DTCM: Joint Optimization of Dark Enhancement and Action Recognition in Videos

Zhigang Tu, *Member, IEEE,* Yuanzhong Liu, Yan Zhang, Qizi Mu, and Junsong Yuan, *Fellow, IEEE*

***Abstract*—Recognizing human actions in dark videos is a useful yet challenging visual task in reality. Existing augmentation-based methods separate action recognition and dark enhancement in a two-stage pipeline, which leads to inconsistently learning of temporal representation for action recognition. To address this issue, we propose a novel end-to-end framework termed Dark Temporal Consistency Model (DTCM), which is able to jointly optimize dark enhancement and action recognition, and force the temporal consistency to guide downstream dark feature learning. Specifically, DTCM cascades the action classification head with the dark augmentation network to perform dark video action recognition in a one-stage pipeline. Our explored spatio-temporal consistency loss, which utilizes the RGB-Difference of dark video frames to encourage temporal coherence of the enhanced video frames, is effective for boosting spatio-temporal representation learning. Extensive experiments demonstrated that our DTCM has remarkable performance: 1) Competitive accuracy, which outperforms the state-of-the-arts on the ARID dataset by 2.32% and the UAVHuman-Fisheye dataset by 4.19% in accuracy, respectively; 2) High efficiency, which surpasses the current most advanced method [1] with only 6.4% GFLOPs and 71.3% number of parameters; 3) Strong generalization, which can be used in various action recognition methods (e.g., TSM, I3D, 3D-ResNext-101, Video-Swin) to promote their performance significantly.**

***Index Terms*—Dark Video Action Recognition, Unified Framework, Dark Temporal Consistency, Representation Learning**

I. Introduction

VIDEO data has explosively grown in recent years, some of them are captured under undesired lighting conditions due to environmental and/or technical constraints. Although deep neural networks have achieved great success in the video action recognition task [2], [3], [4], [5], [6], recognizing human actions in low-light conditions is still a challenging problem. Compared with videos that are recorded under normal illumination, videos shot under low illumination have special characteristics [7]: *low brightness and low contrast*. These limitations result in subpar video quality, which significantly undermining the performance of action recognition methods [8], [9], [10]. These methods were primarily designed for standard-quality videos, thus the limitations obstruct their applicability in low-light video environments [11], [12], [13] in applications involving dark videos. Consequently, exploring effective action recognition method that works well in dark videos is an urgent task.

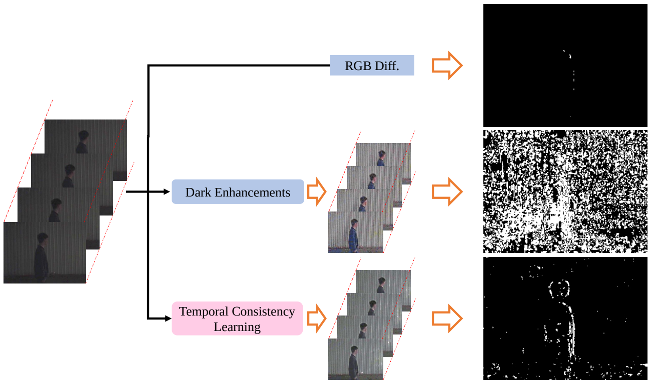

Fig. 1. An illustration of using RGB-Difference to capture temporal continuity of the video sequence, where the difference before and after dark enhancement is presented. It can be found that the dark enhancement often breaks the temporal continuity of the video and can be relieved by learning the temporal consistency. Here is an example of the action “Turn” from the ARID dataset [7].

