DarkLight Networks for Action Recognition in the Dark

Rui Chen[*], Jiajun Chen[*], Zixi Liang, Huaien Gao, Shan Lin[†]
Guangzhou Xi Ma Information Technology Company
101 Waihuan Xi Road, Da Xue Cheng, Guangzhou, Guangdong, China 510006
{2584775140, 1311642081, 781311601, 1496609}@qq.com; alice333.happy@163.com

Abstract

Human action recognition in the dark is a significant task with various applications, e.g.*, night surveillance and self-driving at night. However, the lack of video datasets for human actions in the dark hinders its development. Recently, a public dataset ARID has been introduced to stimulate progress for the task of human action recognition in dark videos. Currently, there are multiple models that perform well for action recognition in videos shot under normal illumination. However, research shows that these methods may not be effective in recognizing actions in dark videos. In this paper, we construct a novel neural network architecture: DarkLight Networks, which involves (i) a dual-pathway structure where both dark videos and its brightened counterpart are utilized for effective video representation; and (ii) a self-attention mechanism, which fuses and extracts corresponding and complementary features from the two pathways. Our approach achieves state-of-the-art results on ARID. Code is available at:* https://github.com/Ticuby/Darklight-Pytorch

1. Introduction

Action recognition (AR) in the presence of dark lighting conditions is a challenging task in computer vision. So far, there is still a lack of relevant research work. Though there has a rise of research interest with the video process tasks in the dark environment, *e.g.* [3, 13], such research focused more on enhancing the visibility of dark videos.

The lack of research for action recognition in dark videos may be attributed to the following two reasons: (i) the lack of sufficient datasets for such an exploration, (ii) ineffective data enhancement methods which cause unexpected data destruction, resulting in a lower classification accuracy. As stated in [33], distinct characteristics of real dark videos cannot be replicated by synthetic dark videos. In other words, it is irrational to transform large numbers of available videos into dark videos for training. It also empirically demonstrated that current frame enhancements which could improve dark video frames visually may not bring consistent improvements for action recognition accuracies of dark videos.

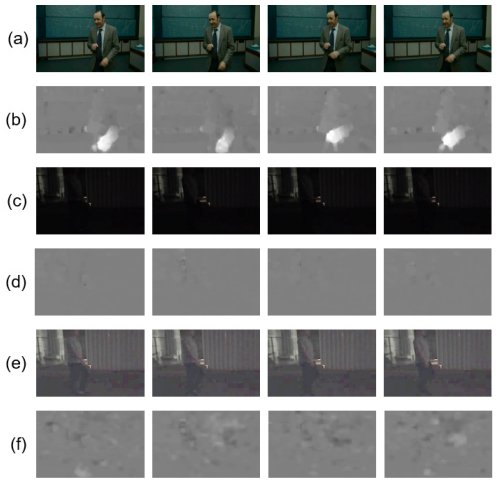

Figure 1. From top to bottom: (a) Sampled frames from HMDB51, where videos are shot under normal illumination, sample from HMDB51; (b) optical flow extracted from (a); (c) sampled frames from ARID, where videos are shot under low illumination; (d) optical flow extracted from (c); (e) enhanced frames of (c), the enhancement is performed with Gamma Intensity Correction (GIC); and (f) optical flow extracted from (e).

Over the past decade, the task of action recognition has received considerable attention from the vision community owing to the flourish of applications relating to the visual domain such as surveillance [34, 15] and smart homes [21, 16]. Most current works for action recognition can be generally classified into two kinds of architectures or frameworks, namely (1) 3D Convolutional Neural Networks (CNN) [12, 29, 23], and (2) two-stream [26, 31, 24]. Methods with two-stream architecture have shown to outperform

[*]Equal Contributions
[†]Corresponding Author

Optical flow - the pattern of apparent motion of image objects between two consecutive frames caused by the movement of object or camera

Complementary features - two people or things that are complementary are different but together form a useful or attractive combination of skills, qualities or physical features

