LIFT-Net: Dual Stream Fourier Attention For Light Invariant Action Recognition

Onkar Susladkar[1,*], Gayatri Deshmukh[1,*], Atharva Chandak[1,*], Santosh Kumar Yadav[2,3,*], Kamlesh Tiwari[1,2], Hari Mohan Pandey[4], and Peter Corcoran[3]

***Abstract*—Recent methods for video action recognition struggle in low-light videos due to the lack of contextual information. Nevertheless, action recognition in low-light videos is crucial due to its numerous applications, including surveillance, autonomous driving at night, *etc.* In this paper, we propose LIFT-Net (Light Invariant Fourier Transformer Network), a novel dual pathway neural network architecture for light-invariant human action recognition. The first (Main-Stream) pathway takes low-light video frames as input, extracts features using the Partial ViT (Vision Transformer) network, and aggregates it along the time axis. The aggregated features are then enhanced by the Feature Fusion module. The second (Luminous) pathway uses gamma-corrected video frames for providing cross-visual attention. Additionally, Fourier Enhancement (FoE) block assists the model by leveraging the Fourier Transform to learn high and low-frequency feature maps, which aids the model to focus on important regions. We evaluated LIFT-Net's ability to recognize actions in videos with low illumination (ARID dataset), normal illumination (HMDB51 and UCF101 datasets), and infrared videos (InfAR dataset). It advances the state-of-the-art by 1.14% on ARID, 2.20% on HMDB51, 0.12% on UCF101, and 2.65% on the InfAR datasets. These results demonstrate the reliability of our proposed model in any illumination condition.**

***Index Terms*—Action Recognition in the Dark, Fourier Transform, Gamma Intensity Correction, Knowledge Distillation, Optical Flow, Vision Transformer, Cross visual attention**

I. Introduction

Humans perform a range of mundane and formal activities, such as driving, cleaning, playing video games, *etc*. To design a human-computer interaction system, it is essential to recognize the human action being performed. Human action recognition is a well-known problem since the 1980s [1]. It encompasses training our machine to infer high-level actions by learning intricate motion patterns from low-level sensing data. It has crucial applications in a vast array of tasks, including gaming, human-robot interaction, rehabilitation, sports analysis, sign language recognition, health monitoring, video surveillance, and robotics [2]. Human body articulations, illumination variations, clothes, background artifacts, clutter, and occlusions present challenges for action recognition systems.

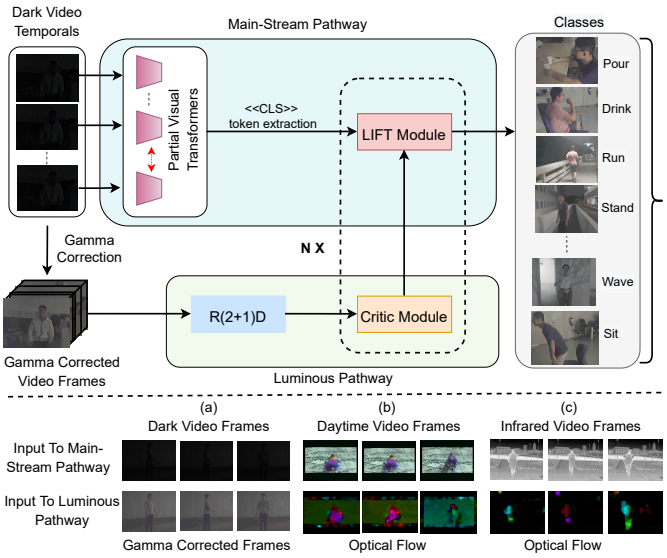

Fig. 1. The proposed architecture's schematic diagram. Firstly the gamma-corrected video frames are created from the given dark video temporals. Then it uses those corrected frames and original temporals, to train a classification network. The symbol $\times N$ indicates that the LIFT and Critic modules repeat N times in the network. Additionally, the red bidirectional arrow indicates that PViTs weights are shared. The dark video frames, daytime video frames, and infrared video frames are inputs to the LIFT-Net's Main-Stream and Luminous pathway, respectively. For the low-light videos, the Luminous pathway uses gamma-corrected input. Optical Flow information is fed to the Luminous pathway for daytime and infrared videos.

These inherent problems have led to the development of numerous approaches for human action recognition employing visual and non-visual sensors. Visual sensors comprise RGB, RGB-D, infrared, time-of-flight, multispectral, light field, and thermal cameras, among others. Non-visual sensors include accelerometers, gyroscopes, magnetometers, audio signals, and electro-thermal responses, among others. This research focuses on computer vision-based light invariant action recognition.

Action recognition has achieved remarkable progress during the last several years and is mostly dominated by vision-based approaches. Most modern methods make use of videos recorded in a wide range of data modalities like RGB, RGB-D, Infrared, *etc*., depending on the intended usage. Due to the low cost, wide availability, and simplicity of capturing videos, RGB video-based datasets are the most widely used for human action recognition. Large-scale publicly available RGB datasets, such as HMDB51 [3], UCF101 [4], Kinetics [5], *etc*., have facilitated major advancements in the field and helped in the creation of numerous deep-learning architectures.

* Equal contributions.

1. Deptartment of Computer Science and Information Systems, BITS Pilani, Rajasthan, India. (e-mail: onkarsus13@gmail.com; dgayatri9850@gmail.com; and f20190062@pilani.bits-pilani.ac.in).

2. CogniX, Madhapur, Hyderabad, Telangana 500081, India. (e-mail: santosh.yadav@ieee.org, kamlesh.tiwari@pilani.bits-pilani.ac.in).

3. College of Science and Engineering, University of Galway, H91TK33, Ireland. (e-mail: peter.corcoran@universityofgalway.ie).

4. Department of Computing & Informatics, Bournemouth University, Poole, Dorset, BH12 5BB, United Kingdom. (e-mail: hpandey@bournemouth.ac.uk).

However, most existing action recognition models are heavily data-driven and struggle significantly in bad conditions such as in low-light videos.

Action recognition in low-light videos has important applications in surveillance and autonomous driving at night. A few studies have tried to recognize human actions in dark videos [6]–[8]. The ARID dataset [9] is the only publicly available benchmark dataset dedicated to action recognition in a dark environment. However, for capturing actions in the dark, thermal imaging-based infrared cameras have traditionally been preferred over visible light. In addition to low-illumination conditions, infrared videos also make it simpler to analyze actions in challenging scenarios like occlusion, complex backgrounds, *etc*. There exist a few publicly available infrared video action recognition datasets like InfAR [10], IITR-IAR [11], and InfDet [12], which are commonly used for evaluating the performance of action recognition models.

A. Recent Advances

Action recognition techniques have evolved rapidly and a number of new approaches have been proposed. Three-dimensional convolutional neural networks (3D CNNs) like R(2+1)D [13] and C3D [14] networks, are among the most popular approaches for video classification tasks. The recent success of Transformer Networks in computer vision led to the development of models such as TimeSformer [15] and VidTr Transformers [16], which better model RGB videos under normal illumination conditions. The attention modules of these architectures allow them to focus on the context better while classifying the videos. For low-light RGB videos, the Darklight [7] network has emerged as the state-of-the-art due to its ability to efficiently represent features by leveraging CNN-based two-stream networks and learning complementary features in each stream, which are then combined using a self-attention module. Due to the challenging nature of the problem, progress on infrared action recognition has been relatively slow. The networks [11], [17] based on 3D CNNs, together with Optical Flow features and actor-driven pose embeddings, have made significant progress in this area.

B. Challenges

The main challenge for light invariant action recognition is to create an architecture that can capture variation in each of the various modalities. Every modality has its own set of distinguishing characteristics, and it is difficult to focus on all of them. Infrared and visible RGB videos, for example, have separate data spaces. Because there is a modality gap between the two, strategies for recognizing actions in visible light are not very effective at recognizing actions in infrared videos.

