Action Recognition in Dark Videos using Spatio-temporal Features and Bidirectional Encoder Representations from Transformers

Himanshu Singh, Saurabh Suman, Badri Narayan Subudhi, Vinit Jakhetiya, and Ashish Ghosh

***Abstract*—Several research works have been developed in the area of action recognition. Unfortunately, when these algorithms are applied to low-light or dark videos, their performances are highly affected and found to be very poor or fall rapidly. To address the issue of improving the performance of action recognition in dark or low-light videos; in this article, we have developed an efficient deep 3D CNN based action recognition model. The proposed algorithm follows two-stages for action recognition. In the first stage, the low-light videos are enhanced using Zero-Reference Deep Curve Estimation (Zero-DCE), followed by the min-max sampling algorithm. In the latter stage, we propose an action classification network to recognize the actions in the enhanced videos. In the proposed action classification network, we explored the capabilities of the $R(2+1)D$ for spatio-temporal feature extraction. The model's overall generalization performance depends on how well it can capture long-range temporal structure in videos, which is essential for action recognition. So we have used a Graph convolutional network (GCN) on the top of R(2+1)D as our video feature encoder which captures long-term temporal dependencies of the extracted features. Finally, a Bidirectional Encoder Representations from Transformers (BERT) is adhered to classify the actions from the 3D features extracted from the enhanced video scenes. The effectiveness of the proposed action recognition scheme is verified on ARID V1.0 and ARID V1.5 datasets. It is observed that the proposed algorithm is able to achieve 96.60% and 99.88% as Top-1 and Top-5 accuracy, respectively, on ARID V1.0 dataset. Similarly, on ARID V1.5, the proposed algorithm is able to achieve 86.93% and 99.35% as Top-1 and Top-5 accuracies, respectively. To corroborate our findings, we have compared the results obtained by the proposed scheme with those of fifteen state-of-the-art action recognition techniques.**

***Impact Statement*—Recognizing human actions in dark-light conditions has many real-time applications like night-smart surveillance systems, elderly people, monitoring in smart homes, military applications, self-driving cars, etc. The current state-of-the-art techniques cannot effectively perform human action recognition in dark-light situations. In the proposed work, we have constructed a novel deep learning architecture that involves an image enhancement module followed by an action classification network, to classify the actions in dark/low-light videos. Our approach surpasses the available state-of-the-art techniques for action recognition in the dark and provides 96.60% Top 1 accuracy.**

***Index Terms*—Action recognition, Image processing, Dark video**

Himanshu Singh, Saurabh Suman, Badri Narayan Subudhi, and Vinit Jakhetiya are with Indian Institute of Technology Jammu, NH44, Nagrota, Jammu-180019, INDIA

Ashish Ghosh is with Indian Statistical Institute Kolkata, 203 B. T. Road, Kolkata-700108, INDIA

I. Introduction

ACTION recognition is one of the most crucial tasks in computer vision. It is widely used in several real-life applications including: intelligent surveillance [1], sports video analysis [2], human-computer interaction [3], and human action analysis [4], elderly people monitoring in smart homes [5], etc. Researchers have reported several pioneering people in the state-of-the-art (SOTA) techniques for action recognition. In conventional practice, [6], [7], [8], [9], [10] techniques focus on action recognition in high-quality daylight video streams or the presence of an adequately illuminated environment rather than in an unfavorable illumination environment or during nighttime. However, many real-world computer vision applications include operating in low-contrast and dim-lighting environments or an environment with poor illumination, such as night security surveillance systems, self-driving cars at night, military applications, etc. Action recognition from such low-light/dark videos is quite challenging to perform with reasonable accuracy.

It is very arduous to recognize action in dark/low-light videos. There is a dearth of studies on action recognition in low-light videos as most of the existing SOTA techniques cannot improve the accuracy of action recognition or are inefficient for dark video data. Though there has been an increase in research interest for action recognition, in dark environments, most of the traditional techniques [11], [12], focused on improving the visibility of dark videos. However, most of the existing SOTA techniques fail due to the poor data augmentation techniques which unexpectedly destroy the data and hence lead to a decrease in classification accuracy. Further, developing a high-efficiency algorithm for action recognition is challenging as it needs to enhance the video scene initially and further classify the same. In real-life instances a large number of videos on which action recognition tasks to be achieved are untrimmed and of various lengths. The required action takes place in only a small part of the video. Cropping or segmenting the more extensive input videos before sending them to the network may cut out the action. As a result, the best technique is to feed the complete video to the network. However, this is not trivial because of device memory restrictions to various video lengths.

It may be noted that the literature on action recognition for dark or low-light video scenes is scarce. Further, the CNN-based state-of-the-art techniques are unable to characterize the spatio-temporal features of the video scene, as unable to

extract the meaningful information from the obscured dark videos. Hence, for a dark or low-light video, the choice of a proper enhancement technique is quite essential. Further, most of the existing graph neural networks (GNNs) typically capture spatial dependencies with the predefined or learnable static graph structure, ignoring the hidden dynamic patterns. Meanwhile, most recurrent neural networks (RNNs) or convolutional neural networks (CNNs) cannot effectively capture temporal correlations, especially for long-term temporal dependencies [13], [14].

With that motivation, We proposed an action recognition technique for dark or low-light videos which follows two stages: image enhancement module (IEM) and action classification network (ACN). In the said architecture, we have used Zero-Reference Deep Curve Estimation (Zero-DCE) [15] followed by the min-max sampling techniques to enhance the dark videos and bring out the inherent details of the video. The Zero-DCE component of IEM adapts nicely to different levels of light conditions. Further, we proposed a new action classification network architecture which is a combination of the $R(2+1)D$ [16] followed by graph convolutional network (GCN) [17] succeeded by Bidirectional Encoder Representations from Transformers (BERT) [18]. Here we proposed GCN as a temporal feature encoder of the features obtained by $R(2+1)D$. The GCN utilizes features obtained by $R(2+1)D$ to model intrinsic temporal relations for providing a robust encoded representation for action recognition. The use of Graph convolutional networks (GCN) on the top of $R(2+1)D$ as our video feature encoder captures the dependencies among the spatial and long term temporal extracted features. We evaluate the proposed scheme on dark light video datasets "ARID V1.0" and "ARID V1.5". We have used two evaluation measures, Top-1 and Top-5 accuracy, to show the effectiveness of the proposed scheme. The evaluation of the proposed scheme is verified using fifteen SOTA techniques and found to be corroborating our findings.

The organization of the remaining portion of this article is as follows. Section 2 gives a brief description of the state-of-the-art techniques. A detailed description of the proposed works is provided in Section 3. Section 4 represents experimental results with quantitative evaluations of the same. Conclusions are drawn in Section 5.

II. State-of-the-art Techniques

Several algorithms for human action recognition have been developed during the last few decades, varying in terms of technology/algorithm utilized, enhanced efficiency, and easy implementation. Although action classification for low light video can be challenging, it has several potential applications, including military deployments, nighttime surveillance, intelligent navigation systems, elederly people monitoring, etc.

