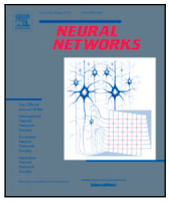
Neural Networks 179 (2024) 106622

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Dark-DSAR: Lightweight one-step pipeline for action recognition in dark videos

Yuwei Yin [a,1], Miao Liu [b,1], Renjie Yang [c,1], Yuanzhong Liu [d], Zhigang Tu [a,*]

[a] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[b] The Department of Pediatrics, Renmin Hospital, Wuhan University, Wuhan, China
[c] The Department of Radiology, Renmin Hospital, Wuhan University, Wuhan, China
[d] Kargobot (Beijing) Technology, Beijing, China

ARTICLE INFO

Keywords:
Action recognition
Dark video
Domain adaption

ABSTRACT

Dark video human action recognition has a wide range of applications in the real world. General action recognition methods focus on the actor or the action itself, ignoring the dark scene where the action happens, resulting in unsatisfied accuracy in recognition. For dark scenes, the existing two-step action recognition methods are stage complex due to introducing additional augmentation steps, and the one-step pipeline method is not lightweight enough. To address these issues, a one-step Transformer-based method named Dark Domain Shift for Action Recognition (Dark-DSAR) is proposed in this paper, which integrates the tasks of domain migration and classification into a single step and enhances the model's functional coherence with respect to these two tasks, making our Dark-DSAR has low computation but high accuracy. Specifically, the domain shift module (DSM) achieves domain adaption from dark to bright to reduce the number of parameters and the computational cost. Besides, we explore the matching relationship between the input video size and the model, which can further optimize the inference efficiency by removing the redundant information in videos through spatial resolution dropping. Extensive experiments have been conducted on the datasets of ARID1.5, HMDB51-Dark, and UAV-human-night. Results show that the proposed Dark-DSAR obtains the best Top-1 accuracy on ARID1.5 with 89.49%, which is 2.56% higher than the state-of-the-art method, 67.13% and 61.9% on HMDB51-Dark and UAV-human-night, respectively. In addition, ablation experiments reveal that the action classifiers can gain ≥1% in accuracy compared to the original model when equipped with our DSM.

1. Introduction

Currently, on the normal action recognition datasets, such as Kinetics (Kay et al., 2017), HMDB51 (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011), etc., action recognition methods (Tu et al., 2018, Gao, Du, & Yang, 2023, Tu, Xie, Dauwels, Li, & Yuan, 2019, Li, Wu et al., 2022, Liu, Yuan & Tu, 2022) have achieved outstanding performance. It is mainly due to most of the videos in these datasets being shot under sufficient lighting conditions, and constrained by the single shooting environment. However, on challenging dark datasets like ARID (Xu et al., 2021), their performance is unsatisfactory. The previous approaches focused more on video action recognition task with adequate illumination, but there are much fewer studies investigated on dark or unbalanced illumination conditions. In brief, videos shot in dark scenes are uneven in color and low quality in brightness. One of the main challenges for dark video action recognition is the migration of training data from the bright domain to the dark domain in a coherent end-to-end learning manner efficiently.

Many efforts have been made to handle this challenge. The methods can be divided into two types based on the model pipeline, *i.e.* two-step and one-step, see Fig. 1 for inference. The two-step methods perform the dark video action recognition task by (1) locally training an image enhancement model offline to enhance each video frame one by one to get the new video data in the bright domain, (2) then using the new enhanced video data to train an action recognition network, individually. Usually, this kind of approach contains the problems of (a) The light domain shift and the action recognition step are trained separately rather than jointly, which disrupts the coherent consistency of the model's functionality. For example, Singh, Suman, Subudhi, Jakhetiya, and Ghosh (2022) first visually enhances all videos in the dataset via utilizing an image enhancement model (Guo et al., 2020) and min–max temporal sampling, and then feeds the enhanced data into a classifier

* Corresponding author.
E-mail address: tuzhigang@whu.edu.cn (Z. Tu).
[1] Yuwei Yin, Miao Liu and Renjie Yang contribute equally, they are co-first authors.

https://doi.org/10.1016/j.neunet.2024.106622
Received 15 November 2023; Received in revised form 29 April 2024; Accepted 7 August 2024
Available online 8 August 2024

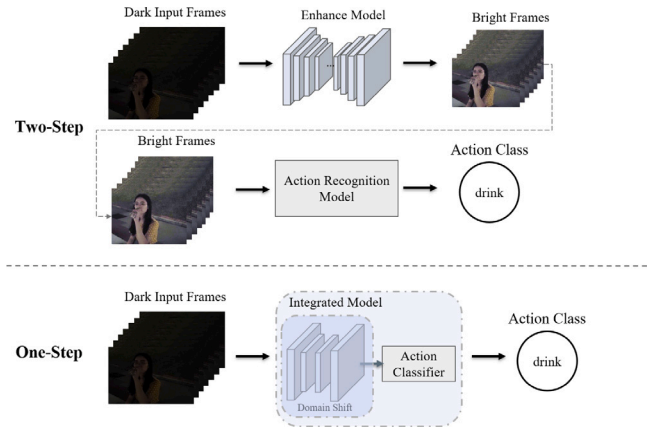


Fig. 1. Two-step methods involve separately training two models for action recognition: an image enhancement model for acquiring bright domain video data and an action recognition model for action classification. This dual-model training is time-consuming and disrupts the cohesiveness of action recognition in dark environments. In contrast, by eliminating the redundant stage of acquiring bright domain data and training a combined model for domain migration and action classification in one-step, the explored one-step method can effectively boost the inference efficiency and recognition accuracy.

with multiple backbone fusion for action recognition. Chen, Chen, Liang, Gao, and Lin (2021) obtains the luminance-enhanced videos through simple gamma correction, and then extracts the features from the dark and bright video data in two separate paths to finish the classification task. The problem of these methods is that the separating training manner decreases the performance of dark video action recognition. (b) The domain shift part of offline training brings additional training time and cost. Some methods try to enhance the original data through target domain data augmentation directly. For instance, Gao, Guo, Wang, and Zhang (2022) applies an image translation module to synthesize new data to reinforce training. Shao, Li, Ren, Gao, and Sang (2020) supplements the synthesized data of the target domain for network input, and increases the amount of information carried by the input to implement image dehazing. Domain alignment on feature space is also a solution. *E.g.*, Liang et al. (2022) constrains the feature space distribution difference between the original domain and the target domain, then minimizing the difference to align different domains. The above-mentioned methods require a large inference time due to synthesizing the target domain data and training with paired data.

The one-step methods handle the challenge by training a whole framework, connecting domain migration and action recognition in a single pipeline, and removing the offline training phase compared to the two-step. Some of the existing methods realize the one-step pipeline through the two-stream strategy, which merges the frame-enhanced part and uses the bright-dark paired data for training, leading to increasing the model's training complexity. For instance, Suman, Naharas, Subudhi, and Jakhetiya (2023) designs a two-stream network that includes a dark stream and a light stream, and the light stream utilizes an image enhanced model to get brighter video frames. But the concatenation of dark and light features increases the training cost. Liang et al. (2022) trains the model in a semi-supervised way and focuses on domain adaptation, where it aligns different dataset domains through feature normalization. However, Motiian, Piccirilli, Adjeroh, and Doretto (2017) requires a large normal light dataset as the source domain, which significantly boosts the complexity of the model. Some other methods discard the two-stream network, like (Tu, Liu, Zhang, Mu & Yuan, 2023), cascading the dark enhancement module and the action classifier, but the enhancement module takes unnecessary light enhancement iterations which causes the inference efficiency to slow. In summary, the one-step method addresses the issue of spatiotemporal inconsistency in (a) (b), but the problem of low inference speed is unsolved.

Motivated by these problems of the existing action recognition methods in solving domain adaptation (dark domain to bright domain), we propose a one-step Transformer-based dark domain shift for action recognition (Dark-DSAR) in dark scenes. The proposed Dark-DSAR is able to conserve the model's coherence in domain migration from bright to dark and action classification, accordingly improving the recognition accuracy and inference speed. Particularly, our Dark-DSAR, which cascades the Transformer-based action classifier MViTv2 (Li, Wu et al., 2022) as the backbone, exploiting symmetric CNN pairs based on Zero-DCE++ (Li, Guo, Loy & Change, 2022) to realize the illumination domain shift, *i.e.* the domain shift module (DSM). In short, DSM first estimates a set of 3-channel enhancement curve parameters that correspond to the R, G, B channel of the frame respectively, and then performs light supplement for each pixel of the frame with these parameters. The step for every pixel is repeated at certain times to obtain good enhancement, and the curve parameters are shared during the whole process to reduce the number of model parameters.

In addition to finding the optimal settings for balancing recognition accuracy and inference speed, we investigated the impact of the spatial resolution compression ratio and the structural setting of the domain shift module (DSM) on the model performance. Remarkably, we found that for the dark action recognition task, high-resolution video is not necessary. To streamline the model to reduce the number of parameters and speed up the inference, we filter the raw input information by resolution compression, which is beneficial for improving efficiency. Moreover, the adopted Zero-DCE++ (Li, Guo et al., 2022) further reduces the computation by sharing the iterative parameters of the enhancement curves. Extensive experiments have been done to verify the presented model Dark-DSAR, and we also explored the factors that affect its performance.

The contributions are summarized as follows.

- A cross-domain one-step dark video action recognition method named Dark-DSAR is proposed, which integrates the illumination domain adaption and action classification coherently into a single stage all-in-one task without costly redundant training consumption.
- A domain shift module (DSM) and a spatial resolution compression phase based on ResKD (Ma et al., 2022) are designed to reduce the computation cost and network parameters of the proposed model, achieving a lightweight and efficient video enhancement process. Besides, the effect of video resolution on action recognition accuracy in the low-light environment is explored.
- Extensive experiments are conducted to test the performance of our model Dark-DSAR, where the results demonstrate that Dark-DSAR obtains superior performance on the task of dark video action recognition in both accuracy and efficiency.