Despite the fact that infrared videos have been widely used for dark action recognition, the major challenges are the necessity for specialized sensors and the need of designing unique architectures. Furthermore, these videos lack color and texture information, making them even more difficult to work with. Therefore, it is important to build effective techniques for low-light RGB videos. One of the primary reasons for the lack of progress in this area is a scarcity of benchmark RGB video-based datasets for action recognition in the dark. The recently made public ARID [9] dataset has expanded the scope of this task. Until now, the majority of known architectures' performance on the ARID dataset has been limited due to these models' inability to generalize in low-light settings.

C. Motivation

Despite recent advances in the field, the emphasis on light invariant action recognition remains limited. To fill the gaps in existing work, we explore how a single model can be used effectively while working with multiple data modalities. Two stream networks have proven to work well by separating two pathways to concentrate on different aspects of the videos. We also use knowledge distillation (KD) with a student-teacher network to improve our model's feature representation learning and training efficiency. Furthermore, Fourier transform-based networks were effective at capturing and incorporating temporal-level features into our designed architecture. The schematic diagram of the proposed architecture is shown in Fig. 1.

D. Our Contributions

In this paper, we present a novel Transformer-based two-stream network that can function effectively even in difficult environments with poor lighting and illumination. Our model uses gamma-corrected video frames to improve the visibility of low-light videos, while the RGB input stream is used to capture spatial details. The major contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the earliest work to target light invariant action recognition in two modalities, $i.e.$, RGB and infrared. Moreover, the input RGB videos could either be well-illuminated or in the dark.
- We present the LIFT-Net (Light Invariant Fourier Transformer Network) model, which consists of two different pathways, $i.e$, Main-Stream and Luminous. The Main-Stream pathway is used to extract the features from the Dark RGB temporals and the Luminous path is designed to provide cross-visual attention in order to acquire complementary features by combining the characteristics of both pathways. In addition, we show how the use of the Fourier transform to extract composite characteristics from the frequency domain can significantly enhance performance (III-E).
- The resiliency of LIFT-Net is demonstrated by conducting rigorous experiments on videos captured in different environments. The proposed model has been evaluated on four publicly available benchmark datasets $i.e.$ ARID [9], HMDB51 [3], UCF101 [4], and InfAR [10] datasets.fff
- LIFT-Net achieves 95.21%, 89.92%, 98.89%, and 93.87% top-1 accuracies on ARID, HMDB51, UCF101, and InfAR datasets, respectively. LIFT-Net improves the state-of-the-art on these datasets by a margin of 1.14%, 2.20%, 0.12%, and 2.65% respectively. These results are impressively high for recommendation toward general adaption of the system.

The rest of the paper is structured as follows: Section II reviews the existing literature on action recognition in various modalities and their approach. Section III, presents the proposed methodology of the LIFT-Net model. We describe the architecture based on dark videos, however, it similarly extends to other sensing modalities too. In Section IV, the training methodology and optimization of the teacher-student network are described. In Section V, the results obtained on the ARID, HMDB51, UCF101, and InfAR datasets are summarised. Section VI presents ablation studies of the different LIFT-Net components. Finally, Section VII concludes our research and highlights the future scope.

II. Related Works

There has been extensive study on computer vision-based action recognition. As we are targeting normal illumination, infrared, and low-light video-based action recognition, this section provides an overview of the most significant developments in these fields. Typically, two-stream networks, 3D-CNNs, and Transformer-based architectures are utilized for normal illumination RGB video-based action recognition. Further, for action recognition in the dark, infrared and low-light video-based methods are outlined. TABLE I provides an overview of the literature review.

A. Two-stream Networks

Two-stream approaches take advantage of the diversity of information in videos by constructing two or more parallel feature extraction pathways, resulting in features with rich information modalities. Simonyan *et al*. [18] developed a two-stream CNN model that included a spatial network in addition to a temporal network. Specifically, given a video, each RGB frame and multi-frame-based Optical Flow were fed to the spatial stream and temporal stream, respectively. Consequently, these two streams learned visual and motion cues for action recognition. Using a two-stream convolutional neural network, Wang *et al*. [35] extracted convolutional feature maps from multi-scale video frames and optical fluxes. A number of other research works [20], [21], [24] have similarly extended the two-stream strategy to extract persistent video-level information for HAR. Nevertheless, because Optical Flow information is widely recognized as a useful complement to RGB information, it has been widely utilized for two-stream approaches [18], [19], [21].

B. 3-Dimensional CNNs

Numerous studies [14], [36], [37] have converted 2D CNNs into 3D structures to simulate the spatial and temporal context data in videos. In an end-to-end learning framework, Tran et al. *et al*. [14] created the C3D, a 3D CNN model for learning spatio-temporal features from raw videos. Carreira et al. *et al*. [22] proposed the Inflated 3D CNN (I3D) by inflating the convolutional and pooling kernels of a 2D CNN to take advantage of the additional temporal dimension. To capture long-term temporal dependencies, Wang *et al*. [25] combined a two-stream 3D CNN with an LSTM model. Consequently, networks such as 3D-ResNet [23] and 3D-ResNext [38] could be deeper because they are more advanced and comprehensive than simple 3D-CNNs; however, their computational time is high. To further enhance the efficacy of video feature extraction, the R(2+1)D [13] architecture has recently been developed to divide 3D Spatio-temporal convolutions into spatial and temporal convolutions.

C. Transformer-based Networks

After observing the success of Transformers in Natural Language Processing (NLP), numerous works, such as [29], [30], [39], have employed them for human action recognition in videos. Using a 3D CNN, a Faster-RCNN model [40], and a Transformer, Girdhar *et al*. [41] presented Video Action Transformer Network for action localization and recognition. After 3D convolutions, Kalfaoglu *et al*. [27] added late temporal modeling by introducing BERT at the end, which allowed the model to learn context from both directions; however, this makes it challenging to apply to real-time videos. In their TimeSformer work, Bertasius *et al*. [15] essentially extended the ViT [39] to the video domain by decomposing videos into a series of frame-level patches. After that, many methods were devised in order to individually apply temporal and spatial attention within each individual block of the model and capture Spatio-temporal attention. In addition, TimeSformer essentially used a technique called divided attention, in which temporal and spatial attention are independently applied within each block. VTN [28] successively employs a Transformer design over the spatial and temporal dimensions. By removing uninformative features along the temporal dimension, VidTr [16] suggested standard deviation-based top-k pooling for attention which decreases computation.

D. Action Recognition in Dark RGB Videos

Due to the lack of benchmark datasets, very few works have been proposed for dark or low-light videos based action recognition. Only one publicly available benchmark dataset named ARID [9] is commonly used. Darklight [7] presented a two-stream network that utilized the raw RGB frames in one pathway and the Gamma-corrected frames in the other. They fuse the two pathways using a self-attention mechanism. In order to compute the Optical Flow, they had a greater need for storage and processing power, so they utilized Gamma Intensity Correction (GIC). In their work on delta-sampling R(2+1)D BERT, Hira *et al*. [33] utilized the zero-DCE image enhancement method and R(2+1)D BERT as the classifier. Z-domain entropy adaptable flex [34] used a similar architecture in the semi-supervised context along with the flexible pseudo-label generation method by successively increasing and decreasing entropy. Various domain adaptation approaches [42], [43] have also been proposed which try to address the recognition in the dark as an out-of-distribution and domain shift problem. However, their success has been limited due to the large gap in the low-level features of the dark videos as against normal videos, which makes it harder to align them.

