DKC3D: Denoising Skeleton for Action Recognition in the dark

Zheng Yilun
School of Electrical and Electronic Engineering
Nanyang Technological University
YILUN001@e.ntu.edu.sg

***Abstract*—Human action recognition(HAR) is a popular research area and many deep neural networks have shown their success. However, the environment in reality might including environment noise such low-light and camera shifting, leading to poorer performance of traditional models. In this work, we propose Denoising Skeleton C3D(DSC3D) to address the aforementioned problem. To be specific, after extracting human skeleton, we first filter joints by confidence score, and then represent the spatial-temporal skeleton structure by sparse trajectory tensors with skeleton attention and feed it to 3D convolution backbone. It outperforms many SOTA models in ARID-small dataset. Comparing with traditional skeleton-based approach, we show sparse skeleton representation is much more better together with denoising operation and skeleton attention mechanism.**

***Index Terms*—Action Recognition, low-light environment, Sparse representation, Attention**

I. Introduction

Nowadays many approaches and datasets have been proposed to address the problem of Human Action Recognition(HAR). However, in reality, due to insufficient light condition or camera shifting, the data distribution would be different from mainstream datasets. Consequently, the additional environment noise will lower the performance of models trained on data sampled from proper lighting. How to effectively tackle the negative effect of environmental noise hasn't been well studied. Solving the aforementioned problem not only makes it more practical to deploy models in real-world scenarios but also improves models' robustness and generalization.

Many deep neural networks have shown success in HAR. From the perspective of modality, we categorize current studies into RGB-based and skeleton-based approaches. Although the RGB-based approaches provide rich information, they are sensitive to viewpoint and background, which is not suitable for HAR in the dark compared with skeleton-based approaches. Most Skeleton-based approaches first extract human skeletons along the time dimension, then feed the skeleton features into models to learn representation, which is more informative and accurate when skeletons are well extracted. Since these approaches directly use joint coordinates as skeleton features, we argue that such dense representation constrains models' capability, instead, sparse representation should be used to further stimulate the potential of skeleton-based approaches. Furthermore, they simply select skeletons according to a fixed threshold, generating insufficient information or more noise information in the context of the low-light environment.

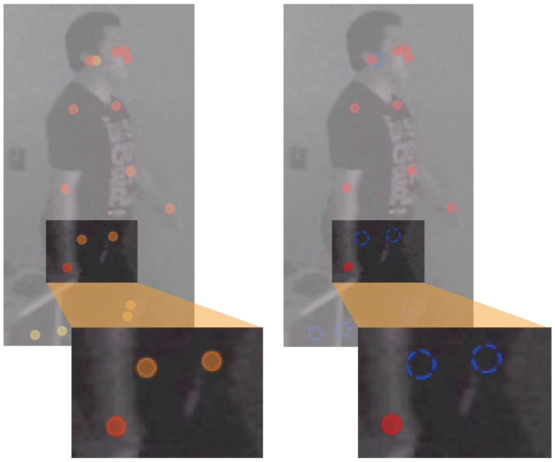

Fig. 1. Example of Denoising in DKC3D. Left: Joints generated by pose estimator. The color gradient of the joint points from red to yellow indicates a decrease in confidence. Right: Joints after denoising. Blue dashed dots denote dropped joints with high environmental noise.

In this work, we propose DSC3D. To address the problem of insufficient information extracted in low-light environment, we use a lower threshold to get more skeleton from low-light videos, which obviously includes more information as well as noise. As illustrated in Figure1, to reduce the dark environment noise, we further drop joints by joint confidence scores to get denoised temporal-spatial skeleton sequences, and then represent the sequences by sparse trajectory tensors with skeleton attention. Finally, we feed the sparse trajectory tensors to a C3D backbone to learn how to recognize human actions in the dark. To validate the effectiveness of DSC3D, our experiments show it outperforms many SOTA models in ARID-small dataset.

The main contributions of this work include three aspects: 1) We propose DSC3D, a robust model for environmental noise in the dark. 2) Our experiments show that among skeleton-based approaches in HAR, sparse skeleton tensor representation is better than coordinates-based representation. 3) Our skeleton attention extracts the relation of joints in both local view and global view.