2. Related works

2.1. Action recognition in the dark video

Most video action recognition algorithms focus on good lighting conditions. Performing action recognition in dark scenes is challenging, and some researchers have started to investigate this issue in recent years. Ulhaq (2018) fusing multiple video stream deep features corresponding to multiple spectra is used to combine visible light sequences with infrared data features in night vision scenes to improve the accuracy of action recognition in dark environments, and some similar works have been investigated (Akula, Shah, & Ghosh, 2018; Anwaar-ul-Haq, Gondal, & Murshed, 2011; Eum, Lee, Yoon, & Park, 2013; Zhang et al., 2022). These studies assisted action recognition in low-quality video with the help of sensor information other than RGB video. These approaches use additional data which is difficult to obtain significantly

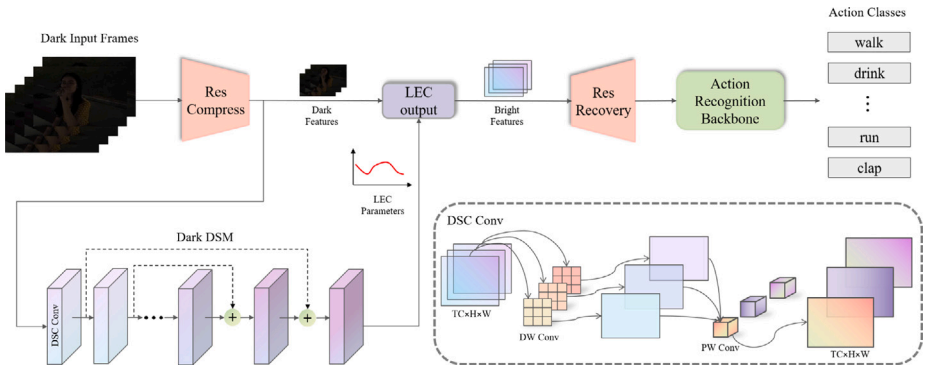


Fig. 2. Overview the structure of the proposed Dark Domain Shift for Action Recognition model Dark-DSAR, comprising three main parts: video resolution compression, dark domain shift, and action recognition. These three components are sequentially connected within a coherent stage and trained together as a single entity. The Res Compress component effectively eliminates the redundant spatial information from the input. The Dark Domain Shift Module (Dark DSM) facilitates the transformation of dark videos into bright domain by learning parameters associated with the lowlight enhancement curve (LEC). The action recognition backbone performs the task of action classification. A schematic of the Depthwise Separable Convolution (DSC Conv) is provided in the lower right corner.

increasing the complexity of the process. Some utilize input data from different modalities to enhance recognition accuracy, *e.g.* Tu, Zhang, Li, Chen and Yuan (2023), Zhang, Jia, Xie and Tu (2022) utilize skeleton information, Tu et al. (2019) employing semantic information to assist convolutional neural networks for extracting action features. There are also ways to handle RGB video directly, Hira, Das, Modi, and Pakhomov (2021) uses image enhancement to first brighten the original domain data, and then incrementally samples the enhanced image as the output of the action recognition network. Hussain et al. (2023) applies a lightweight pyramid network for dark video frame enhancement, and the action features are extracted and classified in the cloud after completing the pre-processing. Singh et al. (2022) enhances the dark video by the IEM module, and then feeds the enhanced video frames into the action classification network ACN for action recognition. The above methods reduce the requirement of raw data, but cannot achieve end-to-end model training. The illumination enhancement and action recognition modules are not a whole and cannot be trained jointly.

Tu, Liu et al. (2023) introduced a novel method for joint light enhancement and action classification in dark action recognition. However, this method employs an enhancement model based on depthwise convolution layers (DW Conv), utilizing 32 channels. Each channel is convolved individually to extract features, and the output is obtained by shuffling the channels. This approach leads to computational and parameter redundancy. The redundancy arises because, firstly, the luminance enhancement of the RGB video does not require excessive channel parameters for enhancing each color channel individually, and high video resolution is not necessary for action classification. Secondly, the continuous enhancement process lacks parameter sharing, resulting in redundant parameter quantities. In contrast, our proposed method is different from those above in that it can be integrated into various backbones to be trained end-to-end, simultaneously addressing the issue of parameter and computational redundancy in Tu, Liu et al. (2023).

2.2. Lowlight image and video enhancement

Enhancement of image and video in low-light environments is beneficial for downstream tasks, *e.g.* classification, detection, semantic segmentation, etc. The existing typical low-light image and video enhancement methods can be divided into traditional methods (Ma, Wang, Zhang, & Zhang, 2023) and deep-learning-based methods. The low-light image enhancement method based on histogram equalization (Pizer, Amburn, Austin, Cromartie, Geselowitz, Greer, ter Haar Romeny, Zimmerman, & Zuiderveld, 1987) and Retinex theory (Land & McCann, 1971) belongs to the traditional method. Histogram equalization is easy to calculate but has low robustness. The method based on Retinex theory needs to consider the illumination and reflection conditions, and has a high computational cost (Li et al., 2022).

Deep-learning-based light enhancement methods can be classified as CNN-based, GAN-based, and Transformer-based according to the backbones they used. CNN-based methods *e.g.*, Wei, Wang, Yang, and Liu (2018) and Lin, Zhang, Wang, and Wang (2023) are based on the Retinex theory (Land & McCann, 1971), which decomposed an image into reflectance and illuminance images by a decomposition network, and fed the decomposition to a light adjustment network. The model is trained using low-light normal-light images. Lv, Lu, Wu, and Lim (2018) enhanced the image features extracted by convolutional layers of different depths separately, and finally fused the enhanced results. Zero-DCE (Guo et al., 2020) is a CNN-based zero-shot light enhancement method that employs a U-Net (Ronneberger, Fischer, & Brox, 2015)-like structure to extract features at different scales and obtains pixel-level exponential parameters to output enhanced images after multiple iterations. Zheng, Li, Yang, and Wu (2021) brightened single images by combining data-driven and model-driven methods rather than employing complex neural networks. The GAN-based method, *e.g.* Jiang et al. (2021) is the first approach to achieve light enhancement without using paired images, which applied U-Net (Ronneberger et al., 2015) as the generator backbone and global-local discriminator as the training guide, and also introduced the self-regularized attention mechanism. Recently Transformer has become prevalent and been applied to various tasks in Computer Vision. The Transformer-based method (Cui et al., 2022; Xu, Wang, Fu, & Jia, 2022), like IPT (Cui et al., 2022) used Transformer to learn the mapping relations among image features to obtain images with a good visual experience. However, these methods usually cost heavily in computation and storage.

2.3. Domain shift and domain adaption

Out of the need to improve learning efficiency, when solving some similar or identical recognition problems, a robust model is expected. Domain adaptation refers to maintaining the performance of the model trained on the original domain when transferred to the target domain. To maintain the performance of the model, we must try to reduce the difference between the original domain and the target domain, and knowledge distillation (Hinton, Vinyals, & Dean, 2015) is a kind of solution. Wang and Deng (2018) divided the domain adaptation problem into two types, *i.e.* one-step domain adaption and multi-step domain adaption. The problem from the bright domain to the dark domain we focus on belongs to the first type, which means the original domain and the target domain are similar or related, no need to establish an intermediate domain, and domain adaption can be completed in one step.

The one-step domain adaption approach can be classified into discrepancy-based, adversarial-based, and reconstruction-based categories (Wang & Deng, 2018). Discrepancy-based (Cai, Wang, He, & Zhou, 2020; Peng, Hoffman, Yu, & Saenko, 2016) directly uses labeled or unlabeled data on the target domain to train the model, and fine-tunes the network parameters of the existing model to reduce the domain bias. The Adversarial-based Ganin and Lempitsky (2015), Tzeng, Hoffman, Darrell, and Saenko (2015) and Shen, Pan, Choi, and Zhou (2023) introduces GAN (Goodfellow et al., 2014)-like domain discriminators and uses the adversarial loss to confound the decisions of the domain discriminators. Reconstruction-based (Kim, Cha, Kim, Lee, & Kim, 2017) implements domain adaption with the aid of data reconstruction in the original or the target domain.

3. Methods

In this section, we describe the proposed model Dark-DSAR in detail, where Fig. 2 shows its overall framework. Specifically, our Dark-DSAR is mainly composed of the explored plug component DSM for light-domain migration and the action classifier MViTv2 (Li, Wu et al., 2022). Inspired by Li, Guo et al. (2022), we construct a domain shift convolutional network between the input video and the classifier to convert the dark domain to the bright domain by learning enhanced curve parameter features meanwhile reducing the computation redundancy.

3.1. Motivations

For dark video human action recognition, we aim to maintain the performance of the model, which is trained under natural light conditions, when encountering illumination change. In particular, we want to address two important issues that the existing methods face.

Complex domain shift stage. The current popular two-step approaches, such as knowledge distillation based (Tu, Liu, & Xiao, 2022) and image enhancement based, split dark video human action recognition into two discrete phases. This manner increases the difficulty and complexity of model training, where a separate teacher model or image enhancement model needs to be trained in advance offline.

Redundant computation cost. This problem exists in both the two-step and one-step methods. For the two-step method, the cost refers to the additional training stage and the increasing inference time that comes with it. For the one-step method, the cost mainly comes from the enhancement module structure, for example, the multiple enhancement curve iterations in the enhanced network.

As shown in Fig. 1, to solve these two problems, a Dark-DSAR model is presented, which has two major innovations: (1) Simplifying the complicated two-step pipeline into a coherent one-step pipeline. We connect domain migration and action classification as a whole by flattening the time dimension, which has the benefit of preserving the coherence of model training to boost recognition accuracy. (2) Optimizing the computing and inferencing procedure jointly, and we reduce redundant computation by spatial resolution compression and light-enhanced network pruning. Details are described below.

3.2. One-step coherence preservation

The key to protecting the coherence of the model lies in the adding time dimension of the video task compared with the image task. Due to this, some relevant assumptions for video tasks are necessary.

Temporal consistency (Tu, Liu et al., 2023). The human action scene of the video subject in a single clip is almost constant, so the magnitude change between the background and the identified objects along adjacent frames is tiny. Meanwhile, the lighting change in the shooting environment is constrained to the scene, therefore the enhancement process remains consistent in time, and the output domain of the same clip is smooth and successive. *Noise artifact*. The movement of video objects over time in space we consider to be relatively small and slow, thus only the noise from lighting conditions is taken into account in the enhancement process, without focusing on the problems *e.g.* motion blur and motion artifacts that are associated with high-speed motion.