TABLE I
SUMMARY OF THE LITERATURE REVIEW

References	Sensing Modality	Learning Methodology	Dataset(s)
Simonyan *et al*. [18], 2014	Visible-RGB (normal illumination)	Two stream networks with spatial and temporal streams	UCF101 and HMDB51
Tran *et al*. [14], 2015	Visible-RGB (normal illumination)	Proposed C3D architecture	UCF101
Wang *et al*. [19], 2015	Visible-RGB (normal illumination)	Learned discriminative deep convolutional features using spatio-temporal and channel normalization to compute Trajectory-pooled Deep convolutional Descriptors (TDDs)	HMDB51 and UCF101
Girdhar *et al*. [20], 2017	Visible-RGB (normal illumination)	Joint space-time aggregation of a video's appearance and motion streams	UCF101, HMDB51, and Charades
Feichtenhofer *et al*. [21], 2017	Visible-RGB (normal illumination)	Two-stream appearance and motion pathways fused by motion-gating	HMDB51 and UCF101
Carreira *et al*. [22], 2017	Visible-RGB (normal illumination)	Two-stream Inflated 3D CNNs	HMDB51 and UCF101
Hara *et al*. [23], 2017	Visible-RGB (normal illumination)	3D CNNs based on ResNets	ActivityNet and Kinetics-400
Diba *et al*. [24], 2017	Visible-RGB (normal illumination)	TLE (Temporal Linear Encoding) layer embedded to capture appearance and motion	HMDB51 and UCF101
Wang *et al*. [25], 2018	Visible-RGB (normal illumination)	Video decomposition into spatial and temporal domains, followed by spatio-temporal pyramid pooling and two-stream 3D fusion	HMDB51, UCF101, and ACT
Tran *et al*. [26], 2018	Visible-RGB (normal illumination)	R(2+1)D Network	Sports-1M, Kinetics-400, UCF101, and HMDB51
Kalfaoglu *et al*. [27], 2021	Visible-RGB (normal illumination)	3D-CNNs with BERT for temporal attention	HMDB51, UCF101, Kinetics-400, and IG65M
Bertasius *et al*. [15], 2021	Visible-RGB (normal illumination)	Spatio-temporal Transformer with divided attention, Sparse local-global attention, and Axial attention	Kinetics-400, Kinetics-600, SS-V2, and Diving-48
Zhang *et al*. [16], 2021	Visible-RGB (normal illumination)	Transformer network with top-k pooling for attention	UCF101, HMDB51, Kinetics, Charades, SS-V2, and Kinetics-700
Neimark *et al*. [28], 2021	Visible-RGB (normal illumination)	Temporal Attention Transformer	Kinetics-400 and Moments in Time
Ryoo *et al*. [29], 2021	Visible-RGB (normal illumination)	Adaptively learn tokens and pairwise model for attention, allowing them to attend to a larger spatial and temporal regions	Kinetics-400, Kinetics-600, Charades, and AViD
Li *et al*. [30], 2022	Visible-RGB (normal illumination)	Transformer architecture combining CNN and self-attention properties	Kinetics-400, Kinetics-600, SS-V1 and SS-V2
Liu *et al*. [31], 2018	Infrared and RGB+D	TSTDDs features extracted using a 3-stream network and classified by SVM	InfAR and NTU RGB+D
Imran *et al*. [11], 2019	Infrared Videos	Generate SSDIs to feed into four-stream network with two CNNs and two CNN-BiLSTM networks	InfAR and IITR-IAR
Khebli *et al*. [17], 2019	Infrared Videos	3D CNNs with Optical Flow	InfAR
Lamghari *et al*. [32], 2022	Infrared Videos	Key pose features identification using clustering and classifying the embeddings	InfAR
Chen *et al*. [7], 2021	Visible-RGB (low illumination)	2-stream network with self-attention and fusion	ARID
Hira *et al*. [33], 2021	Visible-RGB (low illumination)	Zero-DCE frame enhancement with R(2+1)D BERT classifier network	ARID
Chen *et al*. [34], 2022	Visible-RGB (low illumination)	R(2+1)D with self-attention for combining features from dual pathways	ARID

E. Action Recognition in Infrared Videos

Infrared videos are mostly preferred for dark video action recognition. Khebli *et al*. [17] proposed the use of Optical Flow maps and an entropy-based algorithm for filtering frames containing human actions. They pass these filtered frames to a 3D CNN which performs the classification on it. TSTDDs [31] put forth a new global temporal representation called Optical Flow stacked difference image (OFSDI) which is based on local, global, and spatial-temporal information. Imran *et al*. [11] put forth a four-stream framework consisting of two CNN and two CNN-BiLSTM networks which take in Stacked Dense Flow Difference Image (SDFDI) and Stacked Saliency Difference Image (SSDI) maps generated from the videos as input. This output is then fused to predict the action class. The latest work on actor-driven pose embeddings introduced in [32], identified the key actors identified using the Optical Flow features. These were then utilized to generate representations of the key poses selected via deep clustering.

III. PROPOSED METHODOLOGY

We build LIFT-Net to recognize activities under different lighting situations, $i.e.$, low-light and normal illumination. Fig. 2 demonstrates that the overall architecture is comprised of two pathways: the Main-Stream pathway and the Luminous pathway. The Main-Stream pathway's input is a series of dimmed night video frames, whereas the Luminous pathway's input is a 3D sequence of stacked frames generated by applying the Gamma Correction technique to dark video frames, thereby enabling the Main-Stream pathway to be seen with the assistance of brightened video frames.

This section is organized as follows. Section III-A presents the overall picture of the proposed architecture. Sections III-B and III-C describes the Luminous and Main-Stream pathways, respectively. The Feature Fusion module is defined in Section III-D, and the FoE (Fourier Enhancement) block is explained in Section III-E.

A. LIFT-Net Architecture

Fig. 2 depicts the inputs of LIFT-Net, which is a dual-stream network. A low-light, dark video serves as the input for the Main-Stream pathway, while gamma-corrected voxels of video frames are provided for the Luminous pathway. The feature vector from the Main-Stream pathway is combined with the output from the Luminous pathway. The Main-Stream subtree

