perspectives. They propose an approach for learning a compact view-invariant embedding space from 2D joint keypoints alone, without explicitly predicting 3D poses. Our goal is to embed 2D poses such that distances in the embedding space correspond to the similarities of their corresponding absolute 3D poses in Euclidean space.

We use the nearest neighbor search with the sequence distance to classify the seen samples. For the nearest K samples of each test sample, we take out the maximum distance of any two samples from the same class and calculate the average value as the **DISTANCE** value, which meets the following conditions:



(a) **DISTANCE** in seen video

(b) actor score in unseen video

Fig. 3. We use **DISTANCE** in Eq. 8 to initialize the cluster center and iterate the cluster center in Eq. 9. We get the actor score in Eq. 11.

$$\textbf{DISTANCE} - fastDTW(video1, video2) \begin{cases} \leq 0, & video1 \in Y_i \ \ \& \ \ video2 \in Y_i, \\ > 0, & video1 \in Y_i \ \ \& \ \ video2 \notin Y_i, \end{cases} \tag{8}$$

where Y_i represents seen class labels and function $fastDTW$ is the distance calculated after fast standard *dynamic time warping* (DTW), which is used to align two action sequences by minimizing the sum of frame matching distances.

We use **DISTANCE** to initialize the cluster centers. Suppose we know that the unseen samples are divided into n classes, and we find n samples whose distances are larger than the **DISTANCE** between each other as the initial cluster center. We use the 1NN algorithm to classify the poses of unseen samples and then iterate the cluster center. Here we suppose C_i is the cluster center of cluster T_i. The new cluster center B_i needs to meet the following condition:

$$L_{max} = max(\sum_{j=1}^{N} fastDTW(B_i, C_j) - fastDTW(B_i, C_i)), \tag{9}$$

where $fastDTW(B, C)$ denotes the distance between sample B and cluster center C. In the process of 1NN classification, we calculate the maximum distance between the point in each cluster and other cluster centers, and then the point is a new cluster center iterating in turn until the cluster centers remain unchanged.

We classify each cluster by semantic textual. The cluster T and its class name $Z_{cluster}$ need to meet the following condition:

$$T = min_{Ti \in T}(\sum_{i=1}^{N} w(e(O), e(Z_{cluster}))). \tag{10}$$

Scoring actor. According to the clustering results, we get the score of the actors,

$$t_{actor} = 1 - w(e(Z), e(Z_{cluster})). \tag{11}$$

3.4. Unified actor score

We compute the unified actor score of a video V given an action class name Z using a max-pooling operation over the scores from all tubes T_V in the video. The predicted class for video V is determined as

Table 1. Results of the ablation study of different components of UAS. Each row shows the average accuracy and the standard deviation over 5 independent runs. L: local object; G: global object; A: actor.

Component	Train	Test	Accuracy
L + G [19]	51	50	40.4 ± 1.0
L + A	51	50	39.8 ± 1.3
G + A	51	50	36.2 ± 5.8
L + G + A	51	50	**43.4 ± 2.1**

the class with the highest combined score:

$$score = arg \max_{z \in Z}(\max_{T \in T_V} t_{emb}(T, Z) + t_{global}(V, Z) + t_{actor}(A, Z)), \quad (12)$$

4. EXPERIMENTAL ANALYSIS

4.1. Datasets

We verify our method on Olympic Sports [12], HMDB-51 [13], and UCF-101 [14], and we can compare the results of our method to recent state-of-the-art models. The three popular datasets Olympic Sports, HMDB51, and UCF101 contain 783, 6766, and 13320 videos with 16, 51, and 101 categories, respectively.

4.2. Implementation details

To calculate the similarities between the semantics of actions and objects, we choose to use the skip-gram network of word2vec [15], which is trained on the meta-data of images and videos from the YFCC100M dataset [16]. To detect both actors and the local objects, we use the Faster R-CNN [17], which are trained on the MS-COCO dataset [18]. To embed poses [3], we choose $\beta = 2$ as the triplet ratio loss margin. During training, we normalize the matching probabilities within $[0.05, 0.95]$ for numerical stability.

4.3. Ablation experiment

Based on the [19], we add actor score to enhance the action recognition accuracy. While comparing with [19], we also test the impact of each component on the UCF-101 dataset.

Compared with [19] in the first row of Table 1, our proposed pose greatly improves the recognition accuracy. We observe that local objects and global objects improve the accuracy in row 2 and row 3 of Table 1, as they explore the information of objects.

4.4. Comparison to state-of-the-art

In this section, a comparison of our proposed framework with the state-of-the-art approaches for the tasks of ZSL and GZSL in action recognition is given. We follow the commonly used 50/50 splits [20], where 50 percent of all classes are seen classes and the other 50 classes are unseen classes.

For ZSL and GZSL, we observe that UAS consistently outperforms most state-of-the-art approaches on the Olympic Sports, HMDB-51, and UCF-101 in Table 2 and Table 3. Generally, privacy-preserving scenes are more suitable for GZSL. UAS not only needs to classify unseen samples but also needs to classify mixed samples (unseen and seen samples). As detailed in the table, UAS outperforms most state-of-the-art approaches. In the last row of Table 2, we replace the human body with his pose in Olympics, HMDB-51 and UCF-101. The setting of "privacy-preserving" adopted in our work reduces the diversity of human subjects, such as the invisibility of their clothes and body shape, this is exactly the purpose of privacy-preserving. From this point of view, the setting of "privacy-preserving" only focuses on the pose, which does reduce the complexity of the task. However, this simplification increases the difficulty of action recognition as the extracted poses may not correctly represent the true human actions. We note that, the pose estimation itself is not well addressed in current days. In these privacy-preserving videos, we also have a good recognition accuracy in ZSL and GZSL.