Recently, with the use of the deep convolutional models, researchers have been able to develop several reliable, low cost and robust action recognition algorithms. Simonyan and Zisserman [19] proposed a two-stream ConvNet architecture consisting of spatial and temporal networks for action recognition. The authors have utilized the multi-frame dense optical flow features for training the ConvNet architecture with a lesser amount of training data. The deep convolutional networks have earned remarkable success for visual recognition in still images. However, the advantages of such methods over traditional methods are not evident for action recognition. In this regard, Wang *et al.* [20] has proposed a temporal segment network (TSN) based on the idea of long-range temporal structure modeling. TSN combines a sparse temporal sampling strategy and video-level supervision, which enables efficient and effective learning using the whole action video. Carreira and Zisserman [21] have proposed a two-stream Inflated 3D ConvNet (I3D) for action recognition. The method starts with a two-dimensional architecture and inflates all filters and pooling kernels. The inflating provides a new dimension to 2D architecture. Filters in 2D models are square, but by inflating them, they become cubic. The 3D architecture is able to extract the spatio-temporal features from the input video while leveraging successful ImageNet architecture. Further, in state-of-the-art techniques, the use of a 3D CNN model is reported for the action recognition task. The 3D-convolutional neural network utilizes 3D convolutions. Using the 3D CNN models, the low-level feature representations are calculated with the help of a three-dimensional filter. Tran *et al.* [22] reported the first use of 3D-CNN model named as C3D for action recognition. More advanced and broad 3D-CNN networks, such as 3D-Resnet [23] and 3D-Resnext [24], also demonstrated satisfactory results. The use of $R(2+1)D$ [16] has been presented for spatial and temporal convolutions to increase features characteristic of action recognition. A video contains different modalities like RGB information, optical flow information, skeleton information, and depth information. Many of the research works used these diverse modalities with the help of two or multi-streams action recognition models. The two-stream action recognition modules relied on the parallel feature extraction, making them more generic. Optical flow information is complementary to RGB information, and it is the most commonly used two-stream methods.

Recently, the Self-attention Mechanism has been widely used for action recognition. The deep learning network is constructed using the standard transformer architecture, with each block customized for temporal and spatial attention individually. Some widely used self-attention-based models are Video Swin Transformer [25], Video-Vision Transformer (ViViT) [26], TimeSformer [8], etc. For improved temporal information, BERT is used to replace the traditional temporal global average pooling layer in the 3D-CNN model. However, the said approaches are unable or yet to be explored to perform in low-light or dark videos. Action recognition in dark light/ night-time videos is comparatively less explored or yet to be explored. Action recognition of dark video data is a challenging task, However, putting efficient models for dark video action recognition into practice would be highly beneficial.

Xu *et al.* [27] have developed a dataset for action recognition in dark (ARID) to work on action recognition in dark/low-light videos. Patel *et al.* [28] introduces a new low-light image processing pipeline that uses ResNets, and a statistical tool to detect human actions in low-light images, paving the way for

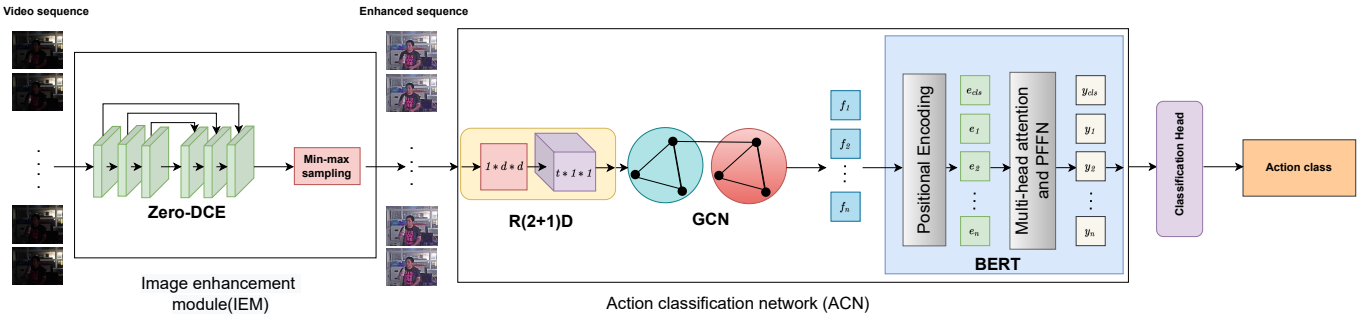


Fig. 1: Framework of the proposed method. The raw frames are obtained first from the given low-light videos. The Zero-DCE approach is then used to enhance these frames. The frames are then sampled using Min-max sampling and passed to feature extractor $R(2+1)D$. These extracted features are then passed to GCN. These obtained features are then fed to BERT, followed by a basic linear layer to get the model's final classification.

learning-based pipelines for human action recognition in dark videos. Dealing with dark video data is a quite challenging task. Further, the distribution of video duration varies by different actions. Learning the action itself may lead to over-fitting, creating a bias towards the frame number at which the most significant changes occur. To solve this problem, Hira *et al.* [29] developed a straightforward but effective delta-sampling-based approach that combines the effective use of ResNet and BERT approaches and outperforms other learning-based benchmarks such as slow-fast networks and temporal attention, as well as other sampling algorithms. Chen *et al.* [30] proposed DarkLight Networks for action recognition in dark/low-light videos. It consists of a dual-pathway structure that employs dark and bright videos for effective video representation. It also uses a self-attention mechanism that fuses and extracts corresponding and complementary information from the two pathways.

III. Proposed Methodology

It may be observed that many of the SOTA techniques employ CNN architectures for action recognition in dark videos. However, existing CNN approaches cannot characterize the spatio-temporal information in low light or dark videos for action recognition. Hence in this work, we seek to explore the capabilities of the $R(2+1)D$, Graph Convolutional Network (GCN), and Bidirectional Encoder Representations from Transformers (BERT) architecture to recognize the actions from the dark or low light videos. This article proposes a unique action recognition network for dark videos, consisting of an image enhancement module (IEM) with an action classification network (ACN), which is a combination of $R(2+1)D$ and GCN followed by BERT.

The framework of our suggested technique is illustrated in Fig. 1. It takes the entire video as input, *i.e.*, $V = \{V_1, ..., V_L\}$, $V \in R^{h \times w}$, where $h \times w$ specifies the spatial size of each frame and L denotes the video length, i.e., the number of frames in the dark video. We employ the IEM module to enhance the dark video frames. The feature for m-frames snippet of V is then extracted using a feature extractor (e.g., the $R(2+1)D$ network). We have taken $m = 64$ into account.