II. Related Work

RGB-based action recognition. RGB videos, containing rich appearance information, are easy to obtain and operate [1]. Early approaches treat RGB images and time dimension individually, they use RNNs to aggregate representation extracted from 2D-CNN model [2] [3]. While treating spatial and temporal view independently does benefit from pre-trained 2D-CNN models, they focus more on 2D image-based recognition and ignore rich semantic relations to temporal modality. To extract spatial and temporal features simultaneously, C3D [4] introduce 3D-CNN with small filter cores. Even though vanilla 3D-CNN shows great capability in video representation, they lose the benefit of pre-trained 2D-CNN models. I3D [5] solve the problem by adding a temporal dimension to inflate 2D structure to 3D structure. To better capture long-range dependencies along the temporal dimension, None-local [6] compute attention scores of different positions with self-attention mechanism. Furthermore, TimeSformer [7] borrow the idea from ViT [8], designing multiple attention to capture relations jointly(Space-time) or independently(Time, width, or height). To catch both stationary and dynamic features, SlowFast [9] propose a slow-pathway for semantic information and a fast-pathway for rapidly changing information, and then a lateral fusion is applied to the two representations.

Skeleton-based action recognition. Human skeleton sequences contain simple structural information, which could be represented by joints or bones. Considering the temporal attribute of skeleton sequences, Du et al. [10] introduce end2end RNN structure and divide the whole joints into 5 parts. To automatically learn co-occurrence connections and inner connections of different joint parts, Zhu et al. add fully connected layers before feeding joints to LSTM. While the above approaches mainly concentrate on the aggregation of joint information, other approaches focus on the design of RNNs structure. To extract dynamical information, dRNN [11] take the derivative of the state and add it to the gates of LSTM. To encourage information sharing between joints, P-LSTM [12] share output gates of LSTM with respect to body parts. Meanwhile, CNN-based approaches are proposed because of their advantage in extracting topology. Kim et al. [13] use TCN to capture both spatial features together with temporal skeleton sequences. To alleviate the bias from the viewpoint and make full use of joint information, Hou et al. [14] project 3D points to three orthogonal planes and feed them to CNNs with a late fusion process. After that, PoseC3D [15] use stacked 2D heatmaps to represent each joint as a channel and feed it to 3D-CNNs. Apart from that, based on Graph Convolution Network(GCN) [16], ST-GCN [17] is proposed to capture spatial and temporal simultaneously with respect to skeleton sequences, the neighbor of one node including its nearby joints of the current moment, last moment and next moment. To address the problem of long-range joint relationship, G3D [18] increase the receptive field of GCN by skip-connections.

Representation of skeleton sequences. Although many skeleton-based approaches have been proposed, there is no general representation of skeleton sequences, which greatly effect the performance of different models. The easiest way to represent skeleton sequences is using coordinates of joints directly [12], [13], [17], [18] or smoothed coordinates with Savitzky-Golay filter [10], [19]. However, using coordinates only may not be sufficient to capture complex spatial and temporal relationships, Veeriah et al. [11] introduces more hand-craft features including position, angle, offset, velocity, and pairwise distances. To enhance the robustness and scalability of skeleton representation, heatmaps are introduced [14], [15]. Hou et al. [14] project 3D skeleton points to three orthogonal 2D heatmaps, while PoseC3D [15] only focus on 2D heatmap in one perspective. Different from [14] which project the whole skeleton in one plane, PoseC3D project each joint independently to one channel, which further increases the sparsity of the representation.

Action recognition in the dark. Few studies [20], [21] focus on action recognition in the dark. In reality, due to environmental noise, current models are more likely to fail [22]. It is necessary to generalize models' capability to different environments. 3D-SDCF [21] apply low-light enhancement techniques, and then feed enhanced images and original images independently into 3D-Convolution structures. Finally, a late fusion technique is applied to the output. Based on the two-stream fusion framework, DarkLight [20] adopt weight sharing in 3D-Convolution and self-attention in the fusion stage.

III. Skeleton-Based C3D

This section introduces our proposed model as well as low-light enhancement techniques. In sectionIII-A, we briefly introduce the proposed model Denoising Skeleton Conv3D(DSC3D). SectionIII-B discusses different enhancements for action recognition in the low-light environment. SectionIII-C and sectionIII-D introduce the denoising technique for skeleton sequences and skeleton attention for joint channels respectively.