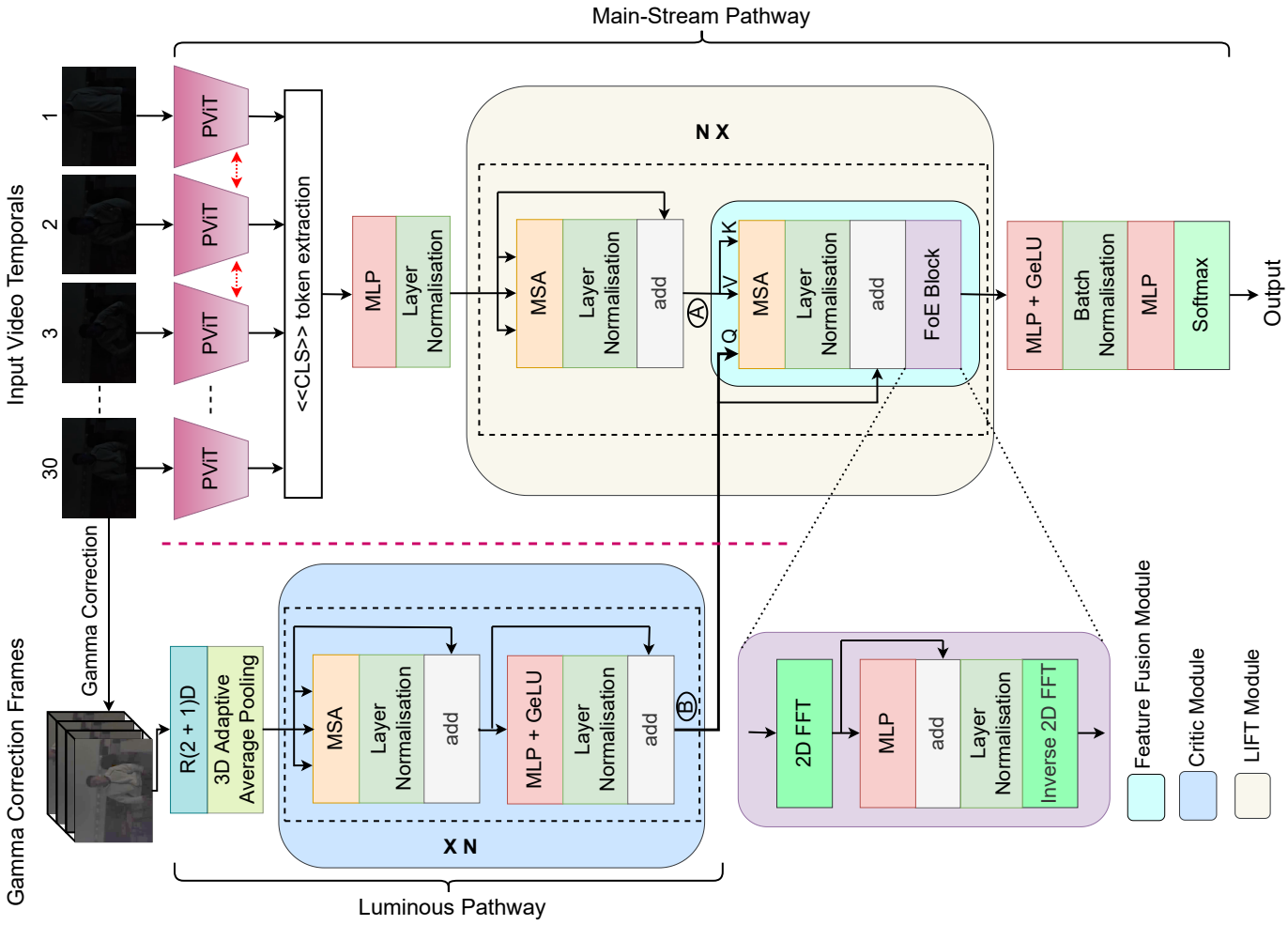

Fig. 2. The overall architecture of LIFT-Net. For classification, the suggested model incorporates dark video frames and visual cues. The network has two branches: the Luminous pathway (lower branch) accepts gamma-corrected video frames as input, while the Main-Stream pathway (upper branch) accepts the original dark video frames as input. Note that the bi-directional red arrows signify that all PViTs (Partial ViTs) share the same weights. In order to enable the model to distinguish actions in low light, features learned by the Luminous pathway are subsequently sent to the Main-Stream to lend visual attention to no-light video frames. The FoE (Fourier Enhancement) block also allows representation learning taking into account the frequency domain. The LIFT and Critic modules are applied N times to capture more complex features.

has a Feature Fusion module where the actual Fusion occurs. This fusing step is repeated N times throughout the network, assisting the model in utilizing the learned information from the Luminous pathway and so attaining better results. The subsequent sections will provide a comprehensive explanation of this fusion component. Following the union of feature maps from both paths, the resulting feature vector is fed into an MLP, where class probabilities are calculated using softmax activation.

B. Luminous Pathway

The input for the Luminous pathway is gamma-corrected video frames, which are 3D inputs of stacked frames. Low-light night-time video frames are processed with the Gamma Intensity Correction (GIC) method to produce these gamma-corrected video frames. By doing so, the local dynamic range of the image is enlarged in low-light regions while compressed in high-light regions. Therefore, adjusting the image's brightness globally. The purpose of utilizing the GIC algorithm is to make night-time video frames more visible. Referring to Fig. 3, which depicts a video frame with no light at night and the corresponding gamma-corrected video frame. It is evident that video frames processed with GIC improve the clarity of the frames.

The stacked video frames obtained from the gamma correction are fed into the Luminous pathway. In the first stage, it goes through an R(2+1)D network that has already been trained using the IG65M dataset. R(2+1)D is preferred because of the emphasis it places on spatiotemporal data, which aids in the efficiency of the learning process. In order to make the R(2+1)D network's output vector a good fit for the critic module, it is fed through a 3D adaptive average pooling. The resulting vector is then used as input by the critic module.

Critic Module: By facilitating visual cross-attention with the aid of the Feature Fusion module in the Main-Stream pathway, the critic module aids in reducing the risk associated with target classification. As this module serves to reduce the number of classification errors made by the Main-Stream pathway, hence the name critic.

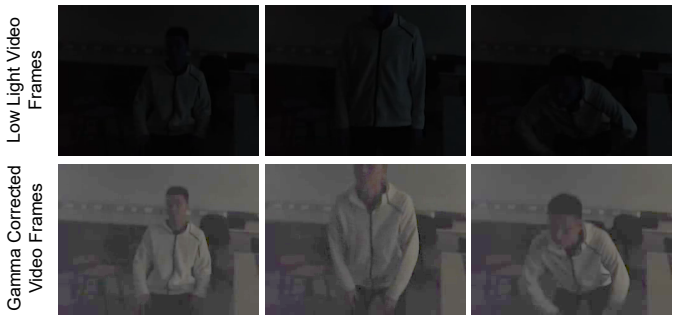

Fig. 3. Sampled frames from the ARID dataset, which contains videos recorded in low light, are displayed in the first row, while the frames from the first row are augmented in the second row using the Gamma Intensity Correction (GIC) technique.

The critic module's input serves as the MSA (Multi-head Self Attention) block's key, query, and value. By highlighting the most important details, MSA aids in the discovery and improvement of key features. In order to maintain the integrity of the network, the concealed representation of MSA is fed into a layer normalization, and the resulting feature map is then added with the input from the critic module. Following the addition, the feature vector is sent through MLP, GeLU (gaussian error linear unit) activation, and layer normalization. Once again, the residual connection is added at this stage, and the output is forwarded on to the Feature Fusion module in the Main-Stream pathway.

C. Main-Stream Pathway

This pathway accepts night-time video-derived temporal as input. Each video temporal image is an RGB image. The PViT (Partial ViT) is used to process all video temporals, as displayed in Fig. 2. PViT makes use of five ViT [39] blocks that have already been trained on the ImageNet-21k [44] dataset. We only keep the first block unfrozen throughout training. As depicted in Fig. 2, weights are shared among all PViT blocks, which are shown by the red bidirectional arrow. As we know, in ViT, the final embedding is obtained by the interaction between the CLS token and the patch token at every Transformer encoder. As a result, we consider CLS to be an operator that summarises all patch tokens, and CLS token extraction is performed to the learned representations of these PViT blocks. The resulting vector is then routed through MLP and layer-normalizing layers before being sent into the LIFT module.

LIFT Module: To begin with, the LIFT input is sent through the module's MSA to generate self-attention. The MSA's hidden representation is then transmitted to layer normalization, and a residual connection is added to keep information flowing throughout the network. The final feature vector is then sent on to the Feature Fusion module, which is discussed in greater detail in Section III-D. Finally, MLP, GeLU, batch normalization, another MLP layer, and a softmax activation function are applied to the output of the Feature Fusion module to derive class probabilities.

D. Feature Fusion Module

When training a model with dark video frames, the model overlooks action specifics while the Luminous pathway pays close attention to action distinctions. Because of this, merging feature maps from the Luminous pathway is an efficient method for building a functional, visually distinct representation. This is how the newly incorporated Feature Fusion module assists the model in learning the distinguishing feature using visual attention gained from the Luminous pathway. Feature fusion module includes MSA, layer normalization, residual connection, and FoE block. The module takes into account the feature vectors from the locations Ⓐ and Ⓑ in Fig. 2, and sends them on to the MSA.

Let H_{MS} be the feature vector obtained from the Ⓐ, which serves as the MSAs key and value. In a similar fashion, H_{LS} will be the hidden representation originating from Ⓑ location that will be served as a query to the MSA block.

Through the integration of the feature vector originating from the Luminous pathway, MSA creates attention-pooled characteristics that strengthen visual action information. The function $F_{attention}$ can be used to derive these pooled attention characteristics by the following formula:

$$F_{attention} = Softmax(\frac{q(H_{LS}) \times k(H_{MS})}{\sqrt{N_{AH}}}) \times v(H_{MS}) \quad (1)$$

where $q(.)$, $k(.)$, and $v(.)$ are linear transformation layers applied on proposal features and represent the query, key, and value, respectively. N_{AH} is the number of attention heads required for normalization. For more details on multi-head attention theory, refer [45]. To further control the gradient flow, the output of $F_{attention}$ is sent through layer normalization followed by residual connection (refer to Equation 2).

$$FE = LN(F_{attention}) + H_{LS} \quad (2)$$

where LN denotes Layer Normalization. The Fourier Enhancement (FoE) block (highlighted in purple in the Fig. 2) receives this FE as an input. The reasoning behind the FoE block is described in detail in Section III-E.