The output of the spatial $R(2+1)D$ branch is a $512 \times 8 \times 7 \times 7$ feature map, which is then average pooled in the spatial domain to reduce it to a dimension of $512 \times 8 \times 1 \times 1$. GCN is mainly used to capture the spatial and long term temporal dependencies among the extracted features from the $R(2+1)D$. The extracted features are fed to GCN to obtain the enhanced feature of dimension 256×8. We further feed the obtained features into BERT followed by a basic linear layer to get the model's final classification result.

A. Image enhancement module (IEM)

In the proposed scheme, the image enhancement module (IEM) is used to increase the perceivable information from dark videos. In the proposed work, we adhered to the use of the Zero-Reference Deep Curve Estimation (Zero-DCE) technique, followed by the min-max sampling strategy to enhance the dark videos. Fig 2 depicts the image enhanced by Zero-DCE.

*1) **Zero-DCE**:* Zero-DCE is a lightweight deep network model. Zero-DCE is a method for estimating a set of best-fitting Light-Enhancement curves from an input image. It generates the enhanced image by applying these curves iteratively by mapping all the pixels of the input image. The key components of Zero-DCE are the light-enhancement curve, DCE-Net, and non-reference loss functions.

Light-enhancement curve (LE-curve): A light-enhancement (LE) curve is a type of curve that can map a low-light image to its enhanced version. The pixel values of the input image are exclusively responsible for the self-adaptive curve parameters. The LE curve accomplishes the following goals:

- It places the improved image's pixel values in a normalized range of [0,1] to minimize information loss due to overflow truncation.
- The nature of the LE curve is monotonous. It also preserves the differences between neighboring pixels.

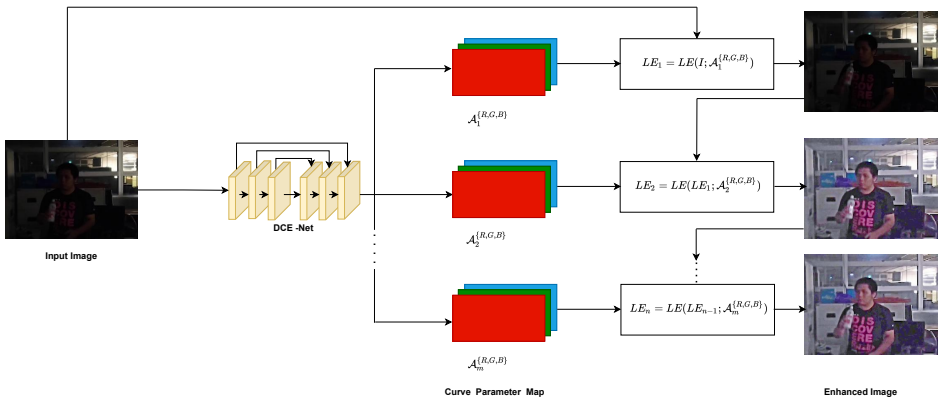


Fig. 2: Graphical illustration of Zero-DCE for low-light image enhancement. A DCE-Net is used to estimate a series of best-fitting Light-Enhancement curves (LE-curves) that are used to enhance a given input image iteratively. Here m is the number of iterations.

- In gradient backpropagation is a basic curve that can be differentiable.

DCE-Net: To learn the mapping between an input picture and its best-fitting curve parameter mappings, Zero-DCE presented a Deep Curve Estimation Network (DCE-Net). The DCE-Net takes a low-light image as input and returns a set of pixel-wise curve parameter mappings for higher-order curves as output. DCE-Net utilizes a Convolutional Neural Network (CNN) of seven convolutional layers. Each layer has the 32 convolutional kernels of size 3×3 and a stride of 1 followed by the ReLU function. It does not include the downsampling and batch normalization layers, which disrupt nearby pixel relationships. For a 256×256 input image with 3 channels, the DCE-Net has just $79,416$ trainable parameters and $5.21G$ Flops. As a result, it is a lightweight model and may be utilized in devices with low processing resources, such as mobile platforms.

Non-reference loss functions:

The Zero-DCE Model is trained using a weighted, linear combination of the following four non-reference loss functions:

a) Spatial consistency loss (L_{spa}): The enhanced result can inherit the spatial consistency from the given input image. In other words, the bright (dark) regions in the input image should keep relatively bright (dark) in the enhanced result. Otherwise, the result would have relatively low contrast. By retaining the difference between contiguous regions between the input image and its enhanced version, the spatial consistency loss aids the spatial coherence of the enhanced image. The spatial consistency loss L_{spa} can be represented mathematically as;

$$L_{spa} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\Omega(i)}(|(Y_i - Y_j)| - |(I_i - I_j)|)^2, \quad (1)$$

where K represents the number of local regions, and $\Omega(i)$ is a symbol for the four neighboring regions (top, down, left, and right) centered on the location i. Y_i and Y_j are the average intensity value of the i^{th} and j^{th} local region in the enhanced image and I_i and I_j is the average intensity value of the i^{th} and the j^{th} local region in the input image, respectively.

b) Exposure control loss (L_{exp}): Exposure control loss is inspired by the measurement of well-exposedness in multi-exposure fusion [31], which stated that the fused result should be near a fixed exposure level. It controls the exposure level to prevent under -/over-exposed areas. This loss measures the distance between the average intensity value of a local region and the well-exposedness level E. The exposure control loss can be expressed as,

$$L_{exp} = \frac{1}{M}\sum\nolimits_{k=1}^{M}|Y_k - E|, \quad (2)$$

where M indicates the number of non-overlapping local regions of size 16×16 and Y_k denotes the average intensity value of the k^{th} local region in the enhanced image.

c) Color constancy loss (L_{col}): The design of color constancy loss, which corrects potential color deviations in the enhanced image, is based on the gray-world color constancy hypothesis [32] that color in each sensor channel averages to grey over the entire image. The color constancy loss L_{col} can be given as,

$$L_{col} = \sum\nolimits_{\forall(p,q)\in\varepsilon}(J^p - J^q)^2, \varepsilon = \{(R,G),(R,B),(G,B)\}, \quad (3)$$

where J^p represents the average intensity value of the enhanced image's p^{th} channel, and (p, q) are a pair of channels.

d) Illumination smoothness loss ($L_{tv_\mathcal{A}}$): An illumination smoothness loss is applied to each curve parameter map $\mathcal{A}$ to preserve the monotonicity relations between neighboring pixels. The illumination smoothness loss $L_{tv_\mathcal{A}}$ can be expressed as:

$$L_{tv_\mathcal{A}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\xi}(|\nabla_x\mathcal{A}_n^c| + \nabla_y\mathcal{A}_n^c|)^2, \xi = \{R,G,B\}, \quad (4)$$

where N denotes the number of iterations and ∇_x is the horizontal gradient operations, and ∇_y represents the vertical gradient operations.

e) Total loss (L_{total}) : The total loss can be computed as:

$$L_{total} = L_{spa} + L_{exp} + W_{col}L_{col} + W_{tv_{\mathcal{A}}}L_{tv_{\mathcal{A}}}, \quad (5)$$

where W_{col} and $W_{tv_{\mathcal{A}}}$ are the weights of the losses.