A. Overview

Figure2 shows the framework of DSC3D. For a video, we apply a low-light enhancement technique to help the pose estimator better recognize joints. Since it is harmful to provide models with additional environmental noise, we drop the joints according to confidence scores, which indicate how much the joints are affected by the environmental noise. Then the joints are projected to the 2D heatmaps with respect to their positions, in which we treat each joint as a channel. After that, a skeleton attention layer is applied to capture the spatial relationships among these joints. Finally, the stacked heatmaps are fed into a 3D convolution module to get the prediction.

B. Enhancement

Pixels of videos in the dark mostly gather around dark tones where human eyes or deep neural models are more insensitive compared with normal illumination conditions. Therefore it is necessary to apply low-light enhancement techniques to

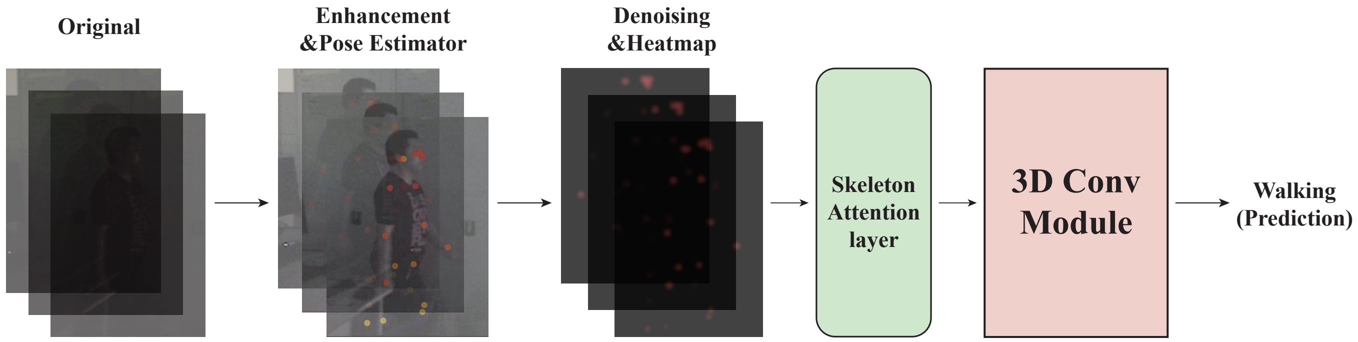

Fig. 2. DSC3D framework. After applying low-light enhancement and pose estimator on the original video, all the joints are projected to 2D heatmaps with denoising operation. Then a skeleton attention layer automatically constructs the joint relation and feeds the representation to 3Dconv Module for model inference.

these videos. There we use a simple enhancement technique called gamma correction, which is effective according to our observation. Although there are many different deep-learning-based enhancement techniques, gamma correction is stable and efficient. The gamma correction process can be represented as

$$V_{out} = cV_{in}^{\gamma} \tag{1}$$

where V_{out}, V_{in}, and c refer to output pixels, input pixels and constant respectively, and γ is a hyper-parameter that controls the degree of correction.

C. Denoising Skeleton

Previous skeleton extraction methods for action recognition mostly select frames according skeleton confidence score. In other word, the selected frames have to satisfy the minimum requirement that properly form a human skeleton. However, videos in the dark contain more environment noise, which makes it harder to recognize human skeleton. In that case, if we still follow the traditional skeleton extraction process, less frames will be utilized, leading to insufficient training samples. To make the full use of low-light videos, we lower down the threshold for skeleton extraction, which inevitably introduces more information as well as irregular skeletons and incorrect joints. Luckily, the quality of different parts of a skeleton is different, that is, each joint has its own confidence score. Based on the observation, we only keep the joints with high confidence scores and drop the joints with high environment noise, which is beneficial for action recognition in the dark since more information is obtained. The skeleton extraction and denoising process can be represented as

$$\begin{aligned} \mathbf{s} &= [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T], where\ s_i > Th_{skeleton} \\ \mathbf{s}_i &= [j_1, j_2, ..., j_K], where\ j_i > Th_{joint} \end{aligned} \tag{2}$$

where $Th_{skeleton}, Th_{joint}$ and $\mathbf{s}_i$ refer to skeleton threshold, joint threshold and the skeleton extracted from one frame respectively.