Gamma correction - responsible for performing nonlinear methods on the pixels of the input image and thereby remodeling the saturation of the image

two-stream

3D-CNNs thanks to their ability in utilizing complementary features extracted parallel with each other. However, conventional two-stream methods usually involve computation or estimation of optical flow, which needs high computational power and large storage resources. In addition, optical flow is useful only when the change of pixel values are rather significant across adjacent frames, which however does not hold for dark videos. As shown in Figure 1, it can be observed that the optical flow extracted from dark videos is unclear visually, and contains little information. Even for frame-enhanced videos, *e.g.*, through Gamma Intensity Correction (GIC) (Figure 1(e)), the optical flow extracted is still of inferior quality compared to that extracted from normal illuminated videos, such as videos from HMDB51. Therefore, instead of using optical flow, we introduce a novel dual-pathway structure for feature extraction. The structure includes two pathways, namely Dark pathway and Light pathway, where each pathway has an input and a feature extractor. The Dark pathway is fed with original input frames, but the input of the Light pathway is pre-processed with a traditional image brightening algorithm, which hypothetically can provide complementary features for the Dark pathway in video representation.

Recently, self-attention has been introduced as an effective module for feature extraction, and has been successfully applied in multiple tasks, *e.g.*, machine translation [30] and image classification [5]. A self-attention module computes the response at a position in a sequence by attending to all position and taking their weighted average in an embedding space. Inspired by the success of self-attention models, multiple works apply it in video action recognition, which can be classified into two forms. One is purely based on the self-attention mechanism [1], the other combines 3D-CNN network with the self-attention blocks [7, 14, 8, 20]. Until now works which completely dependent on self-attention mechanism performance have been inferior to that of CNN in action recognition. Nevertheless, hybrid architectures [14, 8] that combine CNN and attention have recently exhibited competitive results. Inspired by [14], we utilize a self-attention block to fuse features from the two pathways.

Overall, our contributions are: (i) We introduce Dark-Light networks (Section 3) including two pathways for learning complementary features and a self-attention block for fusing the features. (ii) We show that the proposed method achieves state-of-the-art results on ARID (Section 4). (iii) We prove that using a self-attention block to fuse and extract features can obtain a higher accuracy (Section 4).

2. Related Works

3D-CNNs for Action Recognition In 3D convolution, filters are designed in a 3D fashion, where channels and temporal information are represented in different dimensions. The first 3D-CNN for AR is C3D [28]. Subsequently, networks such as 3D-ResNet [10] and 3D-ResNext [11] are deeper and larger 3D-CNN. More recently, R(2+1)D [29] architecture is proposed to decompose 3D spatio-temporal convolutions into spatial and temporal convolutions, which further improves the effectiveness of video feature extraction.

Two-stream Methods for Action Recognition Videos contain a wide range of information with different modalities. The diverse modalities have been made use of by many research work, which includes depth information [25], RGB information [18], skeleton information [17], and optical flow information[24]. Two-stream methods utilize the diversity of information in videos, creating two or even more pathways for parallel feature extraction, resulting in video features containing rich modalities of information. Among the different modalities, optical flow information has been used widely for two-stream methods [26, 31, 24], as optical flow information is widely recognized as an effective complementary towards RGB information. However, modalities apart from RGB information would be obtained easily. Depth information needs to be obtained from a special sensor, while both skeleton and optical flow require high computation cost and large storage. In this work, two-stream architecture is also adopted with the goal of obtaining complementary features to the RGB features, which may contain inadequate information. To achieve this, we observe that frame-enhanced videos, which are visually clearer, ought to contain more information with regards to the action. Therefore, such frame-enhanced videos are utilized for our two-stream architecture.

Self-attention Mechanism The primary work which introduces self-attention as an exclusive building block for video understanding is [1]. The network is built from the standard Transformer architecture adapted temporal attention and spatial attention separately within each block. However, models based solely on the self-attention mechanism require excessive amount of training data. Instead, more recent methods tend to integrate self-attention into feature extractors, *e.g.*, in [20] each frame is characterized by a 2D spatial network, and then use a temporal attention-based encoder to gain a classified (CLS) token for classification. The work [14] adopts BERT [4] to replace the conventional temporal global average pooling layer at the end of 3D-CNN for using better temporal information. Inspired by previous works with self-attention, we employ self-attention as an effective module for fusing the features obtained by the two pathways.