The input RGB dark video sequence is defined as:

$$I(x) \in \mathbb{R}^{3\times T\times H\times W}, \tag{1}$$

x denotes the pixel of one video frame, T is the number of frames in each clip, H is the height and W is the width of the video. Unlike training an additional enhancement module locally, which disrupts models' functional succession and is harmful to video action recognition in the dark environment, we overcome this drawback by taking the temporal feature into account. Following the prior work Singh et al. (2022), Tu, Liu et al. (2023), instead of downsampling along the temporal dimension, *i.e.*, sampling a certain number of frames for each clip, we concatenate the temporal dimension of the video to its RGB color representation. $I(x)$ is updated to $I'(x) \in \mathbb{R}^{C\times H\times W}$, where C denotes the channel after being flattened. At this point, $I'(x)$ can be integrated with other backbone networks related to video processing for joint training, which reduces the training complexity and maintains a one-step manner. $I'(x)$ is then compressed in the spatial domain and feds into the domain shift module for extracting the enhancement curve parameter features (Guo et al., 2020) and performing domain migration.

3.3. Dark domain shift

We propose a dark domain shift component named DSM to implement the domain migration from dark to bright. We find that the computation cost in Tu, Liu et al. (2023) increases because of layer-by-layer and pixel-by-pixel convolution with a large number of video frames, which is detrimental to training efficiency. To solve this problem, we make some improvements to construct a faster and lighter-weight structure. Specifically, inspired by the work (Li, Guo et al., 2022), we replace DW Conv with the lighter Depthwise Separable Convolution (DSC Conv) (Chollet, 2017). Fig. 3 illustrates the schematic of the convolution process. Given an input feature map $f \in \mathbb{R}^{C\times h\times w}$, firstly, we perform DW Conv, which uses a $3 \times 3 \times 1$ convolution kernel for each channel cc to generate a feature map with the same shape as the input. Secondly, a pointwise convolution (PW Conv) operation is applied by using a $[1 \times 1\times C\times N]$ convolution kernel to fuse different channel feature maps on top of the DW Conv output. The final feature map has the shape $[N, h, w]$. Finally, the obtained map is passed through the ReLU (Glorot, Bordes, & Bengio, 2011) activation layer. DSC Conv eliminates the convolution process of each kernel across multiple channels, thereby reducing the number of parameters and computation. Besides, this operation also mitigates the risk of overfitting.

DSM consists of 7 DSC conv, and its architecture is shown in Fig. 3. We modify the Guo et al. (2020) component by adjusting the scale factor at the front of the network to achieve a better match between the input and the network. DSM has symmetric CNNs similar to the U-Net (Ronneberger et al., 2015) structure. To endow the model with multi-scale perceptual capability, the feature maps from symmetrically located positions are concatenated and used as the input for the next layer, *e.g.* DSC Conv1 & DSC Conv6, DSC Conv2 & DSC Conv5, DSC Conv3 & DSC Conv4. The output of DSC Conv7 is $P(x)$ which has a shape of $[3, h, w]$, ranging from -1 to 1. $P(x)$ is the parameter of the pixel-level enhancement curve. As defined in (2). α is the scale factor. i denotes the horizontal position of the pixel, j denotes the horizontal position of the pixel, and $color$ indicates the color channel of the pixel.

$$\underset{0<i<h,0<j<w,color\in R,G,B}{P\left(x_{i,j,color}\right)} = \mathrm{DSM}\left(\alpha I\left(x_{(i,j,color)}\right)\right), \tag{2}$$

In the process of light domain adaptation, the goal is to convert darker scenes into brighter ones. To achieve this, we must compress

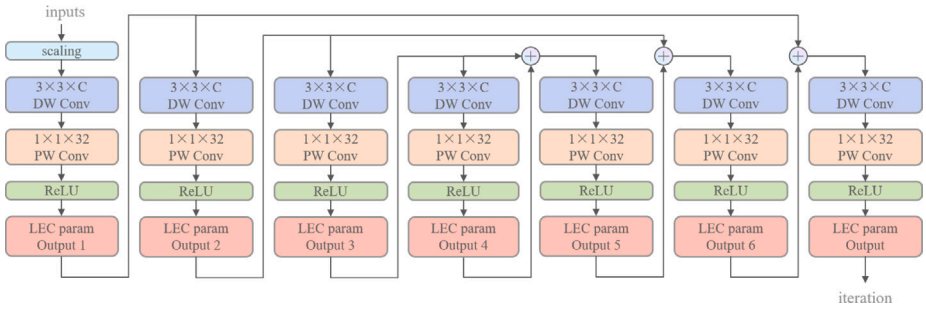


Fig. 3. Architecture of the Dark Domain Shift Module. "scaling" indicates the input resizing operation. C represents the number of channels. DW Conv and PW Conv denote depthwise convolution and pointwise convolution, correspondingly. "LEC param output*" pertains to the output of the internal conv. layer.

the color space with higher color values and extend the color space with lower color values. During this process, the spatial contrast for higher color values decreases while the spatial contrast for lower color values increases. This results in an overall increase in image brightness, completing the brightness conversion. To fulfill this criterion, the transformation function should mimic a gamma transform with a gamma value less than 1 while also avoiding the introduction of undue computational complexity. Therefore, a direct application of the exponential curve is not suitable. Taking all these considerations into account, the quadratic parabolic curve is a suitable choice. It is easier to differentiate compared to the exponential transformation, involves lower computational requirements, and exhibits a growth pattern similar to the exponential transformation, thus ensuring it achieves the desired brightness transformation.

With $P(x)$, we enhance $I'(x)$ through a quadratic curve represented by Eq. (3):

$$E(x) = I'(x) + P(x)I'(x)\left(1 - I'(x)\right), \tag{3}$$

where $E(x)$ refers to the enhanced result after one enhancement iteration. The quadratic curve employed is differentiable, and the enhancement process involves pixel-wise operations on the feature maps. To achieve more satisfactory luminance enhancement, we conduct multiple iterations on the feature maps, utilizing higher-order curves based on Eq. (3). Higher-order curves provide a superior representation of detailed texture features when compared to lower-order curves, leading to improved detail estimation and reconstruction. Denote the number of iterations as 'n'. Consequently, the result of the nth iteration for enhancing the feature map with higher-order curves from the $n-1^{th}$ can be expressed as Eq. (4):

$$E^n(x) = E^{n-1}(x) + P(x)E^{n-1}(x)\left(1 - E^{n-1}(x)\right). \tag{4}$$

The number of iterations influences the model's performance. We have ultimately chosen to set the iteration parameter at 8. This choice satisfies the need for illumination supplement in predominantly dark scenes while maintaining a balance between computational complexity and enhancement effectiveness. A more detailed explanation of our choice of iteration parameters will be provided in the experimental section.

It is important to note that our approach differs from the estimation curves used in the enhancement model of Tu, Liu et al. (2023). We employ the exact same pixel-by-pixel enhancement parameter, denoted as $P(x)$, for each iteration, which means that, as explained earlier the DSM outputs a 3-channel (for R, G, B) enhancement curve parameter. In contrast, Tu, Liu et al. (2023) generates 3-channel (totaling $3N$ channels) enhancement parameters for every iteration, which possesses three times as many parameters as our model. Our approach maintains a lighter model with fewer parameters while avoiding significant degradation in performance.

During the backpropagation training of the network, we consider constraints related to the exposure intensity, the RGB color offset, and the monotonic relationship between neighboring pixels. Specifically, we impose these constraints on the DSM using the complete loss function proposed in Guo et al. (2020). The notation for this loss function Eq. (5) is defined as L_{DSM}. W_{exp}, W_{tv} and W_{color} are set to 1, 20, and 0.5.

$$L_{DSM} = W_{\text{exp}}L_{\text{exp}} + W_{tv}L_{tv} + W_{color}L_{color} \tag{5}$$

The exposure loss, represented by L_{exp}, quantifies the difference between the average intensity of a local area and the intensity of a properly exposed region. Partition each enhanced image into some 16×16 regions. Use Y_i to represent the average grayscale value of the ith region, and E to denote the value of appropriate exposure intensity (set at 0.6 same as Guo et al. (2020)), and K is the number of regions, the loss L_{exp} can be expressed as:

$$L_{\text{exp}} = \frac{1}{k}\sum_{i=1}^{K}\left|Y_i - E\right|, \tag{6}$$

The illumination smoothing loss, L_{tv}, promotes uniform lighting across the image. N is the total iterations of DSM, Ω is the color channel space, P_n^c is the enhancement parameter of the cth color channel obtained in the nth iteration, ∇_x and ∇_y are the horizontal and vertical gradient respectively:

$$L_{tv} = \sum_{n=1}^{N}\sum_{c\in\Omega}\left(|\nabla_x P_n^c| + |\nabla_y P_n^c|\right)^2, \qquad \Omega = \{R, G, B\}, \tag{7}$$

On the other side, the RGB color correction loss, L_{color}, addresses color discrepancies by maintain the gray value consistency of different channels:

$$L_{\text{color}} = \left(J^R - J^G\right)^2 + \left(J^R - J^B\right)^2 + \left(J^B - J^G\right)^2, \tag{8}$$

J is the average intensity of the enhanced result of three channels (R, G, B channel).

3.4. Action classification

The proposed module DSM can be integrated into any action classifier for training coherency. Here we utilize MViTv2 (Li, Wu et al., 2022) as the action classifier baseline. MViTv2 is a video processing Transformer (Vaswani et al., 2017) backbone with multi-scale feature stages, that enable spatio-temporal recognition. By incorporating a multiscale feature pyramid network, cross-attention mechanism, and class-wise attention mechanism, MViTv2 effectively captures the spatial and temporal information in videos, making it achieves state-of-the-art performance for action recognition. The structure of MViTv2 base (MViTv2-B) is analyzed in Table 1 for reference.

3.5. Resolution compression

In this subsection, we present a qualitative analysis of the relationship between video resolution and action recognition accuracy. Low-quality videos or figures have a negative impact on the accuracy

Table 1
Structure of the MViTv2-B backbone. MHPA refers to the Multi Head Pooling Attention. MLP denotes the multi-layer perception. "Head Nums" designates the heads in every basic block for the four stages.