Table 2. Results on ZSL. Comparison to the state-of-the-art for zero-shot action recognition. PS: privacy-preserving scenes.

Method	Olympics	HMDB51	UCF101
SJE [21]	28.6 ± 4.9	13.3 ± 2.4	9.9 ± 1.4
IAF [22]	39.8 ± 11.6	19.2 ± 3.7	22.2 ± 2.7
Bi-Dir GAN [22]	40.2 ± 10.6	21.3 ± 3.2	21.8 ± 3.6
GGM [23]	41.3 ± 11.4	20.7 ± 3.1	20.3 ± 1.9
WGAN [24]	47.1 ± 6.4	29.1 ± 3.8	25.8 ± 3.2
OD [25]	50.5 ± 6.9	30.2 ± 2.7	26.9 ± 2.8
PS-GNN [26]	61.8 ± 6.8	32.6 ± 2.9	43.0 ± 4.9
UAS (ours)	**62.0± 3.5**	**38.1±3.5**	**43.4 ± 2.1**
UAS in PS	51.4± 4.9	33.7±3.9	35.9 ± 3.4

Table 3. Results on GZSL. Comparison to the state-of-the-art for generalized zero-shot action recognition. PS: privacy-preserving scenes.

Method	Olympics	HMDB51	UCF101
IAF [22]	30.2 ± 11.1	15.6 ± 2.2	20.2 ± 2.6
Bi-Dir GAN [22]	32.2 ± 10.5	7.5 ± 2.4	17.2 ± 2.3
SJE [21]	32.5 ± 6.7	10.5 ± 2.4	8.9 ± 2.2
GGM [23]	42.2 ± 10.2	20.1 ± 2.1	17.5 ± 2.2
WGAN [24]	46.1 ± 3.7	32.7 ± 3.4	32.4 ± 3.3
PS-GNN [26]	52.9 ± 6.2	24.2 ± 3.3	35.1 ± 4.6
OD [25]	53.1 ± 3.6	36.1 ± 2.2	37.3 ± 2.1
UAS (ours)	**54.2 ± 2.5**	**36.8 ± 2.1**	**39.5 ± 4.1**
UAS in PS	48.7 ± 2.2	32.1 ± 3.4	34.6 ± 2.7

5. CONCLUSIONS

In this paper, we focus on the privacy-preserving issues of action recognition for humans. We propose using action recognition accuracy to evaluate pedestrian privacy-preserving methods and propose UAS to recognize action in seen videos and unseen videos. Finally, experiments show that each component of UAS improves the results and UAS is better than the state-of-the-art methods. We also do experiments in privacy-preserving scenes, where we replace the human body with his pose in the datasets. The results show that UAS is suitable for these privacy-preserving scenes.

6. REFERENCES

[1] K. Yagi, K. Hasegawa, and H. Saito, "Diminished reality for privacy protection by hiding pedestrians in motion image sequences using structure from motion," in *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2017.

[2] Erdelyi, Adam, Winkler, Thomas, Rinner, and Bernhard, "Privacy protection vs. utility in visual data an objective evaluation framework," *Multimedia tools and applications*, vol. 77, no. 2, pp. 2285–2312, 2018.

[3] J. J. Sun, J. Zhao, L. C. Chen, F. Schroff, H Adam, and T. Liu, "View-invariant probabilistic embedding for human pose," in *European Conference on Computer Vision*, 2020.

[4] A. Badii, A. Al-Obaidi, M. Einig, and A. Ducournau, "Holistic privacy impact assessment framework for video privacy filtering technologies," *Signal & Image Processing*, vol. 4, no. 6, pp. 13–32, 2013.

[5] Vle Junior, H. Pedrini, and D Menotti, "Zero-shot action recognition in videos: A survey," 2019.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Y. Fu, T. M. Hospedales, X. Tao, and S. Gong, "Attribute learning for understanding unstructured social activity," *Springer, Berlin, Heidelberg*, 2012.

[9] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, *Script Data for Attribute-Based Recognition of Composite Activities*, Computer Vision – ECCV 2012, 2012.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality arxiv : 1310 . 4546v1 [cs . cl] 16 oct 2013," 2013.

[11] J. Lin, "Lin, j.: Divergence measures based on the shannon entropy. ieee transactions on information theory 37(1), 145-151," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[12] J. C. Niebles, C. W. Chen, and F. F. Li, "Modeling temporal structure of decomposable motion segments for activity classification," in *European Conference on Computer Vision*, 2010.

[13] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre, "Hmdb: A large video database for human motion recognition," *Springer Berlin Heidelberg*, 2013.

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

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[16] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[18] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick, "Microsoft coco: Common objects in context," *Springer International Publishing*, 2014.

[19] P. Mettes and Cgm Snoek, "Spatial-aware object embeddings for zero-shot localization and classification of actions," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[20] X. Xu, T. Hospedales, and S. Gong, "Transductive zero-shot action recognition by word-vector embedding," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 309–333, 2017.

[21] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and G. Ha, "Exploring semantic inter-class relationships (sir) for zero-shot action recognition," 2015.

[22] A. Mishra, A. Pandey, and H. A. Murthy, "Zero-shot learning for action recognition using synthesized features," *Neurocomputing*, vol. 390, 2020.

[23] A. Mishra, V. K. Verma, Msk Reddy, S. Arulkumar, and A. Mittal, "A generative approach to zero-shot and few-shot action recognition," *WACV, 2018*, 2018.

[24] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[25] D. Mandal, S. Narayan, S. Dwivedi, V. Gupta, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] J. Gao, T. Zhang, and C. Xu, "Learning to model relationships for zero-shot video classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2020.