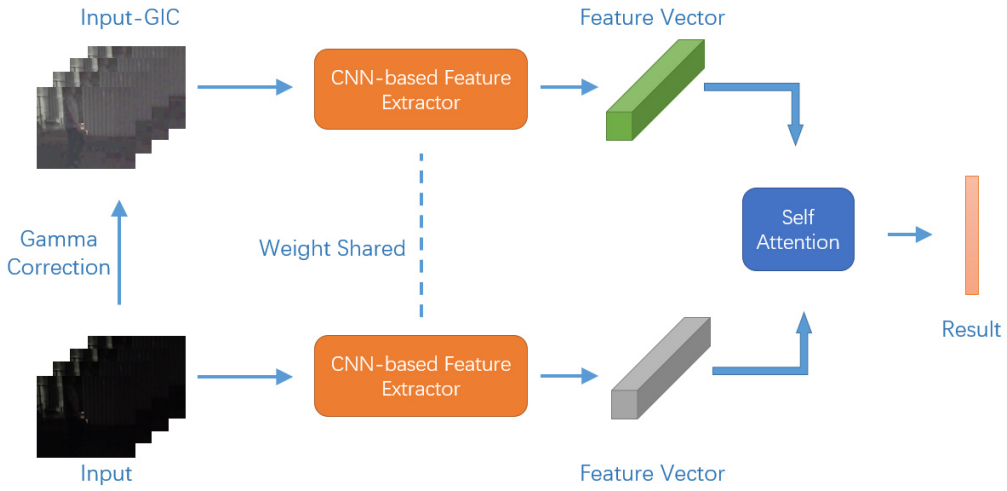

Figure 2. The architecture of DarkLight Network. The input is a sequence of dark frames for the first pathway, denoted as the Dark pathway. The dark video input is simultaneously enhanced through Gamma Intense Correction (GIC), forming the input for the second pathway, denoted as the Light pathway. Taken them into a weight-shared CNN feature extractor, feature vectors are extracted separately to input a self-attention blocks in parallel to fuse and extract more useful spatio-temporal features. Finally, the result is obtain from the output Y_{cls} in self-attention blocks.

3. Method

In this section, the proposed method, DarkLight networks, will be introduced. The overall framework, as show in Figure 2, is based on two pathways: the Dark pathway and the Light pathway. The input of the Dark pathway is a sequence of dimmed frames and Light pathway is raw frames applying GIC. After a weight shared CNN feature extractor, the feature vector is extracted separately and inputted in parallel to a self-attention module. Finally, the result is obtained from the output Y_{cls} in self-attention blocks. In Section 3.1, we present how to construct the Light pathway, and how to obtain features from dual-pathway. In Section 3.2, we explain the principles of self-attention blocks and their effects in this work.

3.1. DarkLight Pathways

There are many image enhancement methods, including traditional Histogram Equalization (HE) [27], Gamma Intensity Correction (GIC) LIME [9], BIMEF [35] and deep learning KinD [36]. Although neural network performs well in visualization, they destroy the distribution of data, which is awful for video understanding. A simple and effective conventional GIC technique is adopted to brighten dimmed frames, derived by Equation 1.

$$GIC(p) = p_{max}(\frac{p}{p_{max}})^{\frac{1}{\gamma}} \tag{1}$$

where p is the value of a pixel with the range of $[0, 255]$, p_{max} is the maximum intensity of the input and γ indicates the degree of luminance increase. When $\gamma > 1$, the overall gray value of the image will become larger, as exhibition in Figure 3.

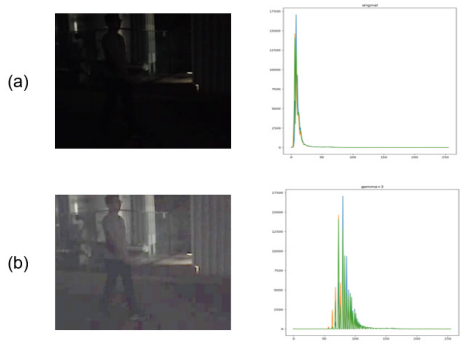

Figure 3. (a) A frame in ARID and its histogram. (b) Apply GIC, gamma = 3, in the frame of (a) and its histogram.

Given the input dark video as a sequence of clips, denoted by $I \in \mathbb{R}^{3\times T_0\times H\times W}$, where number 3 indicates the RGB channels, T_0 is the number of video frames, H, W are the height and width. By GIC, pixels of each input frame are computed by Equation 1, resulting in $I_{GIC} \in \mathbb{R}^{3\times T_0\times H\times W}$. Subsequently, both I and I_{GIC} are separately inputted to the weight-shared CNN feature extractor. From the features extractor, we get two feature vectors containing the complete spatio-temporal information from the