Stage	Layers	Head nums	Output size
input	stride4 × 1 × 1	–	3 × 16 × 224 × 224
cube1	3 × 7 × 7, 96 stride 2 × 4 × 4	–	96 × 8 × 56 × 56
sacle2	$\begin{bmatrix} MHPA(192) \\ MLP(768) \end{bmatrix} \times 2$	1	96 × 8 × 56 × 56
scale3	$\begin{bmatrix} MHPA(192) \\ MLP(768) \end{bmatrix} \times 3$	2	192 × 8 × 28 × 28
scale4	$\begin{bmatrix} MHPA(384) \\ MLP(1536) \end{bmatrix} \times 16$	4	384 × 8 × 14 × 14
scale5	$\begin{bmatrix} MHPA(768) \\ MLP(3072) \end{bmatrix} \times 3$	8	768 × 8 × 7 × 7

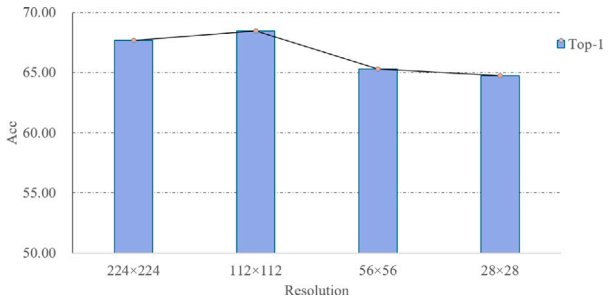


Fig. 4. Accuracy of Video Swin Transformer on ARID1.5 with different resolution frames.

of the classification network (Ma et al., 2022) and may appear to be an intuitive cause of the accuracy decrease. However, this does not imply that low-resolution videos lack critical information for action recognition. To elucidate this, we conduct a quantitative analysis of video resolution in the input network.

Initially, the original video size is uniformly scaled to 224 × 224. Subsequently, downsampling is performed at various scales to obtain low-resolution inputs, *i.e.* $112 \times 112, 56 \times 56, 28 \times 28$. Similar to ResKD (Ma et al., 2022), we simulate the information loss process through downsampling and subsequently resize the obtained frames back to the size of 224 × 224 by using bilinear interpolation. This approach helps eliminate the influence of video scale changes on the results. Fig. 4 shows that the decreased resolution of the input video does not lead to a decline in recognition accuracy (resolution 112 × 112). This observation suggests that the information conveyed by the low-resolution video is adequate for the classification task.

We optimize the proposed cascade network based on this observation. In particular, firstly, we introduce a downsampling stage at the network input to eliminate redundant information presented in the video. Then, the video is passed through the domain migration module to enhance its illumination before being fed into the classifier. Finally, the video is upsampled back to the original input resolution using double linear interpolation to preserve scale invariance.

Using $RDS()$ to denote the spatial resolution compression function and $R(x)$ to represent the result of domain migration. Eq. (1) is updated as Eq. (9):

$$I'(x) = RDS(I(x)) \tag{9}$$

The output of the DSM is represented as Eq. (10):

$$R(x) = \widetilde{RDS}\left(E^n(x)\right) \tag{10}$$

$R(x)$ as the feature map output of DSM, is sent as the input to the action classification network which completes the human action feature extraction and classification. The final recognition result will be output by the classification head.

3.6. Joint training loss

For action classification we choose the cross entropy loss and the formula for action classification loss L_{AR} is shown below:

$$L_{AR} = -\sum_{i=1}^{K} \left(y_i \log \hat{y}_i\right) \tag{11}$$

K is the number of action classes, y_i is the ground truth label, and the $\hat{y}_i$ is the prediction result of the action classifier. To achieve a better balance of different parts of the loss function, we modify the loss of action classification and the dark domain shift module loss during the joint training procedure. We assign distinct weight coefficient β to each part of the loss function based on their respective functions. The complete loss function is defined as Eq. (12):

$$L = \beta L_{AR} + (1 - \beta) L_{DSM} \tag{12}$$

4. Experiments

4.1. Experimental details and datasets

The proposed model is implemented using PyTorch and trained on 4 NVIDIA 2080Ti GPUs. We fixed the learning rate as $2.5e^{-4}$ and set the batch size to 4 throughout the training phase. We used AdamW (Loshchilov & Hutter, 2017) to optimize the learning process of our model. Unless otherwise specified, the network weights are pre-trained based on the kinetics dataset (Kay et al., 2017). We resize the training images to 224 × 224, and each training video frame undergoes a center crop to 256 × 256. We conduct extensive experiments on 3 representative action recognition datasets under dark conditions: ARID (Xu et al., 2021), UAV-human-night (Li et al., 2021), HMDB51-Dark. Top-1 accuracy and top-5 accuracy are used to evaluate the experimental results. We directly quote the recognition results on the relevant datasets from the original papers, and for those that lack experimental results, we reproduce the model and use * for distinguishing.

ARID (Xu et al., 2021) *dataset*. The ARID dataset is specially used to analyze human behavior in dark environments. It contains 11 common action classes, the video frame rate is 30 fps, and the resolution is 320 × 240. The videos are all shot at night, and the shooting scenes are divided into indoor and outdoor. The ARID dataset has two versions, ARID1.0 and ARID1.5. *ARID1.0* consists of 3784 video clips. The ratio of the training set and the testing set is 7:3, and there are three splits. *ARID1.5* is composed of 5572 video clips, each action class contains more than 320 clips, three indoor and outdoor shooting scenes are added, and there are two ways to divide the training and testing sets. Compared with large-scale action recognition datasets, the biggest difference of ARID is that all its videos are shot under low-light conditions, which brings great challenges to the action recognition task.

UAV-human (Li et al., 2021) *dataset*. UAV-human is a large-scale UAV (Unmanned Aerial Vehicle) human action analysis dataset, which can be used in action recognition, pose estimation, pedestrian re-identification, and other visual tasks. Its data is collected from three different sensors carried by UAV: Azure DK, fisheye camera, and night vision camera. Data from night vision is used for our experiments. The UAV-human-night contains 155 classes of human activities, the video frame rate is 25 fps, and the resolution is 640 × 480. There are 22,476 video clips in total, and two splits are provided.

Fig. 5. Comparative examples all datasets. For HMDB51, the original HMDB51 (up) and the synthesized HMDB51-Dark (down).

HMDB51 (Kuehne et al., 2011) *dataset*. HMDB51 is a small dataset for action recognition, which contains 51 classes of human activities. The videos are basically shot during the day or in an indoor room with sufficient room. There are 6849 video clips in total, with a resolution of 320 × 240, and the frame rate is 30 fps. It supplies three splits. Since the videos in HMDB51 are not under dark conditions, we correct the brightness of the videos according to the gamma correction principle of the image. We follow the method of Guo et al. (2020), synthesizing the HMDB51-Dark dataset, and the synthesized examples are shown in Fig. 5. Specifically, we synthesize the dark videos according to Eq. (13), where γ obeys a Gaussian distribution with a mean of 0.2 and a variance of 0.07 to simulate different lighting conditions.

$$f(I) = I^{\gamma} \tag{13}$$

4.2. Comparison with state-of-the-arts

In this section, the main experimental results are presented. We compared the classical action recognition backbones (Carreira & Zisserman, 2017; Feichtenhofer, Fan, Malik, & He, 2019; Hara, Kataoka, & Satoh, 2017; Lin, Gan, & Han, 2019; Tran et al., 2018; Wang et al., 2016), and some currently advanced action recognition backbones (Li, Wang et al., 2022; Li, Wu et al., 2022; Liu et al., 2022) on four datasets. In addition, we compared with the-state-of-the-art (SOTA) dark action recognition methods (Chen et al., 2021; Liang et al., 2022; Singh et al., 2022; Tu, Liu et al., 2023), of which the first three belong to the two-step pipeline, and Tu, Liu et al. (2023) belongs to the one-step pipeline.

Experiments on the ARID dataset. Table 2 shows our experimental results on ARID1.5. The results of 3D-ResNet-18, I3D-RGB, I3D Two-stream, R(2+1)D-GCN+BERT, and Darklight+R(2+1)D-34 are directly copied from the corresponding paper (Singh et al., 2022). The results of other models are replicated by us. We have compared 2D CNN-based (Lin et al., 2019; Wang et al., 2016), 3D CNN-based (Carreira & Zisserman, 2017; Feichtenhofer et al., 2019; Hara et al., 2017; Tran et al., 2018), and Transformer based action recognition methods (Bertasius, Wang, & Torresani, 2021; Li, Wang et al., 2022; Li, Wu et al., 2022), and also the methods specifical for low-light environments (Chen et al., 2021; Singh et al., 2022). Our Dark-DSAR enhances the Top-1 accuracy by 2.56% (89.49% vs. 86.93%) when compared to the previous best-performing model (Singh et al., 2022). Compared with Darklight+R(2+1)D-34, our Dark-DSAR uses 76.3% parameters but gains 7.6% and 1.5% improvement on the Top-1 accuracy and Top-5 accuracy, respectively (89.49% vs. 84.13%, 98.77% vs. 97.34%). We have fine-tuned the reproduced models to enhance their performances. Due to the number of action classes in the ARID dataset are small, some CNN-based and Transformer-based models are not robust enough, and the performance is unsatisfied. Our Dark-DSAR surpasses TimeSformer by 44.84% (89.49% vs. 49.36%), and UniFormerV2 by 31.38% (89.49% vs. 61.41%). Meanwhile, our Dark-DSAR has fewer parameters (50.94M vs. 196M, 50.94M vs. 115M). Remarkably, on ARID1.5, our method achieves the Top-1 accuracy of 89.49%, outperforming the other models by at least 2%. Table 3 provides the results on the ARID1.0 dataset. The proposed method achieves an average Top-1 accuracy of 96.99% and Top-5 accuracy of 99.77% on the three splits. Compared to the existing SOTA action recognition method DTCM (Tu, Liu et al., 2023), the Top-1 accuracy of Dark-DSAR is improved by 0.63% (96.39% vs. 96.99%), reaching the highest. In particular, as shown in Table 4, compared to DTCM, we have reduced the plug-in memory consumption by nearly one-third (828M vs. 2073M), while maintaining a high recognition accuracy (96.99% vs. 96.36%).