E. Fourier Enhancement (FoE) Block

In order to transform the feature map $FE \in \mathbb{R}^{B \times S \times E}$ into the frequency domain, we first execute a 2D fast Fourier transform (FFT) along the sequence S and embedding E dimensions.

$$Y = F(FE) \quad (3)$$

in which $F(.)$ stands for the 2D FFT. It is important to note that Y is a complex tensor that describes the FE spectrum. That is, Y gives the frequency distribution across the combined sequence of the Main-Stream and Luminous pathways. To make these frequency characteristics trainable, we feed Y to an MLP (multilayer perceptron) and then apply a residual connection and Layer Normalization (LN). Finally, we utilize the inverse Fast Fourier Transform (FFT) to revert the learnable modulated spectrum Y to the FE in the spatial domain:

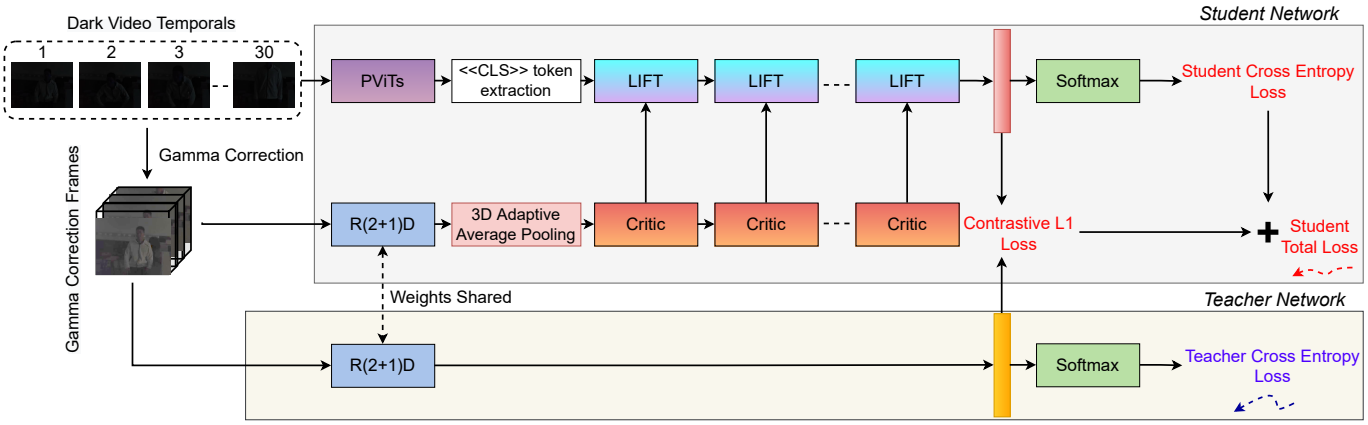

Fig. 4. Adapted teacher-student training paradigm for training LIFT-Net. In the training environment, R(2+1)D network plays the role of teacher, while LIFT-Net is the student. A black directional arrow indicates that the instructor network shares its weight with the R(2+1)D model present in the student network. In addition, the figure's blue and red arrows depict the backward propagation of teacher and student losses.

$$FE = F^{-1}(Y) \tag{4}$$

By boosting local information and dampening down the constraints caused by frames with insufficient lighting, the FoE block is able to get composite features.

IV. Model Training

This section describes our training method and the losses used to train our model. LIFT-Net is trained using a Knowledge Distillation (KD) learning strategy. The term Knowledge Discovery (KD) is used to describe a technique that aids in the training of a student network (LIFT-Net) by a teacher network (R(2+1)D model). Fig. 4 depicts the proposed training paradigm, which consists of two main parts: the first (lower part in Fig. 4) acts as a teacher network, utilizing gamma-corrected video frames (the advantaged information in this case); the second (upper part in Fig. 4) is a student network in the distillation framework, which takes dark video temporals and a gamma-corrected video frame as inputs.

The initial step in training is to train the teacher network. The R(2+1)D model, after being initialized with weights from the pre-trained IG65M dataset, seeks to minimize the cross entropy loss. Let IG be the 3D input of stacked, gamma-corrected video frames that have been fed through the teacher model $T(.)$, as indicated in the Fig. 4. In order to determine the class probabilities, the model first generates a set of hidden representations ($h1$), which are then used in conjunction with softmax to produce a vector (T_{output}) that is used to compute the teacher loss given as –

$$T_{LOSS} = -\sum_{x=i} y_i \times \log(T_{output_i}) \tag{5}$$

where y is the label of the ground truth. In this way, the teacher can learn to accurately foretell which action classes their students will attend, hence facilitating more efficient model learning. This occurs because the student network shares its weight with the teacher.

In this second phase of training, we will be primarily concerned with transferring knowledge from the teacher model to the student network. As input, the student network receives dark video temporal and gamma-corrected video frames. The model is trained using the loss that is defined in Equation 6, while also making use of the knowledge that was learned from the teacher network. In our experience, this facilitates the learning procedure. Let $S(.)$ be the student network and $h2$ is the feature map obtained from $S(.)$. Further, let S_{output} be the output vector after applying softmax to $h2$, then, the student loss is computed as:

$$S_{LOSS} = -\sum_{x=i} y_i \times \log(S_{output_i}) + \sum_{x=i} |h1_i - h2_i| \tag{6}$$

The first term of S_{LOSS} makes use of the cross entropy function in an effort to reduce the student classification loss, and the second term makes use of the L1 loss in an effort to minimize the absolute difference that exists between the student vector and the teacher vector.

V. Experimental Results

A. Dataset Details

LIFT-Net is trained and tested across four datasets: ARID (Action Recognition in the Dark) [9], HMDB51 [3], UCF101 [4], and InfAR [10]. The ARID dataset comprises of dark, no-light video frames demonstrating the model's capacity to distinguish actions in no-light conditions. On top of that, we demonstrate our model's efficacy on the HMDB51 and UCF101 datasets, both of which have videos that have been exposed to visible light as well as the InfAR dataset containing infrared videos, to highlight that it is not limited to identifying actions in low-light scenarios.

Action Recognition in the Dark (ARID): The purpose of this dataset is to facilitate the process of identifying actions in low-light videos. With at least 110 clips in each class, the ARID dataset has a total of 3,784 video clips. There are 11 different types of common human actions. Person actions with objects include drinking, picking, pouring, pushing, sitting, and standing. Person actions without objects include jumping, running, turning, walking, and waving.

HMDB51: The HMDB51 dataset consists of realistic videos drawn from a wide range of media, including movies and online videos. The dataset includes 6,766 video clips, each of which falls into one of the 51 action categories. Each action category has at least 100 clips included in it.

UCF101: The UCF101 dataset is a collection of realistic action videos from YouTube that have been labeled into one of 101 action classes; each class has at least 100 video clips. UCF101 has a greater diversity in terms of actions, with 13320 clips spanning 101 action categories. Additionally, there are considerable changes in camera movements, object appearance and position, object scale, viewpoint, cluttered background, illumination circumstances, and so on.

InfAR: The InfAR dataset is made up of action recognition videos captured using infrared cameras. There are a total of 600 videos in 12 different action categories that make up the dataset. The videos are typically 4 seconds long, with a 25 frame-per-second frame rate, and a 293×256 resolution. All the videos were shot against one of three identical static backdrops. The actions are handclap, walk, one hand wave (wave1), multiple hands wave (wave2), jog, skip, handshake, jump, hug, push, punch, and fight.