*2) **Min-max sampling**:* A typical video consists of action preparation, the start, and end of the action. The distribution of video length is variable for different actions. It introduces a bias toward the frame number where the most significant variance occurs overfits the model. We suggest a simple yet effective Min-max sampling approach to overcome this problem. Let us consider an input video $V = (V_1, ..., V_L)$, V_L represents the L^{th} frame and L represents the total number of frames in the original video. Let L_r be the required number of frames to be given as an input, then the maximum step size for the sampling S_{max} is considered to be,

$$S_{max} = \lfloor \frac{L}{L_r} \rfloor. \quad (6)$$

Hence, the required step size of the sampling S_{req} is obtained as,

$$S_{req} = Max[Min[S_{max}, F_1], F_2], \quad (7)$$

where F_1 and F_2 represent numbers of frames. We have considered the values of F_1 and F_2 are 4 and 1, respectively.

B. *Proposed Action Classification Network (ACN)*

The proposed action recognition model consists of a feature extractor $R(2+1)D$ with a Graph Convolutional Networks (GCN) followed by Bidirectional Encoder Representations from Transformers (BERT). Here the $R(2+1)D$ is used for extracting the clip-level features from a given video. Then, the GCN is used to extract the inherent spatio-temporal relations of the obtained features. Further, the BERT is used to provide a robust encoded network representation for action recognition.

It is a crucial task to capture the spatio-temporal information in low-light videos for action recognition. The model's overall generalization performance get affected by this. To overcome this issue, we adhered to GCN on the top of $R(2+1)D$ followed by BERT. For example, for a video, V with n frames, the feature extractor $R(2+1)D$ maps the clips into the appropriate sequence of features, which does not extensively include the video's complex temporal structure. Therefore, we employ the GCN, which builds a fully connected graph with learnable edge weights on top of the clip-level information using a parameterized adjacency matrix. We apply a graph convolutional network with two layers to these graph representations and then employ BERT on top of it with a basic linear layer to get the model's final classification result.

1) $R(2+1)D$: ResNet [33] has excelled in the field of 2D convolutional neural networks. However, the action recognition task involves video data. The 3D-CNN model has a better performance with good false-positive reduction compared with its 2D counterparts. $R(2+1)D$ architecture is a ResNet-style architecture that explicitly factors 3D convolution into two independent and sequential operations: 2D spatial convolution and 1D temporal convolution. Let us assume V represents an input video of $L \times h \times w \times 3$, where L is the number of frames in the video, h and w denote the frame height and width, and 3 denotes the RGB channels. Let z_i be the tensor produced by the residual network's i^{th} convolutional block. In 3D CNNs, temporal information is propagated via the network layers. The tensor z_i is a 4D of size $N_i \times L \times h_i \times w_i$, here N_i denotes the number of filters used in the i-th block. Every filter has 4-dimensions and has the size $N_{i-1} \times t \times d \times d$, where d denotes the spatial width and height, and t represents the temporal extent of the filter. $R(2+1)D$, replaces the N_i 3D convolutional filters of size $N_{i-1} \times t \times d \times d$ with a $R(2+1)D$ block consisting of M_i 2D convolutional filters of size $N_{i-1} \times 1 \times d \times d$ and N_i temporal convolutional filters of size $M_i \times t \times 1 \times 1$. By increasing the number of channels, the hyperparameter M_i makes the $R(2+1)D$ have the same number of parameters as the complete 3D convolution. This decomposition adds a nonlinear operation between the 2D and 1D convolutions, which increases the non-linearity by two times. Factoring 3D convolution into spatial and temporal components isolates the optimization process and makes it more accessible. Compared to 3D convolutional networks of the same capacity, this results in reduced training error.

*2) **Graph convolutional networks (GCN)**:* Graph convolutional networks (GCN), is a multi-layer convolutional neural networks, invented by Kipf and Welling [17]. The kernel operation in GCN is similar to CNN. GCN operates in the principle of the fusion of information between node neighborhoods. In GCN, the application of convolution operation to the Laplacian matrix of a graph establishes a connection between the spatial and temporal domains. The convolution operation in GCN is a kind of Laplacian smoothing. Hence in GCN, several layers are used repeatedly for over smoothing. In the proposed scheme, the long dependencies between the spatial and the temporal domain features extracted by $R(2+1)D$ schemes are achieved by the GCN network.

GCN learns each node's embeddings by iteratively gathering data from its neighbors. Given a network $G = (V, E)$, let define the node feature matrix $X \in R^{n \times d}$, where $n = |V|$ represents the number of nodes, d is the dimension of features, and E is the number of edges. Let $A \in R^{n \times n}$ be the network's adjacency matrix and D the associated node degree matrix where $D_{ii} = \sum_j A_{ij}$. Let's assume that each node is linked to itself, i.e., $\tilde{A} = A + I$ (where I represents the identity matrix). Then, by using an effective re-normalization approach, such as, $\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ (where $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$), the typical two-layer GCN may be defined as:

$$Z = f(X, A) = \mathrm{softmax}(\tilde{A}\,\mathrm{ReLU}(\tilde{A}XW^{(0)})W^{(1)}), \quad (8)$$

where $W^{(0)}$ and $W^{(1)}$ represent the weight parameter, ReLU (and softmax) the non-linear activation function, and Z the final output for the assignment of node labels.

*3) **Graph Convolutional Network as a Temporal Graph Encoder**:* We use a similarity graph to express a video in

our approach, as in [14] since action recognition depends on the ability to capture long-range temporal structure in videos. Given an input video $\mathbf{V}_n = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$ with n clips, the corresponding clip-level feature vector representations as $\mathbf{f}_n = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$, extracted by $R(2+1)D$, with d dimensions. By taking into account the pairwise affinity between two feature vectors, we create a fully-connected network X with n nodes from $\mathbf{f}$ as:

$$F(\mathbf{f}_i, \mathbf{f}_j) = \phi(\mathbf{f}_i)^\top \phi'(\mathbf{f}_j) \tag{9}$$

where, $\phi(.)$ and $\phi'(.)$ are defined as $\phi(\mathbf{f}) = \mathbf{w}\mathbf{f}$ and $\phi'(\mathbf{f}) = \mathbf{w}'\mathbf{f}$ and represent two different transformation functions of the original feature vectors. The weights $\mathbf{w}$ and $\mathbf{w}'$ of dimension $d \times d$ are used to parameterize the transform. In order to fully utilize the video's rich temporal information, such modifications help in learning the long-range correlations between the feature vectors. By using eq. 9, we compute the affinity for each possible pair to produce a similarity matrix $A^{similar}$ of size $n \times n$. Then we apply a softmax function to normalize the matrix as follows:

$$A_{ij}^{similar} = \frac{\exp(F(\mathbf{f}_i, \mathbf{f}_j))}{\sum_{j=1}^{n} \exp(F(\mathbf{f}_i, \mathbf{f}_j))} \tag{10}$$

The learnable weights $\mathbf{w}$ and $\mathbf{w}'$ enable us to learn the edge weights between the nodes by back-propagation using the normalized matrix $A^{similar}$ as the adjacency matrix for the similarity graph.