Experiments on the UAV-human-night dataset. UAV-human-night includes some videos taken by night vision cameras during the daytime, resulting in the models' performance being less stable than those on ARID. The accuracy of the two splits varies significantly because of the different subset divisions. As shown in Table 5, on this dataset, the proposed Dark-DSAR still obtains the best performance on two splits, whose accuracy outperforms the other models by at least 1%. Taking the mean on these two splits, our Dark-DSAR achieves 57.16% Top-1 accuracy and 82.69% Top-5 accuracy respectively, 0.82% and 1.16% higher than the second place method MViTv2 (Li, Wu et al., 2022). The 2D model's poor performance is evident in Table 5. On split 1, our method shows a 21.63% improvement over TSM (the highest-performing 2D CNN-based model). Additionally, our Dark-DSAR increases the Top-1 accuracy by 0.95% compared to MViTv2-B (the top performer among the Transformer-based models). On split 2, our Dark-DSAR also spurpasses all of the other models.

Experiments on the HMDB51-Dark dataset. To further verify the performance of the proposed method, we synthesized an HMDB51-Dark version that simulates a dark environment by applying gamma correction. On the synthetic dark dataset, we present the results in Table 6. In terms of model performance, the hierarchy holds true: Transformer-based models outperform 3D CNN-based models, which in turn outperform 2D CNN-based models. When compared to the highest performing 2D model — TSM, the proposed method shows a 20.11% enhancement in Top-1 accuracy without employing the pre-training weights on the SICE dataset (Cai, Gu, & Zhang, 2018), and a 21.33% improvement with the pre-training weights integrated. Among the 3D models, SlowFast achieves the highest accuracy, but our Dark-DSAR presents a 19.44% improvement in comparison to it. For the Transformer-based models, we choose the MViTv2-B as the baseline, and our Dark-DSAR increases the Top-1 accuracy by 1.03% to it. Notably, our method achieves the top-1 accuracy of 61.9%, which is higher than other methods by more than 1%.

We compared the feature heat maps of Dark-DSAR on the ARID1.5 dataset for some types of action video after DSM processing and without DSM processing, as shown in Fig. 6. From the figure, we can see that the original input is in a relatively dark lighting condition, the color and detail information of the unDSM-processed image is not rich enough, and the scene and human features are not aggregated enough, and after DSM processing, these problems are improved at some degree.

4.3. Ablation study

Effectiveness of our DSM. We explore the effectiveness of the exploited DSM module for the task of action recognition on dark

Table 2
Comparison with the state-of-the-arts on the ARID1.5 dataset. Models are pretrained on the Kinetics400 (Kay et al., 2017) (K400) or the ImageNet (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009) dataset.

Model	Backbone	Pretrained	Top-1 (%)	Top-5 (%)	Params (M)	GFlops
TSN (Wang et al., 2016)	ResNet-50	K400	34.04	88.29	23.53	131.84
TSN (Wang et al., 2016)+DSM	ResNet-50	K400	36.57	81.52	23.54	136.15
TSM (Lin et al., 2019)	ResNet-50	K400	61.37	91.80	23.53	131.83
3D-ResNet-18* (Hara et al., 2017)	–	–	31.16	90.49	–	–
I3D-RGB* (Carreira & Zisserman, 2017)	–	ImageNet	48.75	90.61	–	–
I3D Two-stream* (Carreira & Zisserman, 2017)	–	–	51.24	90.95	–	–
SlowFast (Feichtenhofer et al., 2019)	ResNet3D-50	K400	66.76	91.58	62.02	97.29
SlowFast (Feichtenhofer et al., 2019)+DSM	ResNet3D-50		67.64	92.84	62.03	101.60
R(2 + 1)D (Tran et al., 2018)	ResNet3D-34	K400	63.52	92.29	63.54	162.85
R(2 + 1)D (Tran et al., 2018)+DSM	ResNet3D-34	K400	66.57	94.25	63.55	167.17
Swin-T (Liu et al., 2022)	–	K400	67.69	97.55	27.51	69.95
Swin-T (Liu et al., 2022)+DSM	–	K400	69.48	97.94	27.54	71.34
TimeSformer (Bertasius et al., 2021)	–	K400	49.36	93.17	196*	122*
UniFormerV2 (Li, Wang et al., 2022)	–	K400	61.41	96.04	115*	1800*
MViTv2-S (Li, Wu et al., 2022)	–	K400	83.38	97.32	34.24	160
MViTv2-S (Li, Wu et al., 2022)+DSM	–	K400	85.65	97.41	34.25	160
MViTv2-B (Li, Wu et al., 2022)	–	K400	87.68	98.55	50.93	225
R(2 + 1)D-GCN+BERT* (Singh et al., 2022)	–	–	86.93	99.35	–	–
Darklight (Chen et al., 2021)+R(2 + 1)D-34*	–	–	84.13	97.34	66.73	674.84
Dark-DSAR	–	K400	**89.49**	**98.77**	50.94	230

Table 3
Comparison with the state-of-the-arts on the ARID1.0 dataset.

Model	Top-1 (%)	Top-5 (%)
TSN (Wang et al., 2016)	52.54	94.17
I3D-RGB (Carreira & Zisserman, 2017)	72.78	99.39
I3D Two-stream (Carreira & Zisserman, 2017)	68.29	97.69
3D-ResNet101 (Hara et al., 2017)	71.57	99.03
R(2+1)D (Tran et al., 2018)	68.89	98.18
MViTv2-B (Li, Wu et al., 2022)	91.43	99.72
Swin-B (Liu et al., 2022)	89.79	99.53
TimeSformer-L (Bertasius et al., 2021)	81.39	98.26
UniFormerV2 (Li, Wang et al., 2022)	73.39	99.15
DANorm (Liang et al., 2022)	80.73	–
Suman et al. (2023)	95.86	**99.87**
UniFormerV2 (Li, Wang et al., 2022)-DSM	74.73	99.39
DarkLight-ResNeXt-101 (Chen et al., 2021)	87.27	99.47
DarkLight-R(2+1)D-34 (Chen et al., 2021)	94.04	99.87
DTCM (Tu, Liu et al., 2023)	96.36	99.02
Dark-DSAR	**96.99**	99.77

Table 4
Comparison of memory consumption and accuracy on the ARID1.0 dataset between DTCM and the proposed Dark-DSAR.

Model	Memory(M)	Top-1 (%)
DTCM (Tu, Liu et al., 2023)	2073	96.36
Dark-DSAR	828	96.99

videos. To verify the effectiveness of the DSM module, we select several widely used action recognition models as the baselines, and compare the accuracy of the model (without DSM, with DSM module the pre-training weight is not loaded, and the DSM module the pre-training weight is loaded) on three datasets. The input downsampling scale of DSM is set to 2, the number of convolutional layers is set to 7, the number of kernel channels is set to 32, and the other hyperparameters of the model are kept consistent with the baseline.

We conduct experiments on kinds of backbones like 2D, 3D, and Transformer, respectively. We choose TSN (Wang et al., 2016), the pioneer of 2D models, as the 2D baseline, SlowFast (Feichtenhofer et al., 2019) and R(2+1)D (Tran et al., 2018) as the 3D baseline, and Video Swin (Liu et al., 2022) and MViTv2 (Li, Wu et al., 2022), as the Transformer baseline. The training settings of the DSM are the same on each baseline, and the only difference on the same baseline is whether the DSM is used or not.

Effectiveness of DSM on the 2D backbone. In the 2D backbone, we choose TSN as the baseline, TSN is one of the typical 2D CNN-based methods for action recognition, it follows the structure of two streams, using RGB image and optical flow as the input of two branches respectively, applying sparse temporal sampling strategy to divide the long series into multiple snippets, and taking one frame from each clip as the final inputs. Results of TSN and TSN+DSM on the three datasets are listed in Table 7. For Top-1 accuracy, TSN with DSM module has 9.02% higher than that without DSM on the ARID1.5 dataset. Similarly, on the UAV-human-night and the synthetic HMDB51-Dark, DSM also brings improvement to the accuracy, 1.05% and 1.85% respectively.

Effectiveness of DSM on the 3D backbone. For the 3D backbones, we choose SlowFast and R(2+1)D as the baseline. SlowFast is a classical 3D CNN-based model for action classification. It employs dual channels with disparate frame rates as inputs, one operating at a faster rate and the other at a slower rate. SlowFast addresses the challenge of imbalanced temporal and spatial information within video action recognition. R(2+1)D captures both temporal and spatial information in videos by breaking down the 3D convolution operation into two distinct steps: a 2D spatial convolution followed by a 1D temporal convolution. The results of SlowFast and SlowFast+DSM on the three datasets are presented in Table 8. For Top-1 accuracy SlowFast with DSM module is 0.88% higher than that without DSM on the ARID1.5 dataset. For UAV-human-night, the DSM implementation leads to a 4.25% increment in the Top-1 accuracy. For Top-1 accuracy on the synthetic HMDB51-Dark, SlowFast with DSM outperforms the version without DSM by 2.76%.

The results of R(2+1)D and R(2+1)D+DSM on the three datasets are presented in Table 9. On the ARID1.5 dataset, the Top-1 accuracy of R(2+1)D with DSM module is 0.88% higher than that without DSM. On UAV-human-night, the DSM implementation leads to a 4.25% increment in the Top-1 accuracy. For Top-1 accuracy on the synthetic dataset HMDB51-Dark, SlowFast with DSM outperforms the version without DSM by 2.76%.

Table 5
Comparison with the state-of-the-arts on the UAV-human-night dataset.