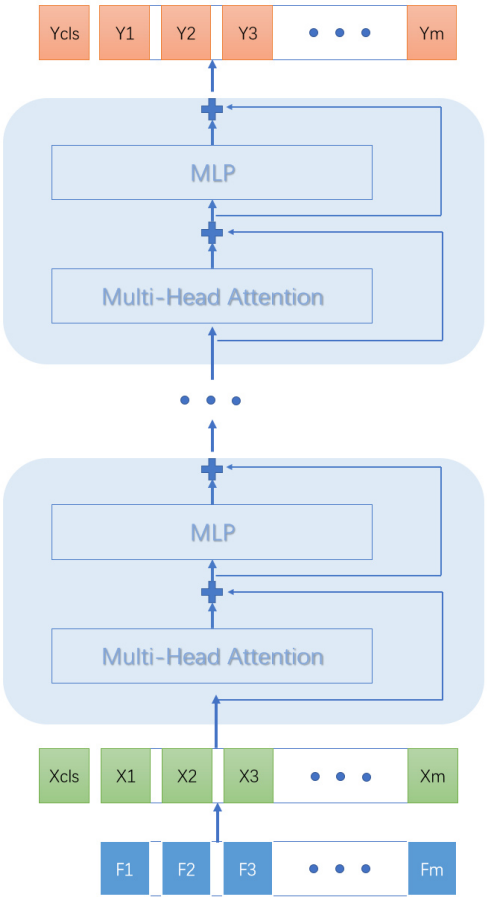

Figure 4. $F_1 \ldots F_m$ denote the features of input. $X_1 \ldots X_m$ are the encoding vectors of features, and X_{cls} is an adding extra learnable classification token. The area of blue is the basic block in self-attention, including Multi-Head and MLP, which will be repeated in the stack. In addition, LayerNorm is omitted for brevity.

Dark and Light ways, following by Equation 2,3. Next, both of them input to self-attention blocks in parallel, the purpose of which is to fuse and extract beneficial features from two pathways.

$$F_{Dark} = f(I) \tag{2}$$

$$F_{Light} = f(I_{GIC}) \tag{3}$$

Here $f()$ refers to CNN-based feature extractor. F_{Dark}, F_{Light} separately mean the features extracted from the pathways Dark and Light.

3.2. Self-attention Blocks

A seminal work applying self-attention mechanism is the Non-local Neural Networks [31] in video classification. A self-attention module computes the response at a position in a sequence by attending to all position and taking their weighted average in an embedding space. Inspired by [14], which studies how to remove temporal global average pooling with BERT, we take the self-attention blocks, as shown in Figure 4, to select more helpful spatio-temporal features from two pathways for action recognition.

After extracting features from two flows, we obtain F_{Dark} and F_{Light}, the shape of both is $D \times m$ where m can be regarded as the dimensionality of temporal and D represents the number of short-term characteristics of adjacent frames, to put them into the self-attention mechanism. First of all, the input features are added to a learnable positional embedding to encode each feature, following by Equation 4:

$$X_i = F_i + e^i_{pos} \quad (i = 1, 2, \cdots, m) \tag{4}$$

Where X_i denotes the encoding vector, which contains location information, and e^i_{pos} is a learnable positional embedding where $e_{pos} \in \mathbb{R}^{D \times m}$. In addition, a learnable CLS token is also considered in the encoding as X^0, $X \in \mathbb{R}^{D \times (m+1)}$. There are L encoding blocks in the self-attention, and each block l, query/key/value vector is computed from the output of the last block, as follows Equation 5,6,7:

$$q_i^{(l,h)} = W_Q^{(l,h)} \mathcal{L}(X_i^{l-1}) \quad \in \mathbb{R}^d \tag{5}$$

$$k_i^{(l,h)} = W_k^{(l,h)} \mathcal{L}(X_i^{l-1}) \quad \in \mathbb{R}^d \tag{6}$$

$$v_i^{(l,h)} = W_v^{(l,h)} \mathcal{L}(X_i^{l-1}) \quad \in \mathbb{R}^d \tag{7}$$

Where $\mathcal{L}$ is LayerNorm, $h = 1, \ldots, H$ is the index of multiple attention heads, and W_q,W_k,W_v are all the weight matrices, and the latent dimensionality $d = D/H$.