*4) **Bidirectional Encoder Representations from Transformers (BERT)**:* BERT is a bidirectional self-attention technique found to be demonstrating outstanding results in a variety of downstream natural language processing (NLP) applications. For sequential data, the bidirectional feature allows BERT to combine contextual information from both directions rather than depending on a single direction. BERT also offers complicated, unsupervised pre-training problems, which result in useful representations for a variety of applications. A self-attention module in BERT architecture computes the response at a position in a sequence by attending to all positions and calculating their weighted average in an embedding space. Inspired by the work of Kalfaoglu *et al.* [10], which investigates eliminating temporal global average pooling with BERT, we use the self-attention blocks, as illustrated in Fig 3, to choose more useful spatio-temporal features for action recognition in dark videos. To put them into the self-attention mechanism, we pass feature $f = f_1, f_2, f_3, ..., f_n$ of size $D \times n$, where n is the dimensionality of temporal and D is the number of short-term characteristics of consecutive frames. To encode each feature, the input features are first added to a learnable positional embedding by,

$$e_i = f_i + e_{pos}^i \quad (i = 1, 2, ..., n), \tag{11}$$

where e_i indicates the i^{th} encoding vector, it provides location information. e_{pos}^i denotes i^{th} learnable positional embedding where ${e_{pos}}^i \in R^{D \times n}$. A learnable CLS token is also considered in the encoding as e_0, $e \in R^{D \times (n+1)}$. In the self-attention, there are L number of encoding blocks, and for

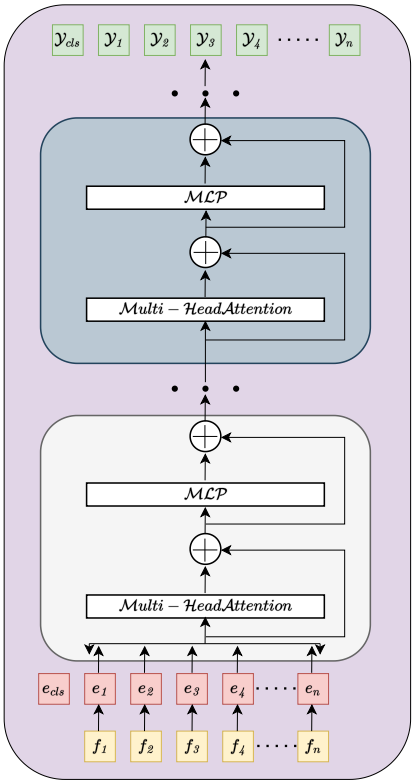


Fig. 3: Graphical illustration of self-attention mechanism

every block l, a query/key/value vector is constructed from the output of the previous block, as shown below

$$Q_i^{(l,h)} = W_q^{(l,h)} \mathcal{L}\left(e_i^{l-1}\right) \quad \in R^d. \tag{10}$$

$$K_i^{(l,h)} = W_k^{(l,h)} \mathcal{L}\left(e_i^{l-1}\right) \quad \in R^d. \tag{11}$$

$$V_i^{(l,h)} = W_v^{(l,h)} \mathcal{L}\left(e_i^{l-1}\right) \quad \in R^d. \tag{12}$$

Here $Q_i^{(l,h)}$, $K_i^{(l,h)}$ and $V_i^{(l,h)}$ represents the i^{th} query, key and value vector respectively, the $\mathcal{L}$ denotes the LayerNorm, $h = 0, 1, 2, 3, 4...H$, represent the index of multiple attention heads. $W_q^{(l,h)}$ $W_k^{(l,h)}$ $W_v^{(l,h)}$ denotes the weight matrices and D repersent the latent dimensionality and evaluated as $d = D/H$. Dot-product is used to determine self-attention weights as follows:

$$\alpha^{(l,h)} = softmax\left(\frac{q_i^{(l,h)T}}{\sqrt{d}} k_i^{(l,h)}\right) (i = 0, \ldots, m). \tag{13}$$

The weighted sum of value vectors using α from each attention head is used to obtain encoding e_i^l at block l, as shown below:

$$s_i^{(l,h)} = \sum_{0}^{m} \alpha_i^{(l,h)} v_i^{(l,h)}, \tag{14}$$

$$e_i'^{(l)} = W_o \begin{bmatrix} s_i^{(l,1)} \\ \vdots \\ s_i^{(l,H)} \end{bmatrix} + e_i^{(l-1)}, \tag{15}$$

$$e_i^{(l)} = MLP\left(\mathcal{L}\left(e_i'^{(l)}\right)\right) + e_i'^{(l)}. \tag{16}$$

The classification token $\mathcal{Y}_{cls}$ acquired from the last block is processed by an FC layer and the argmax function to deliver the final prediction result, as shown below:

$$Result = Argmax\left(FC\left(\mathcal{Y}_{cls}\right)\right). \tag{17}$$

The proposed technique uses BERT-based temporal pooling. The primary benefit of BERT is that it learns the most efficient subspace in which the attention mechanism operates, as well as the classification embedding, which learns how to appropriately attend to the spatio-temporal features of the $R(2+1)D$ and GCN architecture.

IV. Experimental Results

The proposed scheme is implemented on a $Core$ i7 system with 128 GB RAM and 32 GB GPU using the open-source machine learning framework PyTorch [34]. The effectiveness of the proposed algorithm is tested on the ARID database. The efficiency of the proposed scheme is validated quantitatively by Top-1 and Top-5 accuracy. The achievement of the proposed technique is evaluated by comparing the results obtained by it against fifteen existing state-of-the-art action recognition techniques.

Fig. 4: Total action classes in ARID dataset

A. ARID dataset

The availability of the dark dataset in real-life is very scarce. In this work, we have devised our strategies on the ARID [27] dataset to prove its efficiency. The ARID dataset has two versions: ARID V1.0 and ARID V1.5. All the videos of the ARID dataset are taken in low light-condition or at night time. ARID V1.0 consists of 3,784 video clips, whereas ARID V1.5 has 6,207 dark videos. Both versions of the ARID dataset have eleven action classes containing videos of lower illumination and dark environment conditions. The dataset comprises eleven action classes: drinking, jumping, picking, pouring, pushing, running, sitting, standing, turning, walking, and waving. Given the dark/low light video, the proposed scheme aims to classify them into eleven classes. The training, testing, and validation split contains 3792, 1768, and 647 videos, respectively, ranging from 33 to 255 frames. Fig 4 shows the different classes inside the ARID dataset. The size of the ARID dataset is minimal. This dataset differs significantly from other freely accessible video datasets such as UCF101, Kinetics [35], and Charades [36] etc. Since most of the videos in ARID are luminously extremely dark. There are several challenges while dealing with the ARID dataset. The main challenge of the ARID dataset is that it is very small compared to other datasets. A further challenge is the scarcity of videos that are notably different. Approximately 95% of the dataset consists of videos with the same subject(s) performing similar actions.