Model	CSV1		CSV2		GFlops	Params (M)
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)		
TSN (Wang et al., 2016)	33.61	61.64	41.64	76.08	131.84	23.53
TSN (Wang et al., 2016)+DSM	35	60.77	42.36	77.26	136.15	23.54
TSM (Lin et al., 2019)	38.18	62.36	55.85	81.93	131.84	23.53
SlowFast (Feichtenhofer et al., 2019)	39.85	65.34	56.17	85.89	97.29	62.02
SlowFast (Feichtenhofer et al., 2019)+DSM	41.67	67.23	62.84	87.89	101.60	62.03
R(2 + 1)D (Tran et al., 2018)	39.19	62.88	54.08	82.37	162.85	63.54
R(2 + 1)D (Tran et al., 2018) +DSM	40.58	62.27	60.12	86.40	167.17	63.55
Swin-T (Liu et al., 2022)	39.88	65.93	62.79	87.35	69.95	27.51
Swin-T (Liu et al., 2022)+DSM	43.45	68.79	62.62	87.21	71.34	27.54
TimeSformer (Bertasius et al., 2021)	34.05	59.83	49.17	78.19	122*	196*
MViTv2-B (Li, Wu et al., 2022)	47.77	73.72	64.91	89.34	225	50.93
UniFormerV2 (Li, Wang et al., 2022)	42.35	69.57	58.68	88.06	1800*	115*
Dark-DSAR	**48.72**	**75.08**	**65.59**	**90.29**	230	50.94

Table 6
Comparison with the state-of-the-arts on the HMDB51-Dark dataset.

Model	Top-1 (%)	Top-5 (%)	GFlops	Params (M)
TSN (Wang et al., 2016)	42.61	71.66	131.84	23.53
TSN (Wang et al., 2016)+DSM	44.46	72.81	136.15	23.54
TSM (Lin et al., 2019)	52.81	80.02	131.84	23.53
SlowFast (Feichtenhofer et al., 2019)	54.08	83.13	97.29	62.02
SlowFast (Feichtenhofer et al., 2019)+DSM	56.84	83.33	101.60	62.03
R(2 + 1)D (Tran et al., 2018)	49.28	76.23	162.85	63.54
R(2 + 1)D (Tran et al., 2018) +DSM	51.81	78.48	167.17	63.55
Swin-T (Liu et al., 2022)	60.85	85.17	69.95	27.51
Swin-T (Liu et al., 2022)+DSM	61.55	85.42	71.34	27.54
TimeSformer (Bertasius et al., 2021)	51.22	78.67	122*	196*
MViTv2-B (Li, Wu et al., 2022)	65.66	87.36	225	50.93
UniFormerV2 (Li, Wang et al., 2022)	54.62	81.07	1800*	115*
Dark-DSAR	**67.13**	**88.28**	230	50.94

Table 7
Ablation study of TSN (Wang et al., 2016) with (w/) DSM or without (w/o) DSM on the three low-light human action datasets.

Datasets	Top-1 (%)		Top-5 (%)	
	w/	w/o	w/	w/o
ARID1.5	43.06 (+9.02)	34.04	88.16 (−0.13)	88.29
UAV-human-night	38.68 (+1.05)	37.63	69.02 (+0.16)	68.86
HMDB51-Dark	44.46 (+1.85)	42.61	72.81 (+1.15)	71.66

Table 8
Ablation study of SlowFast (Feichtenhofer et al., 2019) with (w/) DSM or without (w/o) DSM on the three low-light human action datasets.

Datasets	Top-1 (%)		Top-5 (%)	
	w/	w/o	w/	w/o
ARID1.5	67.64 (+0.88)	66.76	92.84 (+1.26)	91.58
UAV-human-night	52.26 (+4.25)	48.01	77.56 (+1.94)	75.62
HMDB51-Dark	56.84 (+2.76)	54.08	83.33 (+0.20)	83.13

Table 9
Ablation study of R(2+1)D (Tran et al., 2018) with (w/) DSM or without (w/o) DSM on the three low-light human action datasets.

Datasets	Top-1 (%)		Top-5 (%)	
	w/	w/o	w/	w/o
ARID1.5	66.57 (+3.05)	63.52	94.25 (+1.96)	92.29
UAV-human-night	50.35 (+3.71)	46.64	74.34 (+1.71)	72.63
HMDB51-Dark	51.81 (+2.53)	49.28	78.48 (+2.25)	76.23

Table 10
Ablation study of Video Swin (Liu et al., 2022) with (w/) DSM or without (w/o) DSM on the three low-light human action datasets.

Datasets	Top-1 (%)		Top-5 (%)	
	w/	w/o	w/	w/o
ARID1.5	69.48 (+1.79)	67.69	97.04 (−0.51)	97.55
UAV-human-night	53.04 (+11.70)	41.34	78.00 (+1.36)	76.64
HMDB51-Dark	61.55 (+0.70)	60.85	85.17 (−0.25)	85.42

The effectiveness of DSM on the Transformer-based backbone. For Transformer-based models, we choose Video Swin Tiny and MViTv2-B as the baseline. Results of Swin-T and MViTv2-B w/DSM or w/o DSM on three datasets are listed in Table 10 and Table 11. Compared to the baselines, when equipped with DSM, Video Swin improves the accuracy by 1.79% (69.48% vs. 67.69%), 1.7% (53.04% vs. 51.34%), 0.7% (61.55% vs. 60.85%), respectively. As shown in Table 11, DSM brings in a minimum of 0.47% improvement in accuracy for MViTv2.

Effect of the scale factor. The amount of information carried by the video data is closely related to the resolution of the video. Generally, the higher the video resolution, the more detailed information it carries. However, the higher the resolution of the original input, the higher the computational consumption is. Consequently, it is extremely important to choose an input size suitable for the model depending on the downstream task, which is very meaningful to reduce unnecessary

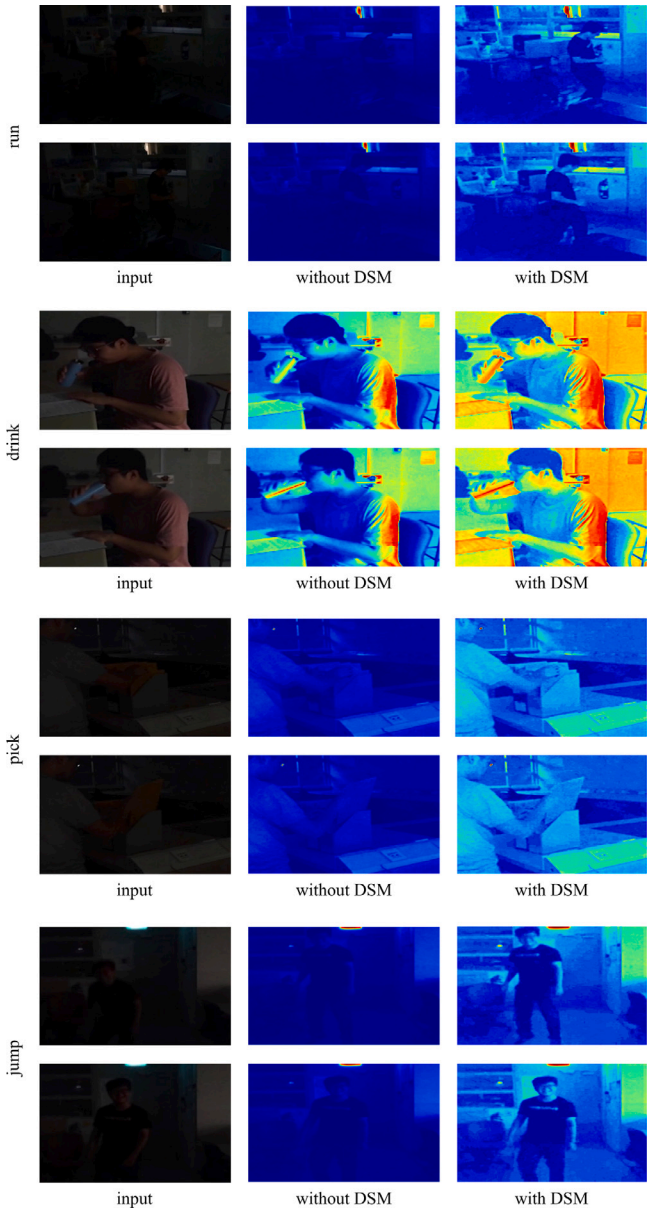


Fig. 6. The figure displays the results of visualizing the feature maps at the front of the action classifier for Dark-DSAR with and without DSM, using four action classes from the ARID1.5 dataset as examples. The left shows the RGB frame of the original input. The middle presents the feature map before inputting the action classifier without DSM, and the right shows the visualization results with DSM.

Table 11
Ablation study of MViTv2-B (Li, Wu et al., 2022) with (w/) DSM or without (w/o) DSM on the three low-light human action datasets.

Datasets	Top-1 (%)		Top-5 (%)	
	w/	w/o	w/	w/o
ARID1.5	88.15 (+0.47)	87.68	98.47 (−0.08)	98.55
UAV-human-night	57.16 (+0.82)	56.34	82.69 (+1.16)	81.53
HMDB51-Dark	67.13 (+1.47)	65.66	88.28 (+0.92)	87.36

memory expenses. Wang et al. (2021) pointed out the existence of significant spatially redundant information in high-resolution videos. Ma et al. (2022) showed that videos do not have obvious information loss during downsampling in a certain range (especially for some tasks like video action recognition), and the reason for the degradation of

Table 12
Changing the scale factor of the DSM module, the accuracy varied accompany with the input videos' resolution.

Scale factor	Memory (M)	Top-1 (%)	Top-5 (%)
1 (baseline)	1469	67.69	97.55
2	828	**68.46**	**97.87**
4	428	65.30	**97.87**
8	334	64.58	95.95

model performance after downsampling is the mismatch between the network structure and the input size. Based on the above knowledge, we changed the input size of the DSM module to explore the optimal input size to match the network. To harmonize with the data processing mechanism of the GPU, we set the scale factor to an exponential level of 2 and studied the change in accuracy when the scale factor varies. During the experiment, the number of the convolutional layers and the feature map channels were set to 7 and 32 respectively.

The results are shown in Table 12. When not scaling or scaling by a small amount, the accuracy is maintained and even gets higher if the scale factor is 2 (68.46% vs. 67.69%). As mentioned above, scaling at this level removes the spatial redundant information in the data and retains the effective information for action recognition, leading to the model performs well. When the scale factor is further increased, the recognition accuracy decreases significantly (65.40% vs. 68.46%, 64.58% vs. 68.46%), because the downsampling at this time no longer only removes the spatial redundancy information, but also removes some important information carried by the pixels.