B. Experimental Settings

Our network is trained in parallel on two NVIDIA RTX A6000 GPUs using the CUDA 11.2 and PyTorch framework. The network receives a sequence of frames with a size of $3 \times 224 \times 224$ and a length of 30. The training is performed with an AdamW optimizer with a batch size of 32 and an initial learning rate of 0.0001. A cosine annealing scheduler is utilized throughout the training process. Augmentations like random flip, color jitter, random affine transform, optical distortion, elastic distortion, random rain, cutmix, and frame shuffle are all applied to the training set.

C. Performance Evaluation

In this Section, we discuss the outcomes of our method and current competitive action recognition-based models. We evaluate LIFT-Net against the current best practices in the field. To illustrate our model's diversity, we tested it in three scenarios: 1. Dark or no-light, 2. daytime or visible, and 3. infrared videos. In all scenarios, which we will elaborate on, our model achieved state-of-the-art performance.

1. Dark or No-Light Scenarios: TABLE II compares the performance of multiple approaches with LIFT-Net on the ARID dataset. Note that in the Luminous sub-branch of the network, we use gamma-corrected video frames for the dark/no-light videos. To detect motion in low-light settings, models like I3D-Two-stream combine Optical Flow data with the original video frames. However, we conduct experiments to show that Optical Flow information is not relevant in videos with low lighting. The reasoning can be found in Section VI. Our findings also show that the gamma-corrected data makes it easier for the model to discern the pattern, rather than the Optical Flow. As far as we are aware, our model has the highest accuracy on ARID, at 95.21%.

TABLE II
TOP-1 ACCURACY RESULT ON THE ARID DATASET [9]. **BLUE** REPRESENTS THE PREVIOUS STATE-OF-THE-ART. **RED** DENOTES THE BEST RESULTS.

Model	Top-1 Accuracy
C3D [14]	39.17%
3D-ShuffleNet [46]	44.35%
3D-SqueezeNet [46]	50.18%
Pseudo-3D-199 [**?**]	71.93%
I3D-Two-stream [22]	73.39%
3D-ResNext-101 [38]	74.73%
DarkLight-ResNext-101 [7]	87.27%
Darklight-R(2+1)D-34 [7]	94.04%
Timesformer + Flow [15]	**94.07%**
LIFT-Net (ours)	**95.21%**

When we put these models next to others like 3D-ResNet-18 and 3D-ResNet-101, we find that adopting a more intricate and in-depth network structure is not enough to improve performance. In addition, DarkLight has a dual pathway and self-attention that helps it attain an accuracy of 94.04%, yet it falls short of LIFT-Net by 1.17%. Similar to LIFT-Net, DarkLight makes use of gamma-corrected frames and self-attention technique in the network as well. However, it employs self-attention at the very end of the network which inhibits it from perceiving and learning the more abstract elements that were present in the earlier layers. This is where LIFT-Net excels. When compared to TimeSFormer, which also uses a Transformer and distributed attention in the network, LIFT-Net achieved a higher score with a margin of 1.14%. LIFT-Net employs the proposed Feature Fusion module to fuse the learned representations from both network branches and repeating this step numerous times allows the model to learn critical information more effectively. In conclusion, the optimal outcome can be achieved by combining the dual pathway with a Feature Fusion module to apply cross-visual attention.

2. Daytime or Visible Scenarios: TABLE III and IV demonstrate that our proposed LIFT-Net produces the highest results on HMDB51 and on UCF101, respectively. Most of the results have been referred from [27]. Both datasets feature videos taken during the day. Due to the nature of the GIC algorithm, which makes video frames visible by altering the brightness of images, it is not necessary to apply this technique for daytime videos due to the inherent clarity of the footage. Therefore, Optical Flow information is used to provide motion information.

Under this configuration, LIFT-Net improves by 2.2% over the previous benchmark on HMDB51, achieving an accuracy of 89.92%. (refer to TABLE III). Furthermore, on the UCF101 dataset, LIFT-Net achieved 98.71% accuracy (refer to TABLE IV). Previously, the best performances on both datasets were obtained by BERT using an R(2+1)D model and TimeSFormer. Scores of 83.99% and 85.10% were attained on the HMDB51 dataset by BERT using RGB R(2+1)D (32f) and (64f), respectively, where 32f and 64f denote the frame lengths. These models were the second highest on UCF01 as well. By factorizing the spatial and temporal convolutions on

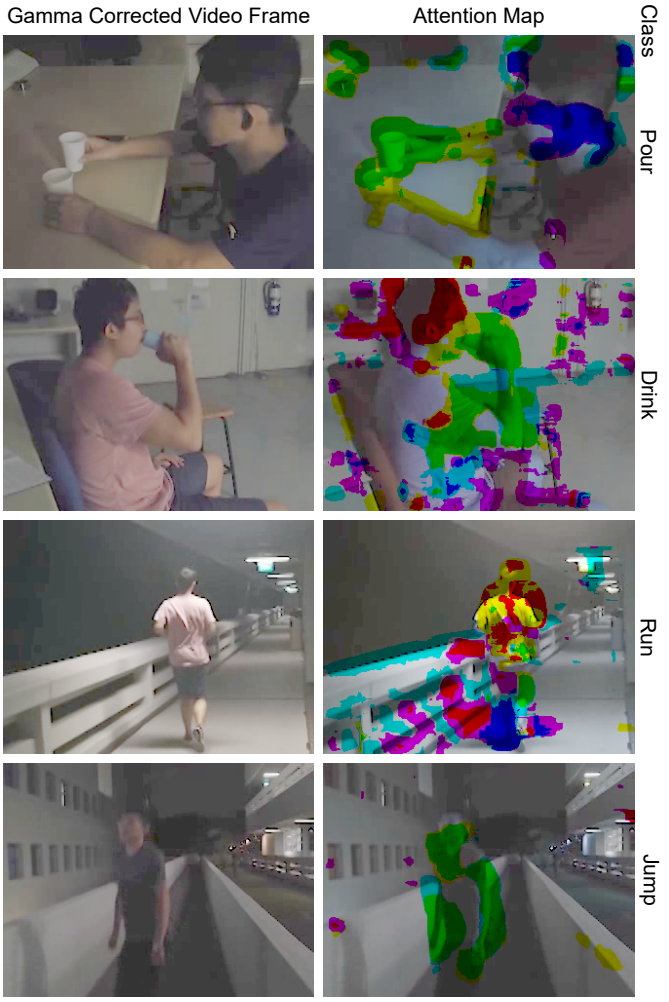

Fig. 5. The visualization of attention maps on each individual frame of the video. Column one shows the video frames after Gamma Intensity Correction (GIC) correction, while column two shows the attention map. Our model is trained to focus on actionable details in videos in order to identify what kinds of tasks are being carried out.

each layer, the R(2+1)D model doubles the number of non-linearities in the network, and by adding BERT, the network is able to achieve temporal embedding information, which in turn improves its performance. However, these models don't pay enough consideration to the crucial component within the frame. TimeSFormer utilizes a space-time-divided attention system that is based on a modified Transformer architecture. This model attained an accuracy of 87.72% and 98.77% on HMDB51 and UCF101 respectively, but, trails LIFT-Net by a substantial margin. Due to its incorporation of the idea of external visual attention, LIFT-Net's score was significantly higher than those of the other top three networks. In addition, the Feature Fusion module instructs the model on how to learn and recognize the crucial features in the frequency domain, allowing it to observe the tiniest of motion changes in the video with the aid of frames that have been transformed into frequency maps. This proves LIFT-Net's supremacy over the previous state-of-the-art methods. As far as we are aware, LIFT-Net is state-of-the-art on the HMDB51 and UCF101 datasets.

TABLE III
QUANTITATIVE RESULTS ON THE HMDB51 DATASET [3]. **BLUE** REPRESENTS THE PREVIOUS STATE-OF-THE-ART. **RED** DENOTES THE BEST RESULTS.