Self-attention weights are computed via dot-product, given by Equation 8,

$$\alpha^{(l,h)} = softmax(\frac{q_i^{(l,h)^T}}{\sqrt{d}} k_i^{(l,h)}) \ (i = 0, \ldots, m) \tag{8}$$

and encoding X_i^l at block l is obtained by the weighted sum of value vectors using α from each attention head, as follows Equations 9,10,11.

$$s_i^{(l,h)} = \sum_0^m \alpha_i^{(l,h)} v_i^{(l,h)} \tag{9}$$

$$X_i^{'(l)} = W_o \begin{bmatrix} s_i^{(l,1)} \\ \vdots \\ s_i^{(l,H)} \end{bmatrix} + X_i^{(l-1)} \tag{10}$$

$$X_i^{(l)} = MLP(\mathcal{L}(X_i^{'(l)})) + X_i^{'(l)} \tag{11}$$

The classification token Y_{cls} obtained from the final block, goes through a FC layer and argmax function to return the final forecast result, as follow Equation 12:

$$Result = Argmax(FC(Y_{cls})) \tag{12}$$

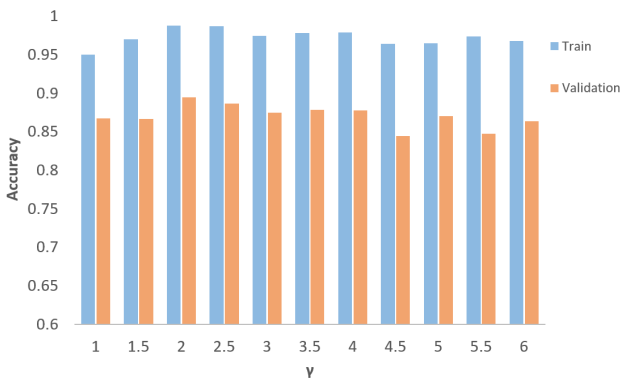

Figure 5. The experiment for selecting the value of the hyper-parameter γ in DarkLight networks.

Method	Top-1	Top-5
C3D	39.17%	94.17%
3D-ShuffleNet	44.35%	93.44%
3D-SqueezeNet	50.18%	94.17%
3D-ResNet-18	54.68%	96.60%
Pseudo-3D-199	71.93%	98.66%
I3D-Two-stream	73.39%	97.21%
3D-ResNext-101	74.73%	98.54%
DarkLight-ResNeXt-101	**87.27%**	**99.47%**
DarkLight-R(2+1)D-34	**94.04%**	**99.87%**

Table 1. The Top-1 and Top-5 accuracy results of a few competitive models and ours.

4. Experiments

4.1. Experimental Details

We conduct experiments on the first benchmark datasets for AR in the dark: ARID [33], which consists of 3,784 video clips in 11 action categories with lower brightness and contrast than other AR video datasets. We report the average Top-1 and Top-5 accuracies of three splits.

Our experiments use PyTorch [22]. The input is a sequence of frames, whose size is $3 \times 64 \times 112 \times 112$. If the frames of the video clip are less than 64, we take the method of looping frames. As for the feature extractor, we adopt ResNeXt-101 [32] and R(2+1)D-34 [29] without the average temporal pooling at the end. ResNeXt-101 is constructed by repeating a building block that aggregates a set of transformations with the same topology. R(2+1)D-34 decomposes 3D convolution into 2D spatial convolution and 1D temporal convolution. And they are separately pre-trained on Kinetics-400 [2] and IG65M [6] to accelerate our training. Extracted by ResNeXt-101 or R(2+1)D-34, both the shape of F_{Dark} and F_{Light} are 512×8. We set $L = 12$, $H = 8$ on the number of the blocks and the heads in self-attention architecture, following by [4]. For training, the ADAMW [19] optimizer with a learning rate 10^{-5} is utilized.

We conducted experiments on the hyper-parameter γ from 1 to 6 at intervals of 0.5, and each experiment is carried out 20 epochs, as shown in Figure 5. We observe that $\gamma = 2$ obtain more excellent results. Therefore, in all subsequent experiments, the $\gamma = 2$ is set in GIC.