Effect of the number of convolutional layers. We hope that the proposed plug will introduce as little computational consumption and fewer additional parameters as possible without affecting the network performance. Based on this, in this section, we explore the effect of convolutional network depth on the DSM structure. The DSM structure uses symmetric connections of U-Net networks, so in our experiments, we change the number of network layers in pairs and delete the feature extraction layers and upsampling layers which have the same scale. The final count of layers is set to 7, 5, and 3, respectively. The scale factor is fixed to 1 and the number of feature map channels is fixed to 32 during the experiments, and the experimental results are shown in Table 13. When other parameters of DSM are fixed to be set, the variable of single convolutional layers has less influence on the accuracy of action recognition. It can be known from the table that there is no significant decreasing trend of recognition accuracy when the number of network layers is reduced (67.70% vs. 67.69%, 66.84% vs. 67.69%), but this does not mean that network layers have no influence on model performance, which is presented in Fig. 7. When the scale factor is fixed to 1, the number of feature map channels is fixed to 24, and the number of convolutional layers is fixed to 7 (3 pairs), the model maintains a satisfactory performance. A small number of channels with a modest scale factor or a low number of channels combined with numerous convolutional layers can also obtain effective recognition results within the appropriate range. From Fig. 7, we can see that the scale factor is a critical determinant of the model performance. This observation aligns with the previously stated conclusion that the input size (resolution) plays an important role in influencing action recognition. The accuracy remains relatively stable with small scaling changes but significantly drops as scaling increases. The convolutional layers, scale factor, and number of channels influence each other, which in turn affects recognition accuracy.

Effect of the convolution layer. In addition to the number of convolutional layers, we explored the internal convolutional composition of our DSM. In this experiment, we fixed the other parameters and structural settings of the DSM and substituted PW Conv and DW Conv. Experimental results are shown in Table 14. The worst results are obtained when only PW Conv is used (64.56% vs. 69.48%), which

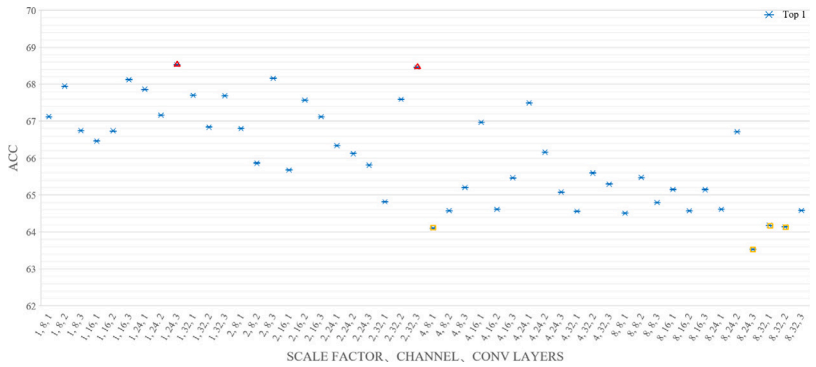


Fig. 7. Top-1 accuracy on the ARID1.5 dataset. Where Video Swin (Liu et al., 2022) is the backbone, DSM with various combinations of scale factor, channel numbers, and convolutional layers. The horizontal coordinates consist of three values separated by ",", representing the scale factor, the number of channels, and the convolutional layers, respectively. △ indicates of relatively high Top-1 accuracy, and □ indicates the low Top-1 accuracy.

Table 13
Changing the convolutional layers, pairs of Layer 1, 2, and 3 represent the DSM with 3, 5, and 7 convolutional layers, respectively.

Layers (pair)	Top-1 (%)	Top-5 (%)
3 (baseline)	67.69	97.55
2	66.84	97.20
1	**67.70**	**97.67**

Table 14
Change the convolution layer of the DSM module.

Conv settings	Top-1 (%)	Top-5 (%)
DW+PW (baseline)	**69.48**	97.94
DW+DW	66.08	97.01
PW+DW	66.09	**98.05**
PW+PW	64.56	95.92

Table 15
Changing the convolution channels of the DSM module, and the scale factor is fixed as 1.

Channels	Top-1 (%)	Top-5 (%)
32 (baseline)	67.69	**97.55**
24	**68.54**	97.53
16	68.13	97.32
8	66.75	96.71

is due to the fact that PW Conv uses window-aware features of 1×1 size, which extracts less spatial neighborhood information compared to 3×3 kernel. Top-1 accuracy is also unsatisfying when using only DW Conv or using PW Conv first and then the DW Conv. This is because the feature representation ability of the model is limited when using only DW Conv. DW Conv uses the same convolution kernel on each channel and can only perform convolution operations on each input channel, but cannot interact information across channels, which limits the model's ability to perceive local features. In addition, when DW Conv is performed after PW Conv, the model has lost some detailed information of the original features in the pre-modeling stage, which also affects the performance of the model. Performing DW Conv first and then using PW Conv can reduce the amount of computation. Besides, it can effectively integrate the depth and channel-associated features, so it has the best recognition performance.

Effect of the number of feature map channels of DSM. The number of channels is related to the order of magnitude of the quadratic operations in the enhancement curve, and here we tried to reduce the computational expense of the proposed plug by decreasing the number of channels. We selected four different channel numbers of 8, 16, 24, and 32, with a fixed scale factor of 1 and a fixed number of convolutional layers of 32. The accuracy of action recognition on the ARID1.5 dataset for the four cases is shown in Table 15. From the results, we can find that decreasing the number of channels degrades the recognition accuracy. The feature map channels in the DSM structure represent different features of the estimated curve that can be extracted, and decreasing the number of channels reduces the curve parameter features that can be learned, and the network cannot fully extract the curve features. As a result, losing the abstract features and the detailed information required for illumination condition recovery, which further affects the action recognition accuracy under dark conditions. Of course, this does not mean that more channels are better. Increasing the number of channels will result in a higher count of network parameters and computational effort, consequently, promoting the risk of overfitting.

Trade-offs between accuracy and inference efficiency. For DSM, a large scale factor, fewer channels, and fewer convolutional layers imply higher computational and inference efficiency, but at the same time recognition accuracy suffers a bit due to unrefined feature extraction. In this part, we give a more detailed description of Fig. 7, where all experiments are performed on 4 NVIDIA 2080Ti GPUs, the learning rate is set to $2.5e^{-4}$, and the number of training epoch is set to 15. It can be seen from 7 that the parameter that has the greatest impact on the accuracy rate is the scale factor. The scale factor gradually increases from left to the right, and the overall trend of the accuracy shows a decreasing state. Table 12 and Fig. 7, reveal that as the scale factor increases, the memory cost gradually decreases. With a scale factor of 8, memory consumption and reasoning time are minimal, but accuracy is correspondingly low. However, reducing the scale factor from 8 to 2 results in a notable increase in accuracy, while the relative rise in memory consumption remains within an acceptable range. When the scale factor is reduced from 2 to 1, the memory consumption increases by more than double, but the accuracy has no significant growth. When the scale factor is reduced from 2 to 1, the memory consumption increases more than twofold and there is no significant increase in accuracy. Based on the observation, we set our scale factor of the DSM module to 2. To compensate for the detail loss caused by the scale factor, we tend to set the number of channels and the number of convolutional layers to larger values, such as 32 and 3.

5. Conclusions

To handle the task of dark video human action recognition, this work proposed a novel model Dark-DSAR, a cross-domain end-to-end framework focusing on action recognition in dark scenes, achieves framework coherence and model lightweight. Our main contributions are three-fold: (1) Designing a domain shift module (DSM) to migrate

videos from dark domain to light-optimized domain, which can be assembled into arbitrary action recognition backbones without introducing large computation. (2) Integrating illumination domain adaption and action classification into a single stream model, which significantly simplifies the complex training stages of the existing two-step pipelines. (3) Exploring the matching relationship between video resolution and the model itself, decreasing the model parameters by compressing the spatial resolution of the videos. Extensive experiments demonstrate that our Dark-DSAR is lightweight and shows superior accuracy on all the three dark video action recognition datasets.

CRediT authorship contribution statement

Yuwei Yin: Conceptualization, Software, Writing – original draft. **Miao Liu:** Data curation, Resources. **Renjie Yang:** Formal analysis, Investigation. **Yuanzhong Liu:** Investigation. **Zhigang Tu:** Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by the Natural Science Fund for Distinguished Young Scholars of Hubei Province under Grant 2022CFA 075. It was also supported in part by the National Natural Science Foundation of China under Grant 62106177 and the Fundamental Research Funds for the Central Universities under Grant 2042023KF0180. The numerical calculation was supported by the supercomputing systems in the Super-computing Center of Wuhan University.

References

Akula, A., Shah, A. K., & Ghosh, R. (2018). Deep learning approach for human action recognition in infrared images. *Cognitive Systems Research*, *50*, 146–154.

Anwaar-ul-Haq, Gondal, I., & Murshed, M. (2011). Contextual action recognition in multi-sensor nighttime video sequences. In *Proc. int. conf. digital image computing: techniq. applic.* (pp. 256–261). The Organization, http://dx.doi.org/10.1109/DICTA.2011.49.

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *Vol. 2*, In *ICML* (3), (p. 4). The Organization.

Cai, J., Gu, S., & Zhang, L. (2018). Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, *27*(4), 2049–2062. http://dx.doi.org/10.1109/TIP.2018.2794218.

Cai, G., Wang, Y., He, L., & Zhou, M. (2020). Unsupervised domain adaptation with adversarial residual transform networks. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(8), 3073–3086. http://dx.doi.org/10.1109/TNNLS.2019.2935384.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 4724–4733). The Organization, http://dx.doi.org/10.1109/CVPR.2017.502.

Chen, R., Chen, J., Liang, Z., Gao, H., & Lin, S. (2021). DarkLight networks for action recognition in the dark. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. workshops* (pp. 846–852). The Organization, http://dx.doi.org/10.1109/CVPRW53098.2021.00094.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 1800–1807). The Organization, http://dx.doi.org/10.1109/CVPR.2017.195.

Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., et al. (2022). You only need 90K parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *BMVC* (p. Vol. 238). The Organization.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 248–255). The Organization, http://dx.doi.org/10.1109/CVPR.2009.5206848.