After the skeleton extraction and denoising process, we project skeletons to 2D heatmaps, in which we borrow the idea proposed in PoseC3D [15]:

$$\mathbf{J}_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}} * c_k \tag{3}$$

each pixel (i, j) in 2D heatmaps of one joint can be calculated by the above equation, where joint coordinate (x_k, y_k) is represented as a gaussian blur with joint confidence score c_k.

D. Skeleton Attention

After getting 2D heatmaps of each joint, we use skeleton attention to capture joint relationships. Compared with handcraft joint connection, our skeleton connection is built based on the attention mechanism, which could capture local and global connections automatically. We use multi-head attention to capture complex joint relations and residual connections to make the model converge faster. The skeleton attention can be represented as

$$\begin{aligned} head_j &= \sigma(\frac{QK_T}{\sqrt{d_k}})\mathbf{s}_i \\ attn &= [head_1 \circ ... \circ head_M]W^O \\ \mathbf{s}_i^{'} &= attn \cdot \mathbf{s}_i + \mathbf{s}_i \end{aligned} \tag{4}$$

After the skeleton attention operation, we feed all the representations into a 3D convolution module to get the prediction.

IV. Experiments

A. Dataset

ARID-small [23]. We train and evaluate all the methods on this dataset. The dataset contains 10 classes (including drinking, jumping, picking, pouring, pushing, running, sitting, standing, turning, and walking) in total. We follow the default splitting ratio of this dataset, where the number of videos in the train set, test set, and validation set is 750, 320, and 450 respectively.

NTU RGB-D Skeleton [24]. To better enhance our model and explore the effect of the pre-tained model with larger datasets, we further pretrain our method on this large dataset. Since both datasets are sampled by the same research group, there should be less difference between these two datasets. This large dataset includes 56880 samples of 60 actions, which include individual people and multiple people.

B. Implementation

For the data preparation, we first split each video into multiple frames and apply GID low-light enhancement where we set the threshold to 2.5. Then we use HRNet [25] to capture human skeletons. We choose the skeleton with a fixed threshold, which is set to 0.3. After capturing human skeletons from videos, we apply the denoising technique and skeleton attention as described in section III-C and section III-D respectively, where the denoising threshold is set to 0.3 and attention head number is set to 10. As for the 3D Convolution module, we choose SlowOnly [9] as our backbone, in which input pass through multiple convolution layers with an increasing channel number and decreasing filter size. Then a global average pooling layer and a fully connected layer is applied to get action prediction.

For the training, we set the training batch size to 16 and use SGD optimizer where the learning rate is set to 0.2, momentum is set to 0.9, and decay weight is set to 3×10^{-4}. Since the extracted skeleton is fast to train, the training epoch is set to 1000. Our experiment is conducted with Ubuntu 18 platform with Intel(R) Xeon(R) Silver 4314 CPU and three NVIDIA RTX A5000 GPUs.

C. Comparison

We compare our method with many baselines and SOTA models, which are briefly introduced as follows:

- **R3D** [26] use multiple 3D convolution layers to capture spatial and temporal features simultaneously.
- **MC3** [26] first use three 3D convolution layers to capture spatial-temporal features and then use 2D convolution layers to extract high-level semantic features.
- **R(1+2)D** [26] decompose 3D convolution into two computing-efficient spatial and temporal convolutions, implemented with 2D convolution and 1D convolution respectively.
- **SlowOnly** [9] is a variant of SlowFast [9] model, where two channel is used to capture spatial and temporal features respectively. SlowOnly only use output representation of slow-path to predict labels.
- **ST-GCN** [9] first construct 3D-GCN with human joints in both spatial and temporal channels, and then construct adjacency matrix with local connections and apply GCN operation to get the final representation.
- **PoseC3D-Limb** [15] first project human limb to 2D heatmap with a gaussian function and then feed the stacked 2D heatmaps to 3D convolution framework.
- **PoseC3D-Joint** [15] is a variant of the former approach, the only difference is PoseC3D-Joint project human joints instead of human limbs into 2D heatmaps.
- **DSC3D +Attn** is our proposed method with skeleton attention.
- **DSC3D +Denoising +Attn** is our proposed method with skeleton attention and denoising operation.