Model	Accuracy
IDT [47]	61.70%
Two-stream [18]	59.40%
Two-stream Fusion + IDT [48]	69.20%
ActionVald + IDT [20]	69.80%
TSN [49]	71.00%
RSTAN + IDT [50]	79.9%
TSM [51]	73.50%
R(2+1)D With Flow [26]	78.70%
I3D [22]	80.90%
MARS + RGB + Flow [52]	80.90%
FcF [53]	81.10%
ResNeXt-101 [38]	81.78%
EvaNet [54]	83.3%
HAF + BoW / FV halluc [55]	82.48%
ResNeXT101 + BERT [27]	83.55%
R(2+1)D Bert 32f pre trained on IG65M [27]	83.99%
R(2+1)D Bert 64f pre trained on IG65M [27]	85.1%
TimeSFormer [15]	**87.72%**
LIFT-Net (Ours)	**89.92%**

TABLE IV
QUANTITATIVE RESULTS ON THE UCF101 DATASET [4]. **BLUE** REPRESENTS THE PREVIOUS STATE-OF-THE-ART. **RED** DENOTES THE BEST RESULTS.

Model	Accuracy
Two-stream [18]	88.00%
Two-stream Fusion + IDT [48]	93.50%
ActionVald + IDT [20]	93.60%
TSN [49]	94.90%
RSTAN + IDT [50]	95.10%
TSM [51]	95.90%
R(2+1)D With Flow [26]	97.30%
I3D [22]	97.80%
MARS + RGB + Flow [52]	98.10%
ResNeXt-101 [38]	97.46%
ResNeXT101 + BERT [27]	97.87%
R(2+1)D Bert 32f pre trained on IG65M [27]	98.65%
R(2+1)D Bert 64f pre trained on IG65M [27]	98.69%
TimeSformer [15]	**98.77%**
LIFT-Net (Ours)	**98.89%**

3. Infrared Scenarios: We also test our model's capability on infrared videos. For the infrared setup, we provide Optical Flow information to the Luminous pathway. A comparison of our suggested method with other state-of-the-art approaches is presented in TABLE V. This comparison was performed on the InfAR dataset. Our proposed two-stream visual attention-based architecture surpasses all other existing models.

D. Visual Results

We apply LIFT-Net to video frames, and the resulting space-time attention visualizations are shown in Fig. 5. It is evident that LIFT-Net has a strong comprehension ability. It has the ability to concentrate on a specific area, making it easier to pinpoint the source of an event.

LIFT-Net's ability to accurately identify videos relies on its ability to pay attention to the intended areas of the frame, and attention maps provide evidence for it. For instance, if

TABLE V
QUANTITATIVE RESULTS ON THE INFAR DATASET [10]. **BLUE** REPRESENTS THE PREVIOUS STATE-OF-THE-ART. **RED** DENOTES THE BEST RESULTS.

Model	Accuracy
HOF [10]	68.58%
DT [47]	68.66%
iDT [47]	71.83%
Two-Stream 2D CNN [10]	76.66%
Two-Stream 3D CNN [56]	77.5%
TSTDD [31]	79.25%
Deep Residual Infrared Action Recognition [11]	83.50%
TimeSformer [15]	90.11%
R(2+1)D [27]	**91.22%**
Ours	**93.87%**

we look at Fig. 5, in the first row, we can see that the model concentrates on the hands of the person holding the cup, which enables the model to more accurately identify the 'pour' class.

VI. ABLATION STUDY

To advance our understanding of the contributions made by various components and to inform the design of the final architecture, we propose a series of experiments. In addition to being consistent with our final design, the findings we draw from these experiments provide significant insight. TABLE VI provides a summary of the outcomes of ablation studies conducted on the ARID dataset.

A. LIFT-Net with R3D Model

At first, we tried the R3D network rather than the R(2+1)D model. We note that the top-1 accuracy with the R3D model was 92.11, which is 3.1% less than with the R(2+1)D model. Despite employing 3D convolution and making efforts to maintain temporal dimension across the network, the R3D model still falls short in terms of model complexity. In comparison to R3D, the R(2+1)D model has greater complexity and nonlinearities, resulting in better results.

B. LIFT-Net with Optical Flow

We tested LIFT-Net using Optical Flow data since motion trajectories offer information that helps with recognition. In this configuration, Optical Flow data is fed into the Luminous pathway instead of gamma-corrected input frames. The rest of the model architecture and training process is unaltered. Compared to the proposed method, which makes use of gamma-corrected video frames, this arrangement yields significantly lower results: an accuracy of 87.22%. Optical Flow does not achieve better results because it cannot find motion when the frame rate is high, the video isn't stable, or there is noise in the video. However, in all of these circumstances, the gamma correction algorithm improves the visibility of the videos by changing the frame contrast, regardless of how quick the movement is or if there is any noise in the video.

TABLE VI
ABLATION STUDY

Model	Accuracy
LIFT-Net with R3D model	92.11%
LIFT-Net with Optical Flow	87.22%
LIFT-Net without KD approach	94.15%
Changing ViT blocks configuration	
ViT (5 block freeze)	94.92%
ViT (5 block unfreeze)	93.97%
Changing hyper-parameter N	
LIFT-Net - 4 (Small)	94.32%
LIFT-Net - 6 (Medium)	94.93%
LIFT-Net - 8 (Large)	95.21%

C. LIFT-Net without KD Approach

We also tried training our model without the KD method, meaning, the training is carried out without a teacher network. The top-1 accuracy of the model, in this case, was 94.15%. This experiment demonstrates that employing a teacher network to teach LIFT-Net (a student) improves the results.

D. Changing ViT Blocks Configuration

While training LIFT-Net, we experimented with varying the number of frozen and unfrozen blocks as well as the amount of pre-trained ViT blocks. With the identical training configuration, when the model was trained by leaving five ViT blocks unfrozen, $i.e.$, the weights of all five blocks were updated throughout training. An accuracy of 93.97% was achieved with this setup. Similarly, when the network is trained by freezing the weights of all blocks, we observe a 0.95% improvement in accuracy. Increment occurs because unfreezing all blocks implies the modification of the weights of pre-trained ViT blocks during back-propagation. In such a scenario, the network requires a substantial amount of time and data to learn. Nonetheless, we found that keeping only the first block unfrozen during training is the optimal setup when dealing with a small dataset.

E. Changing Hyper-Parameter 'N'

In this study, we tested out three different versions of LIFT-Net - small, medium, and large versions. In small versions, we maintain the N value constant at 4. Similarly, it was 6 and 8 for the medium and large versions, respectively. This N determines how many times the fusion of feature vectors from two paths should be repeated. TABLE VI, showcases the outcomes for the small, medium, and large-scale versions. By a wide margin, the best performance comes from LIFT-Net - 8 (big version). This is due to the fact that network complexity increases with the increasing value of N. However, when compared to the prior benchmarks shown in TABLE II, it is notable that the small and medium versions of LIFT-Net likewise produce state-of-the-art results. This demonstrates LIFT-Net's capacity to distinguish actions in low-light settings.