4.2. Results and Comparisons

The results of our method and current competitive 3D-CNN based model in action recognition are recorded in Table 1, most of the data come from [33]. We notice that the Top-5 accuracy is relatively high in all methods because of the small number of classes in ARID dataset and our network achieves the best results on the benchmark datasets. From Table 1, we find that different CNN-based feature extractors can work effectively and R(2+1)D-34 is 6.77% higher than ResNeXt in Top-1 accuracy. More specifically, comparing I3D-Two-stream architecture [2] using the optical stream and the original frames as the input, we find that the Top-1 accuracy of our best is increased 20.65% by the I3D-Two-stream network, which proves not only the proposed method is powerful but the optical flow may not be useful for AR in the dark. Meanwhile, in comparison with 3D-ResNet-18 and 3D-ResNet-101, we discover that the deeper network layers, the higher effect could be achieved, but the performance of DarkLight-R(2+1)D-34 structure with 34 layers is 19.31% better than the 101 layers in 3D-ResNet-101. In summary, the comparisons illustrate DarkLight network is far better than many other excellent models based on 3D-CNN or two-stream in Top-1 accuracy.

We further perform ablation experiments on DarkLight-R(2+1)D-34 to explore the role of each part, as shown in Table 2. Without a self-attention mechanism, taking the dual-pathway as the input increases 1.93% from only using the Dark pathway and 0.83% from only using the Light pathway in Top-1 accuracy. This shows that the features from the Dark pathway and Light pathway have consistent and complementary information. With the self-attention module, taking a dual-pathway as input obtains an improvement of 1.6% by only using the Dark pathway and 1.29% by only using the Light pathway in Top-1 accuracy. It illustrates that self-attention blocks can catch more important spatio-temporal features for AR. In conclusion, applying the dual-pathway with a self-attention block can achieve the best result.

5. Conclude

In this work, we propose a new architecture for action recognition in the dark while avoiding the use of optical flow. The traditional image process GIC, which improves the brightness of dimmed images, is taken to form another

Method	Top-1	Top-5
R(2+1)D-34-Dark	90.45%	98.11%
R(2+1)D-34-Light	91.55%	99.51%
R(2+1)D-34-DarkLight	92.38%	99.17%
R(2+1)D-34-Dark-SA	92.44%	99.70%
R(2+1)D-34-Light-SA	92.75%	99.51%
R(2+1)D-34-DarkLight-SA	94.04%	99.87%

Table 2. Ablation experiences, where -Dark, -Light, -DarkLight, means using the corresponding view of Dark, Light or both. -SA denotes using self-attention mechanism to fuse and select features from two pathways.

pathway named Light that provides complementary information for the Dark pathway. Meantime, a self-attention mechanism is applied to fuse and select more beneficial spatiotemporal information from dual-pathway, namely Dark and Light. The experiments indicate the proposed method is powerful.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 2

[2] 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

[3] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3185–3194, 2019. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019. 5

[7] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 2

[8] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. *arXiv preprint arXiv:2012.10671*, 2020. 2

[9] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016. 3

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017. 2

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 2

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 1

[13] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7324–7333, 2019. 1

[14] M Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. *arXiv preprint arXiv:2008.01232*, 2020. 2, 4

[15] Karani Kardas and Nihan Kesim Cicekli. Svas: surveillance video analysis system. *Expert Systems with Applications*, 89:343–361, 2017. 1

[16] Jonathan S Lee, Sukjae Choi, and Ohbyung Kwon. Identifying multiuser activity with overlapping acoustic data for mobile decision making in smart home environments. *Expert systems with applications*, 81:299–308, 2017. 1

[17] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8561–8568, 2019. 2

[18] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Unsupervised learning of view-invariant action representations. *arXiv preprint arXiv:1809.01844*, 2018. 2

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[20] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 2

[21] Alessandro Ortis, Giovanni M Farinella, Valeria D'Amico, Luca Addesso, Giovanni Torrisi, and Sebastiano Battiato. Organizing egocentric videos of daily living activities. *Pattern Recognition*, 72:207–218, 2017. 1

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 5

[23] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 1

[24] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12056–12065, 2019. 1, 2

[25] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443, 2016. 2

[26] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 1, 2

[27] Panos E Trahanias and Anastasios N Venetsanopoulos. Color image enhancement through 3-d histogram equalization. In *11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis,*, volume 1, pages 545–548. IEEE Computer Society, 1992. 3

[28] 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. 2

[29] 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. 1, 2, 5

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[31] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. 1, 2, 4

[32] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[33] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. *arXiv preprint arXiv:2006.03876*, 2020. 1, 5

[34] Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. Carefi: Sedentary behavior monitoring system via commodity wifi infrastructures. *IEEE Transactions on Vehicular Technology*, 67(8):7620–7629, 2018. 1

[35] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *International Conference on Computer Analysis of Images and Patterns*, pages 36–46. Springer, 2017. 3

[36] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1632–1640, 2019. 3