B. Implementation details

We have conducted experiments on both versions of ARID benchmark datasets for action recognition in the dark, i.e., ARID V1.0 and ARID V1.5. We report the Top-1 and Top-5 accuracies on ARID V1.0 and ARID V1.5. The input frames size is considered to be $3 \times 64 \times 112 \times 112$. The input frames have been enhanced by Zero-DCE. As for the feature extractor, we have utilized $R(2+1)D$ -34 without the average temporal pooling at the end, which was pre-trained on the $IG65M$ [6] dataset. $R(2+1)D$ -34 decomposes 3D convolution into 2D spatial convolution and 1D temporal convolution. Output from the feature extractor has the dimension of $512 \times 8 \times 7 \times 7$. After that, an average pooling layer is applied which provides the output of dimension 512×8. After that we have transposed it to a dimension of the size 8×512 which is the input to the GCN (Temporal Graph Encoder). We have used two layer GCN that provides the output of dimension 8×256. This is further fed to the BERT that provides the feature vector of dimension 9×256. Which is then reduced to the dimension of 256 and given to the classification head to classify the action. We have used eight attention heads and one transformer block in the BERT architecture. The ADAMW [37] optimizer with a learning rate of 10^{-5} is utilized for training.

C. Quantitative Evaluation and Discussions

We have conducted experimentation of the proposed technique on the V1.0 and V1.5 versions of the ARID database. The effectiveness of the proposed approach is tested by comparing it against fifteen SOTA techniques: VGG-TS [19], TSN [20], I3D-TS [21], C3D [22], Separable-3D [38], 3D-ShuffleNet [39], 3D-SqueezeNet [40], 3D-ResNet-18 [16], I3D-RGB [21], 3D-ResNet-50 [24], 3D-ResNet-101 [24], Pseudo-3D-199 [41], 3D-ResNext-101 [24], DarkLight-ResNeXt-101 [30] and DarkLight-$R(2+1)D$-34 [30].

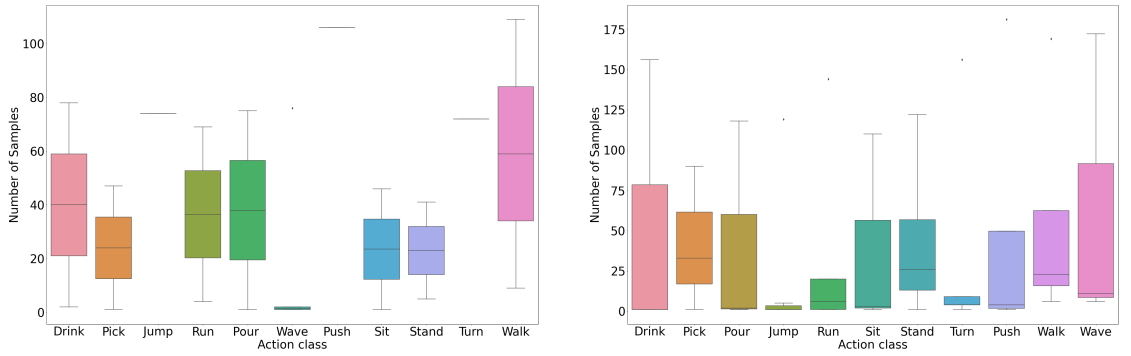
In the proposed scheme, we have utilized Zero-DCE with min-max sampling for visual enhancement of all the videos of the ARID database. Fig. 5 represents the dark image frames and corresponding visually enhanced results. The Zero-DCE scheme has brought out much intended visual information of the dark videos. This can be clearly observed in Fig. 5. Further, these frames are given as the input to the $R(2+1)D$. The GCN is used as the temporal feature encoder to utilize the long-range relationships among the features obtained from $R(2+1)D$ and are then fed to BERT, followed by a linear layer to obtain an action classification result.

We have used two metrics: Top-1 accuracy and Top-5 accuracy, for evaluating our model performance quantitatively. Top-1 accuracy measures the proportion of examples for which the predicted label matches the single target label. In contrast, Top-5 accuracy considers a classification correct if any of the five predictions match the target label. The quantitative evaluation and comparison of the proposed technique on ARID V1.0 with those of the considered fifteen SOTA techniques are provided in Table I. It can be seen that the suggested approach outperforms fifteen other SOTA techniques. A similar analysis of results on ARID V1.5 is shown in Table II.

Dark light frames of ARID dataset

Visual Enhancement results of Zero-DCE on all action classes of ARID dataset

Fig. 5: Visual illustration of Zero-DCE based image enhancement on all classes of ARID dataset



Box-Plot of our proposed method's action classification on the ARID V1.0 dataset

Box-Plot of our proposed method's action classification on the ARID V1.5 dataset

Fig. 6: Confusion matrix representation using box plot on ARID database

TABLE I: The Top-1 and Top-5 accuracy results on ARID V1.0 of a few competitive models and ours.

Models	Top-1 Accuracy	Top-5 Accuracy
VGG-TS	32.08%	90.76%
TSN	57.96%	94.17%
C3D	40.34%	94.17%
Separable-3D	42.16%	93.44%
3D-ShuffleNet	44.35%	93.44%
3D-SqueezeNet	50.18%	94.17%
3D-ResNet-18	54.68%	96.60%
I3D-RGB	68.29%	97.69%
3D-ResNet-50	71.08%	99.39%
3D-ResNet-101	71.57%	99.03%
Pseudo-3D-199	71.93%	98.66%
I3D Two-stream	72.78%	99.39%
3D-ResNext-101	74.73%	98.54%
DarkLight-ResNeXt-101	87.27%	99.47%
DarkLight-$R(2+1)D$-34	94.04%	99.87%
$R(2+1)D$-GCN+BERT(Proposed)	**96.60%**	**99.88%**

TABLE II: The Top-1 and Top-5 accuracy results on ARID V1.5 of a few competitive models and our proposed method.

Models	Top-1 Accuracy	Top-5 Accuracy
3D-ResNet-18	31.16%	90.49%
I3D-RGB	48.75%	90.611%
I3D Two-stream	51.24%	90.95%
Darklight+$R(2+1)D$-34	84.13%	97.34%
$R(2+1)D$-GCN+BERT(Proposed)	**86.93** %	**99.35%**

Due to the limited number of classes in the ARID dataset, the Top-5 accuracy is reasonably good in all SOTA approaches. Our proposed technique achieves the best results on both versions of the ARID benchmark dataset. The proposed work uses all eleven classes of ARID datasets for the evaluation of Top-1 and Top-5 accuracy. Table I demonstrates how well the various CNN-based feature extractors perform. In terms of Top-1 accuracy, our suggested method is 9.33% better than DarkLight-ResNeXt-101. In terms of Top-1 accuracy, the proposed ACN surpasses the I3D-Two-stream network, which employs both RGB and flow features as input, by 23.82%. This demonstrates that the ACN is not just potent but that the optical flow may be ineffective for action recognition in the dark. Meanwhile, when compared to 3D-ResNet-18 and 3D-ResNet-101, we find that the deeper the network layers, the greater the effect, but the performance of the suggested $R(2+1)D$-GCN and BERT design with 34 layers is 21.87% better than the 101 layers in 3D-ResNet-101.