Method	***Top1 Acc***	***Top5 Acc***
R(1+2)D	0.2656	0.7312
MC3	0.2781	0.6906
R3D	0.4750	0.8688
SlowOnly	0.2094	0.6531
ST-GCN	0.4719	0.9031
PoseC3D-Limb	0.7063	0.9563
PoseC3D-Joint	0.8469	0.9844
DSC3D	0.8625	0.9844
DSC3D +Pretrain	**0.8781**	**0.9906**

TABLE I
Top-1 accuracy and Top-5 accuracy in test set of our experiment.

Method	***Top1 Acc***	***Top5 Acc***
PoseC3D-Joint	0.8222	**0.9889**
DSC3D	0.8511	0.9867
DSC3D +Pretrain	**0.8533**	**0.9889**

TABLE II
Top-1 accuracy and Top-5 accuracy in validation set of our experiment.

Table IV-C shows the performance of our proposed DSC3D and other methods on test set, where we use Top-1 accuracy and Top-5 accuracy to evaluate the performance. Among all the other approaches, DSK3D achieves the best performance both in Top-1 accuracy and Top-5 accuracy, which shows our improvement on skeleton-based approach is effective. We also notice that, except from R3D, Traditional RBG-based approaches behave much poorer than skeleton-based approaches, even if they can get good result in other datasets with normal light condition, which shows the additional noise introduced by dark environment is harmful for these traditional models. For skeleton-based approaches ST-GCN and PoseC3D, their performance in Top-1 accuracy and Top-5 accuracy is much better. It shows the skeleton-based approaches is much robust to environment noise, since the prediction is generated based on extracted skeleton. We also find that PoseC3D is much better than ST-GCN, indicating the effectiveness of stacked 2d heatmaps. We infer the huge improvement of heatmap representation shows that traditional approaches, where joint representation is represented directly by coordinates, are sensitive to distortion.

Table IV-C shows the experiment result in the validation set. Since the limited time is given, we only compare the validation result on several competitive methods. It shows the top-1 accuracy of DSC3D is better than PoseC3D-Joint, indicating that DSC3D is still effective under the setting of the validation set. Besides, there is a negligible difference between the test set and the validation set, which shows our training process is reasonable and has no label leaking.

As for DSC3D, we find that the Top-1 accuracy is improved after applying skeleton attention and denoising technique, which indicates the effectiveness of capturing the connections of different joints and removing additional useless environment noise.

V. Conclusion

In this work, we propose DSC3D for action recognition in the dark. Our experiment shows that it is helpful to apply the denoising technique and skeleton attention in skeleton sequences in a low-light environment. As for the skeleton-based approach, we conclude that heatmap representation is helpful compared with coordinate representation. However, the challenge of the low-light environment still remains unsolved and its performance still needs to be improved. We believe 2D heatmaps still lack depth estimation, which could provide additional information. What's more, all of the approaches should be evaluated in larger and more challenging datasets to eliminate the influence of the sampling environment and explore the robustness of models in different environments.

References

[1] Z. Sun, J. Liu, Q. Ke, H. Rahmani, M. Bennamoun, and G. Wang, "Human action recognition from various data modalities: A review," *CoRR*, 2020.

[2] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021,Virtual Event, Austria, May 3-7, 2021*, 2021.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.

[11] V. Veeriah, N. Zhuang, and G. Qi, "Differential recurrent neural networks for action recognition," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015.

[12] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

[13] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.

[14] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, 2018.

[15] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[17] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020.

[19] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, 2016.

[20] R. Chen, J. Chen, Z. Liang, H. Gao, and S. Lin, "Darklight networks for action recognition in the dark," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops,CVPR Workshops 2021, virtual, June 19-25, 2021*, 2021.

[21] A. Ulhaq, "Action recognition in the dark via deep representation learning," in *IEEE International Conference on Image Processing, Applications and Systems, IPAS 2018, Sophia Antipolis, France, December 12-14, 2018*, 2018.

[22] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A new dataset for recognizing action in the dark," *CoRR*, 2020.

[23] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A new dataset for recognizing action in the dark," *CoRR*, vol. abs/2006.03876, 2020.

[24] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1010–1019, IEEE Computer Society, 2016.

[25] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[26] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.