Eum, H., Lee, J., Yoon, C., & Park, M. (2013). Human action recognition for night vision using temporal templates with infrared thermal camera. In *Int. conf. ubiquitous robots ambient intell.* (pp. 617–621). The Organization, http://dx.doi.org/10.1109/URAI.2013.6677407.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In *Proc. IEEE/CVF int. conf. comput. vis.* (pp. 6201–6210). The Organization, http://dx.doi.org/10.1109/ICCV.2019.00630.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *Vol. 37*, In *Proc. int. conf. machine learning* (pp. 1180–1189). The Organization.

Gao, X., Du, S., & Yang, Y. (2023). Glimpse and focus: Global and local-scale graph convolution network for skeleton-based action recognition. *Neural Networks*, *167*, 551–558. http://dx.doi.org/10.1016/j.neunet.2023.07.051.

Gao, H., Guo, J., Wang, G., & Zhang, Q. (2022). Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 9903–9913). The Organization, http://dx.doi.org/10.1109/CVPR52688.2022.00968.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proc. int. conf. artificial intelligence and statistics* (pp. 315–323). The Organization.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Vol. 2*, In *Proc. int. conf. neural infor. process. syst.* (pp. 2672–2680). The Organization.

Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., et al. (2020). Zero-reference deep curve estimation for low-light image enhancement. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 1777–1786). The Organization, http://dx.doi.org/10.1109/CVPR42600.2020.00185.

Hara, K., Kataoka, H., & Satoh, Y. (2017). Learning spatio-temporal features with 3D residual networks for action recognition. In *Proc. IEEE/CVF int. conf. comput. vis. workshops* (pp. 3154–3160). The Organization, http://dx.doi.org/10.1109/ICCVW.2017.373.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hira, S., Das, R., Modi, A., & Pakhomov, D. (2021). Delta sampling R-BERT for limited data and low-light action recognition. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. workshops* (pp. 853–862). The Organization, http://dx.doi.org/10.1109/CVPRW53098.2021.00095.

Hussain, A., Khan, S. U., Khan, N., Rida, I., Alharbi, M., & Baik, S. W. (2023). Low-light aware framework for human activity recognition via optimized dual stream parallel network. *Alexandria Engineering Journal*, *74*, 569–583.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., et al. (2021). EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, *30*, 2340–2349. http://dx.doi.org/10.1109/TIP.2021.3051462.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., et al. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. *Vol. 70*, In *Proc. int. conf. machine learning* (pp. 1857–1865). The Organization.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. In *Proc. IEEE/CVF int. conf. comput. vis.* (pp. 2556–2563). The Organization, http://dx.doi.org/10.1109/ICCV.2011.6126543.

Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, *61*(1), 1–11. http://dx.doi.org/10.1364/JOSA.61.000001.

Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.-M., Gu, J., et al. (2022). Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9396–9416. http://dx.doi.org/10.1109/TPAMI.2021.3126387.

Li, C., Guo, C., & Loy, C. C. (2022). Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 4225–4238. http://dx.doi.org/10.1109/TPAMI.2021.3063604.

Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., & Li, Z. (2021). UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 16261–16270). The Organization, http://dx.doi.org/10.1109/CVPR46437.2021.01600.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., et al. (2022). Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. arXiv preprint arXiv:2211.09552.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., et al. (2022). MViTv2: Improved multiscale vision transformers for classification and detection. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 4794–4804). The Organization, http://dx.doi.org/10.1109/CVPR52688.2022.00476.

Liang, Z., Chen, J., Chen, R., Zheng, B., Zhou, M., Gao, H., et al. (2022). Domain adaptable normalization for semi-supervised action recognition in the dark. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. workshops* (pp. 4250–4257). The Organization, http://dx.doi.org/10.1109/CVPRW56347.2022.00470.

Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal shift module for efficient video understanding. In *Proc. IEEE/CVF int. conf. comput. vis.* (pp. 7082–7092). The Organization, http://dx.doi.org/10.1109/ICCV.2019.00718.

Lin, F., Zhang, H., Wang, J., & Wang, J. (2023). Unsupervised image enhancement under non-uniform illumination based on paired CNNs. *Neural Networks*, http://dx.doi.org/10.1016/j.neunet.2023.11.014.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., et al. (2022). Video swin transformer. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 3192–3201). The Organization, http://dx.doi.org/10.1109/CVPR52688.2022.00320.

Liu, Y., Yuan, J., & Tu, Z. (2022). Motion-driven visual tempo learning for video-based action recognition. *IEEE Transactions on Image Processing*, *31*, 4104–4116. http://dx.doi.org/10.1109/TIP.2022.3180585.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lv, F., Lu, F., Wu, J., & Lim, C. (2018). MBLLEN: Low-light image/video enhancement using CNNs. *Vol. 220*, In *BMVC* (1), (p. 4). The Organization.

Ma, C., Guo, Q., Jiang, Y., Luo, P., Yuan, Z., & Qi, X. (2022). Rethinking resolution in the context of efficient video recognition. *Vol. 35*, In *Proc. int. conf. neural infor. process. syst.* (pp. 37865–37877). The Organization.

Ma, J., Wang, G., Zhang, L., & Zhang, Q. (2023). Restoration and enhancement on low exposure raw images by joint demosaicing and denoising. *Neural Networks*, *162*, 557–570. http://dx.doi.org/10.1016/j.neunet.2023.03.018.

Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *2017 IEEE international conference on computer vision* (pp. 5716–5726). http://dx.doi.org/10.1109/ICCV.2017.609.

Peng, X., Hoffman, J., Yu, S. X., & Saenko, K. (2016). Fine-to-coarse knowledge transfer for low-res image classification. In *IEEE int. conf. image processing* (pp. 3683–3687). The Organization, http://dx.doi.org/10.1109/ICIP.2016.7533047.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., et al. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, *39*(3), 355–368. http://dx.doi.org/10.1016/S0734-189X(87)80186-X.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Proc. int. conf. medical image computing computer-assisted intervention* (pp. 234–241). The Organization.

Shao, Y., Li, L., Ren, W., Gao, C., & Sang, N. (2020). Domain adaptation for image dehazing. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 2805–2814). The Organization, http://dx.doi.org/10.1109/CVPR42600.2020.00288.

Shen, X., Pan, S., Choi, K.-S., & Zhou, X. (2023). Domain-adaptive message passing graph neural network. *Neural Networks*, *164*, 439–454. http://dx.doi.org/10.1016/j.neunet.2023.04.038.

Singh, H., Suman, S., Subudhi, B. N., Jakhetiya, V., & Ghosh, A. (2022). Action recognition in dark videos using spatio-temporal features and bidirectional encoder representations from transformers. *IEEE Transactions on Artificial Intelligence*, 1–11. http://dx.doi.org/10.1109/TAI.2022.3221912.

Suman, S., Naharas, N., Subudhi, B. N., & Jakhetiya, V. (2023). Two-streams: Dark and light networks with graph convolution for action recognition from dark videos (student abstract). In *Proc. AAAI conf. artif. intell.* (pp. 16340–16341). The Organization.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 6450–6459). The Organization, http://dx.doi.org/10.1109/CVPR.2018.00675.

Tu, Z., Liu, X., & Xiao, X. (2022). A general dynamic knowledge distillation method for visual analytics. *IEEE Transactions on Image Processing*, *31*, 6517–6531. http://dx.doi.org/10.1109/TIP.2022.3212905.

Tu, Z., Liu, Y., Zhang, Y., Mu, Q., & Yuan, J. (2023). DTCM: Joint optimization of dark enhancement and action recognition in videos. *IEEE Transactions on Image Processing*, *32*, 3507–3520. http://dx.doi.org/10.1109/TIP.2023.3286254.

Tu, Z., Xie, W., Dauwels, J., Li, B., & Yuan, J. (2019). Semantic cues enhanced multimodality multistream CNN for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(5), 1423–1437. http://dx.doi.org/10.1109/TCSVT.2018.2830102.

Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R. C., Li, B., et al. (2018). Multi-stream CNN: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, *79*, 32–43.

Tu, Z., Zhang, J., Li, H., Chen, Y., & Yuan, J. (2023). Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Transactions on Multimedia*, *25*, 1819–1831. http://dx.doi.org/10.1109/TMM.2022.3168137.

Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *Proc. IEEE/CVF int. conf. comput. vis.* (pp. 4068–4076). The Organization, http://dx.doi.org/10.1109/ICCV.2015.463.

Ulhaq, A. (2018). Action recognition in the dark via deep representation learning. In *IEEE int. conf. image process. applic. syst.* (pp. 131–136). The Organization, http://dx.doi.org/10.1109/IPAS.2018.8708903.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Vol. 30*, In *Proc. int. conf. neural infor. process. syst.*. The Organization.

Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., & Huang, G. (2021). Adaptive focus for efficient video recognition. In *Proc. IEEE/CVF int. conf. comput. vis.* (pp. 16229–16238). The Organization, http://dx.doi.org/10.1109/ICCV48922.2021.01594.

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, *312*, 135–153. http://dx.doi.org/10.1016/j.neucom.2018.05.083.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., et al. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *Proc. eur. conf. comput. vis.* (pp. 20–36). The Organization.

Wei, C., Wang, W., Yang, W., & Liu, J. (2018). Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560.

Xu, X., Wang, R., Fu, C.-W., & Jia, J. (2022). SNR-aware low-light image enhancement. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 17693–17703). The Organization, http://dx.doi.org/10.1109/CVPR52688.2022.01719.

Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., & See, S. (2021). ARID: A new dataset for recognizing action in the dark. In *Deep learning for human activity recognition* (pp. 70–84). The Organization.

Zhang, J., Jia, Y., Xie, W., & Tu, Z. (2022). Zoom transformer for skeleton-based group activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(12), 8646–8659. http://dx.doi.org/10.1109/TCSVT.2022.3193574.

Zhang, J., Ye, G., Tu, Z., Qin, Y., Qin, Q., Zhang, J., et al. (2022). A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Transactions on Intelligence Technology*, *7*(1), 46–55. http://dx.doi.org/10.1049/cit2.12012.

Zheng, C., Li, Z., Yang, Y., & Wu, S. (2021). Single image brightening via multi-scale exposure fusion with hybrid learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(4), 1425–1435. http://dx.doi.org/10.1109/TCSVT.2020.3009235.