VII. CONCLUSION

Low-light action recognition is a challenging but essential task that has been understudied to date. In this study, we

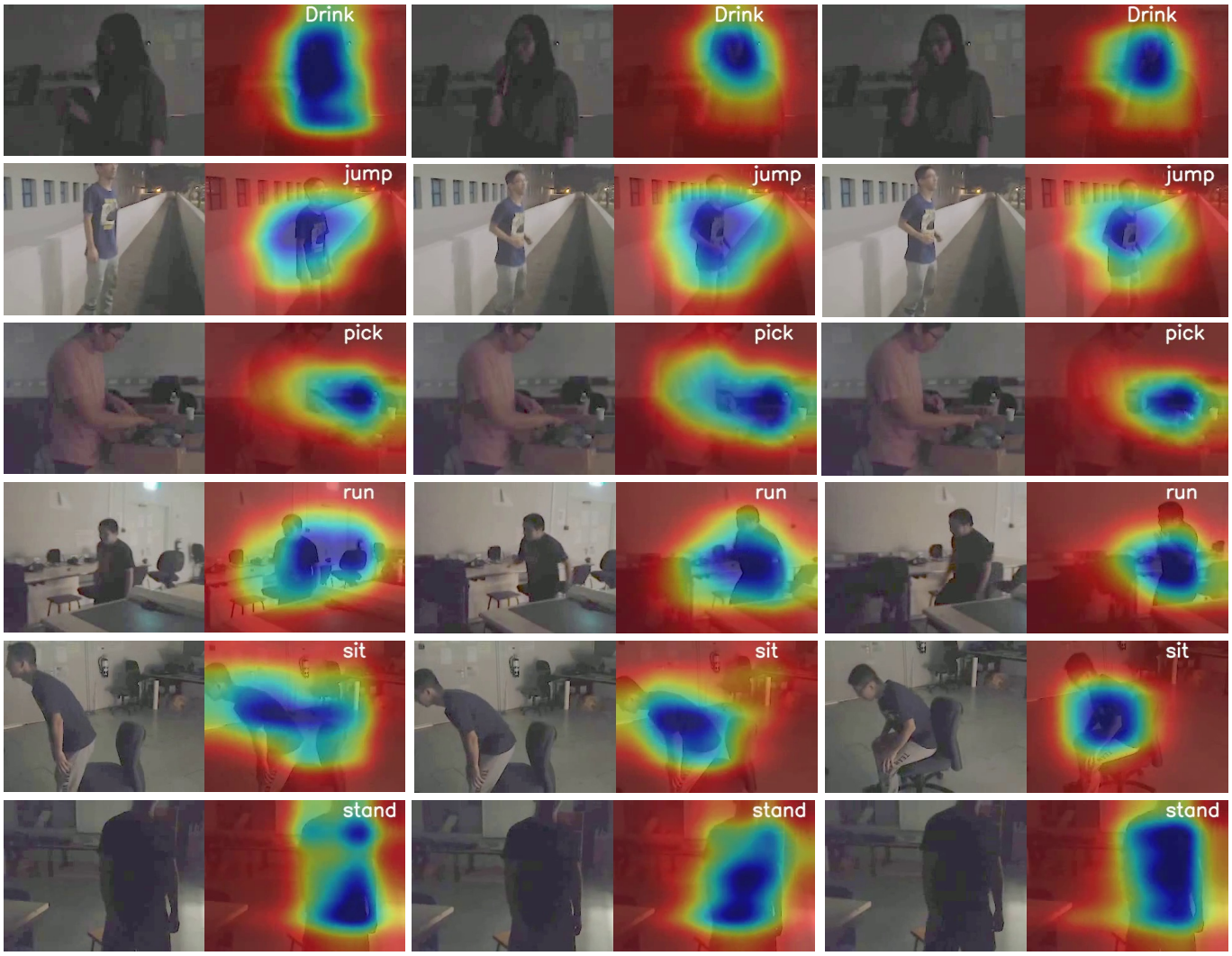

Fig. 6. An example of the class activation map that is produced by our model after the Feature Fusion block. The first section of each image displays the gamma-corrected input frames, while the second section displays the activation maps. On the activation map, the class of each frame is displayed in white. The links (https://youtu.be/0r9cihCMvUY and https://youtu.be/9n_LFvie0dg) provide demo videos that demonstrates the use of LIFT-Net for light-invariant action recognition in real-time on daytime, nighttime, and infrared videos.

introduced a model capable of detecting activity in a dark environment. Our method delivers state-of-the-art performance on the ARID dataset, which consists of dark videos. In addition, we examined the model's performance in both normal illuminations and infrared videos to confirm its robustness. LIFT-Net achieves state-of-the-art performance on the HMDB51, UCF101, and InfAR datasets under these illumination circumstances. Extensive experimentation and thorough ablation research further validate our methodology and demonstrate the efficacy of LIFT-Net with both qualitative and quantitative results. These findings confirm the significance of cross-visual attention and the merits of using feature vectors in the frequency domain to enhance learning. In the future, more light-weight Transformer networks leveraging features from the frequency domain could be explored for faster light invariant action recognition

REFERENCES

[1] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[2] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[4] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[5] 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.

[6] Y. Xu, J. Yang, H. Cao, J. Yin, Z. Chen, X. Li, Z. Li, and Q. Xu, "Going deeper into recognizing actions in dark environments: A comprehensive benchmark study," *arXiv preprint arXiv:2202.09545*, 2022.

[7] 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*, 2021, pp. 846–852.

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

[9] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "Arid: A comprehensive study on recognizing actions in the dark and a new benchmark dataset," *arXiv preprint arXiv:2006.03876*, 2020.

[10] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "Infar dataset: Infrared action recognition at different times," *Neurocomputing*, vol. 212, pp. 36–47, 2016.

[11] J. Imran and B. Raman, "Deep residual infrared action recognition by integrating local and global spatio-temporal cues," *Infrared Physics & Technology*, vol. 102, p. 103014, 2019.

[12] X. Chen, C. Gao, C. Li, Y. Yang, and D. Meng, "Infrared action detection in the dark via cross-stream attention mechanism," *IEEE Transactions on Multimedia*, vol. 24, pp. 288–300, 2021.

[13] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[14] 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*, 2015, pp. 4489–4497.

[15] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[16] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe, "Vidtr: Video transformer without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 577–13 587.

[17] A. Khebli, H. Meglouli, L. Bentabet, and M. Airouche, "A new technique based on 3d convolutional neural networks and filtering optical flow maps for action classification in infrared video," *Journal of Control Engineering and Applied Informatics*, vol. 21, pp. 43–50, 2019.

[18] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[19] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.

[20] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 971–980.

[21] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4768–4777.

[22] 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*, 2017, pp. 6299–6308.

[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 workshops*, 2017, pp. 3154–3160.

[24] A. Diba, V. Sharma, and L. Van Gool, "Deep temporal linear encoding networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2329–2338.

[25] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634–644, 2017.

[26] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[27] M. Kalfaoglu, S. Kalkan, and A. A. Alatan, "Late temporal modeling in 3d cnn architectures with bert for action recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 731–747.

[28] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3156–3165, 2021.

[29] M. S. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: What can 8 learned tokens do for images and videos?" *arXiv preprint arXiv:2106.11297*, 2021.

[30] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.

[31] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Global temporal representation based cnns for infrared action recognition," *IEEE Signal Processing Letters*, vol. 25, no. 6, pp. 848–852, 2018.

[32] S. Lamghari, G.-A. Bilodeau, and N. Saunier, "Actar: Actor-driven pose embeddings for video action recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 398–407, 2022.

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

[34] Z. Chen, "Z-domain entropy adaptable flex for semi-supervised action recognition in the dark," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4258–4265, 2022.

[35] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[36] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 525–542.

[37] X. Li, B. Shuai, and J. Tighe, "Directional temporal modeling for action recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 275–291.

[38] 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.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[41] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 244–253.

[42] Z. Liang, J. Chen, R. Chen, B. Zheng, M. Zhou, H. Gao, and S. Lin, "Domain adaptable normalization for semi-supervised action recognition in the dark," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4250–4257, 2022.

[43] Y. Xu, J. Yang, H. Cao, Q. Li, K. Mao, and Z. Chen, "Partial video domain adaptation with partial adversarial temporal attentive network," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9312–9321, 2021.

[44] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[46] 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. 0–0.

[47] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[48] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[49] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[50] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347–1360, 2017.

[51] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[52] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7882–7891.

[53] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9945–9953.

[54] A. Piergiovanni, A. Angelova, A. Toshev, and M. S. Ryoo, "Evolving space-time neural architectures for videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1793–1802.

[55] L. Wang, P. Koniusz, and D. Q. Huynh, "Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8698–8708.

[56] Z. Jiang, V. Rozgic, and S. Adali, "Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 115–123.