Table II shows the proposed ACN performance on the benchmark dataset ARID V1.5. The proposed Action classification network achieves 86.93% & 99.35% Top-1 and Top-5 accuracies, respectively. It can be observed that the proposed ACN network surpassed the I3D-Two-stream network by 35.69% in Top-1 accuracy. It leaves behind the 3D-ResNet-18 by 55.77% in terms of Top 1 accuracy. The confusion matrix score obtained by the proposed action recognition technique on ARID V1.0 and ARID V1.5 datasets are presented in Fig 6. For ARID V1.5, it is observed that for a few action classes the proposed techniques are getting confused, and hence the box plot indicates the misclassification whereas a better result is obtained with less misclassification in ARID V1.0. It is happening because in ARID V1.5 number of similar-action videos is more than in ARID V1.0. It may be observed that the proposed scheme provides better results than fifteen SOTA techniques on ARID V1.0 dataset and comparative results with the Darklight-$R(2+1)D$-34. The comparisons illustrate that the proposed ACN is far better than many other excellent

models based on 3D-CNN or two-stream in Top-1 accuracy.

D. Ablation Study

In this section, we have provided an ablation study of the different blocks used in the architecture of the proposed action recognition techniques. We have considered different algorithms: Video Swin Transformer with Gamma Intensity Correction (GIC), Video Swin Transformer with GIC and GCN, Zero-DCE with Video Swin Transformer along with BERT, Zero-DCE with $R(2+1)D$ followed by GCN and BERT. The Top-1 and Top-5 accuracy for all the mentioned techniques with proposed Zero-DCE with $R(2+1)D$ followed by the GCN and BERT techniques are provided in Table III. It may be observed that in comparison to all the mentioned architecture, the proposed architecture (Zero-DCE with $R(2+1)D$ followed by GCN and BERT) has provided a higher action recognition accuracy in terms of both Top-1 and Top-5 accuracy.

TABLE III: Ablation Study of the Top-1 and Top-5 accuracy results on ARID V1.5 Dataset

Models	Top-1 Accuracy	Top-5 Accuracy
GIC + Video Swin Transformer	70.07%	95.50%
GIC + Video Swin Transformer + GCN	71.51%	96.52%
Zero-DCE + Video Swin Transformer + BERT	74.71%	98.53%
Zero-DCE + $R(2+1)D$ + BERT + GCN	83.90%	98.72%
$R(2+1)D$ **+ GCN + BERT(Proposed)**	**86.93** %	**99.35**%

We have also conducted an ablation study to see the performance of the GCN with different layers and found that 2-layer GCN performs better than 1 and 3-layer GCN. The Tables IV-V provide the quantitative comparison of proposed networks using different GCN layers on ARID V1.0 as well as ARID V1.5 datasets. It may be observed that two layers of GCN are found to be providing better results than 1 and 3 layers.

TABLE IV: Performance comparison of proposed network with different layers of GCN on ARID V1.0 database.

Models	Top-1 Accuracy	Top-5 Accuracy
Proposed method with single layer GCN	96.35%	99.88%
Proposed method with two-layer GCN	**96.60**%	**99.88**%
Proposed method with three-layer GCN	95.87%	99.88%

TABLE V: Performance comparison of proposed network with different layers of GCN on ARID V1.5 database.

Models	Top-1 Accuracy	Top-5 Accuracy
Proposed method with single layer GCN	86.65%	99.26%
Proposed method with two-layer GCN	**86.93**%	**99.35**%
Proposed method with three-layer GCN	85.41%	98.87%

We have also performed an ablation study for the use of min-max sampling against delta sampling [29] and network without sampling. Table VI provides the quantitative comparison of the above said approaches. It may be observed that the proposed scheme's top-1 accuracy of 96.60 surpassed the proposed network without sampling as well as the delta sampling.

TABLE VI: Top-1 accuracy results of different sampling methods on ARID V1.0 dataset

Models	Top-1 Accuracy
Delta Sampling Resnet-BERT	90.46%
Proposed method without Min-max sampling	95.63 %
Proposed method with Min-max sampling	**96.60** %

E. Discussions and Future Works

Hira et al. [29] proposed an action recognition framework for dark videos, where the Zero-DCE is used to enhance the input low-light/dark video sequences which were further sampled using delta-sampling technique and fed to $R(2+1)D$ to extract the features. These features are further classified into different actions. Although the said approach is found to be worthy but does not utilize the long-range temporal structure of videos which plays a vital role in action recognition

Similarly, Chen et al.[30] proposed a dark-light network that uses firstly gamma intensity correction to enhance the raw low-light/dark videos which are future fed to a feature extractor (ResNext-101 as well as $R(2+1)D$) to extract the features. The said extracted features are used for action classifications. It is observed that the gamma correction does not improve the image enhancement result irrespective of the value of gamma. It happens due to the varying degrees of darkness across the dataset which cannot be captured properly by the gamma correction table. Further, the feature extractor used in dark-light is not able to effectively capture temporal correlations, especially for long-term temporal dependencies of features in dark videos.

As compared to the above-mentioned approaches, in the proposed scheme we have adhered to the use of Zero-DCE to enhance the low-light/dark videos followed by the min-max sampling. Further, we consider a $R(2+1)D$ network to extract the spatio-temporal features. A GCN is used as the temporal feature encoder to utilize the long-range relationships among the features obtained by the $R(2+1)D$ network. It models intrinsic temporal relations which provides a robust encoded representation for action recognition. Further BERT is used for action classifications.

In the proposed scheme, we have used GCN on top of a $R(2+1)D$ convolutional neural network as our video feature encoder. Specifically, for a video V with n clips, the feature extractor maps the clips into the corresponding sequence of features, which alone do not incorporate the rich temporal structure of the video. Therefore, we use a temporal graph encoder that constructs a fully connected graph on top of the clip-level features, with learnable edge weights through a parameterized adjacency matrix, as in [13]. With these graph representations, we apply a graph convolutional neural network with two layers over all the node features to output the encoded representation of the video V.

Capturing long-range temporal structure in videos is crucial for action recognition, which in turn affects the overall generalization performance of a model. We adopted a fully connected GCN with reference to some prior works - e.g. [14] and used it to capture long-range relationships among the clip features. Also, we would like to mention here that, in the use of graphs it is observed that smoothing is a natural effect

of adding more layers to the GCN model. Over-smoothing happens when the receptive field of two GCN nodes overlaps, which results in similar embedding to those nodes. This means the nodes contain the same information which causes the over-smoothing. However we are not learning the node embeddings, as we have given output features of $R(2+1)D$ as the node features to the GCN, and our network learns the edge weight only. We are using only two layers of GCN and not getting the adverse effects of over smoothing. We also tried with different layers of GCN and checked the performance and adhered to the use of 2 layer GCN to avoid over smoothing.

The Zero-DCE method produces an enhanced result through image-specific curve mapping. Such a strategy enables light enhancement on images without creating unrealistic artifacts [42]. Hence we used Zero-DCE to enhance the low-light/dark video frames in our case and we do not face noise amplification problems.

The creation of a comprehensive dataset for dark video action recognition might be one of the goals of future research. In future work, we are also planning to use the K-nearest neighbor graphs, to explore its capabilities for action recognition. Further, We would like to employ the meta-learning mechanism to extract the distinctive and domain invariant features to represent the visual cues in the dark/ low light videos. This enables transfer of action classification models across the videos captured with diverse environment and action configurations.

V. Conclusions

This paper addresses the unexplored task of action recognition in dark or low-light videos. We first explored an image enhancement technique using Zero-DCE, followed by the min-max sampling technique. Further, we proposed an action classification network to classify the actions in enhanced video scenes. In the proposed action classification network, we have used the $R(2+1)D$ network to extract both spatial and temporal features. To capture the long term temporal dependencies of the extracted features, we adhered to the use of GCN to characterize the input video by robust spatio-temporal features. We have used BERT for the classification of the actions in a video scene. The proposed scheme is verified with two versions of ARID dark video database: ARID V1.0 and ARID V1.5. The efficiency of the proposed scheme is verified by comparing it against fifteen SOTA techniques. We corroborate our findings by taking two performance evaluation measures: Top-1 and Top-5.

References

[1] L. Zhang, S. Z. Li, X. Yuan, and S. Xiang, "Real-time object classification in video surveillance based on appearance learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[2] Z. Cai, H. Neher, K. Vats, D. A. Clausi, and J. Zelek, "Temporal hockey action recognition via pose and optical flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 2543–2552.

[3] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–6.

[4] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1623–1631.

[5] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, 2011, pp. 611–622.

[6] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 046–12 055.

[7] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 056–12 065.

[8] "Is space-time attention all you need for video understanding?"

[9] I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph, "Revisiting ResNets: Improved training and scaling strategies," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 22 614–22 627.

[10] M. Kalfaoglu, S. Kalkan, and A. A. Alatan, "Late temporal modeling in 3D CNN architectures with BERT for action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 731–747.

[11] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3185–3194.

[12] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7324–7333.

[13] A. Sahoo, R. Shah, R. Panda, K. Saenko, and A. Das, "Contrast and mix: Temporal contrastive video domain adaptation with background mixing," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 23 386–23 400.

[14] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 399–417.

[15] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1780–1789.

[16] 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)*, 2018, pp. 6450–6459.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.

[21] 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)*, 2017, pp. 6299–6308.

[22] 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)*, 2015, pp. 4489–4497.

[23] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017, pp. 3154–3160.

[24] K. Hara., H. Kataoka, and Y. 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*, 2018, pp. 6546–6555.

[25] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3202–3211.

[26] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6836–6846.

[27] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "Arid: A new dataset for recognizing action in the dark," in *International Workshop on Deep Learning for Human Activity Recognition*. Springer, 2021, pp. 70–84.

[28] H. R. Patel and J. T. Doshi, "Human action recognition in dark videos," in *International Conference on Artificial Intelligence and Machine Vision (AIMV)*, 2021, pp. 1–5.

[29] S. Hira, R. Das, A. Modi, and D. Pakhomov, "Delta sampling R-BERT for limited data and low-light action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 853–862.

[30] R. Chen, J. Chen, Z. Liang, H. Gao, and S. Lin, "Darklight networks for action recognition in the dark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 846–852.

[31] T. Mertens *et al.*, "Exposure fusion: A simple and practrical alterrnative to high dynamic range photography," *Computer Graphics Forum*, vol. 28, no. 1, pp. 161–171, 2009.

[32] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Institute*, vol. 310, no. 1, pp. 1–26, 1980.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2021, pp. 8024–8035.

[35] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[36] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proceeding of the European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[38] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[39] O. Kopuklu, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3d convolutional neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 1910–1919.

[40] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[41] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D Residual networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5533–5541.

[42] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1780–1789.

Himanshu Singh received the B.Tech. degree in Electrical engineering from the Uttar Pradesh Technical University (UPTU), Lucknow, India, in 2011, and the M.E. degree in Signal Processing from Indian Institute of Science (IISc), Banaglore, India, in 2013. He is currently pursuing a Ph.D. degree in Electrical Engineering with the Indian Institute of Technology Jammu. His research interests include image signal processing, computer vision, and machine learning.

Saurabh Suman is pursuing a B.Tech. Degree in Electrical engineering from the Indian Institute of Technology, Jammu, India. He will be graduating in the year 2023. His research interest includes Action recognition in videos and Image processing.

Badri Narayan Subudhi (S07, M17, SM19) received B.E in Electronics and Communication Engineering from Bijupatnaik University of Technology, Odisha, and an M.Tech. in Electronics System & Communication from the National Institute of Technology, Rourkela, India, in 2008-09. He worked for his Ph.D. in Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India, in 2014. Currently, he serves as an Assistant Professor at the Indian Institute of Technology Jammu, India. Before this, he worked as an Assistant Professor at the National Institute of Technology Goa from July 2014 to March 2017. His research interests include Video Processing, Image Processing, Machine Learning, Pattern Recognition, and Remote Sensing Image Analysis.

Vinit Jakhetiya received the B.Tech. degree in computer and communication engineering from the LNM Institute of Information Technology, Jaipur, India, in 2011, and the Ph.D. degree in electronics and computer engineering from the Hong Kong University of Science and Technology, in 2016. From January 2015 to December 2015, he was a Visiting Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Later, he joined as a Project Officer with the same university. Currently, he is working as an Assistant Professor of computer science at the Indian Institute of Technology, Jammu, India. His research interests include image/video processing, image quality assessment, and visual perceptual modeling.

Ashish Ghosh (Senior Member, IEEE) is a Professor with the Machine Intelligence Unit, Indian Statistical Institute. He has authored or coauthred around 250 research papers in internationally reputed journals and refereed conferences, and has edited eight books. His research interests include pattern recognition, machine learning, data mining, image and video analysis, soft computing, neural networks, evolutionary computation, and bioinformatics. He was the recipient of the the Young Scientists Award from the Indian Science Congress Association in 1992 and the Indian National Science Academy in 1995. He is acting as a Member of the Editorial boards of various international journals.