Despite some efforts have been done [1], [7], [14], [15], [16], the problems for dark video action recognition are far from being solved. For example, the way to learn dark features directly from dark videos relies on large, well-labeled datasets, but such a desired dataset is not available at present. Another way is the domain adaptive feature learning, e.g., [17], [18], [19],but they are computational expensive, as involving training the large-scale video datasets (e.g., Kinetics400 [4], Something-Something [20]). Benefiting from the intuitiveness and efficiency, augmentation-based methods [1], [7], [21] are widely used. They perform dark video action recognition in two-stage: 1) conducting frame dark enhancement for dark videos, 2) using the enhanced results to execute action recognition. However, isolating action recognition and dark enhancement in a two-stage pipeline may result in suboptimal recognition performance. Studies [7], [22] have found that some enhancement methods can be regarded as artifacts or even adversarial attacks for video action recognition, since they focus more on enhancing the visibility rather than augmenting the visual features for recognition. On the other hand, the two-stage pipeline methods, which directly apply image enhancement to dark videos, often break the temporal consistency among the input dark video frames and their enhanced version. Because the enhancement methods [23], [24], [25]

Zhigang Tu and Yuanzhong Liu are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 430079 Wuhan, China. (Zhigang Tu and Yuanzhong Liu contributed equally; Yuanzhong Liu and Yan Zhang are co-corresponding authors).

Yan Zhang is with the Department of Clinical Laboratory, Renmin Hospital of Wuhan University, 430060 Wuhan, China.

Qizi Mu is with CHN Energy Digital Intelligence Technology Development (BeiJing) CO. LTD, Beijing, China.

Junsong Yuan is with the Computer Science and Engineering department, State University of New York at Buffalo, USA.

are designed for images, ignoring the spatio-temporal correlation among the image sequence. The temporal inconsistency decreases the action recognition performance, since it affects the spatio-temporal representation learning. Consequently, it is crucial to learn the temporal consistency for downstream spatio-temporal representation learning in an end-to-end way to boost the performance of dark video action recognition.

One representative method is [26], which infers the motion prior for single image and enforces the temporal consistency for low-light video enhancement. However, it can only be trained on the synthetic images and its performance is not satisfactory with respect to real videos. How to make full use of the neighboring frames to improve the enhancement capacity and accelerate the processing speed is remains an unsolved issue [27].

To address this issue, we propose an end-to-end unified framework named Dark Temporal Consistency Model (DTCM), which can exploit the temporal consistency information between neighboring video frames, learn temporal representation interactively with dark video action recognition, and expedite the enhancement speed in a one-stage pipeline. As shown in Fig. 2, the explored DTCM cascades the enhancement network and the action recognition network to form a single stream model, in which the dark video frames are well enhanced to be used for action classification. Specifically, an improved Zero-DCE [28] is adopted for dark enhancement and the 3D-ResNext-101 [29] is chosen as the action classifier based on extensive experimental examination. The transfomer-based action recognition models *i.e.* TimeSformer [30] and VideoSwin [31] are not used here due to their expensive computation cost. Remarkably, the proposed DTCM is flexible, where any darkness enhancement networks and action recognition networks can be easily integrated with it.

Moreover, to extract the temporal information for temporal consistency learning, we explore a temporal consistency loss to maintain the temporal smoothness of the consecutive video frames. Besides, we design a modified spatial consistency loss with the usage of more local spatial information to strengthen the spatial stability inspired by Zero-DCE [28]. The temporal consistency is learned directly from the real-world dark videos, and the optimized spatio-temporal features are captured for action recognition. As shown in Fig. 1, after temporal consistency learning, the original temporal continuity between successive video frames is preserved and robust video feature learning is benefited.

In addition, to accelerate the dark video enhancement speed, we share the computation between adjacent video frames during enhancement based on the illumination invariant assumption. We also lighten the dark enhancement network of [28] to further improve the efficiency, where both the number of parameters and the computation cost are reduced by about 70%.

On the other side, to promote the research of action recognition in dark videos, we construct a dataset named Dark-48 with 8815 extremely dark videos in 48 categories, which is publicly released at the website https://github.com/yzfly/Dark48. Compared to the most widely used dark video dataset ARID [7], our Dark-48 is about 2.3× (8815 videos vs. 3784 videos) larger, and also contains richer semantic features of dark actions (48 classes vs 11 classes). Videos in our Dark-48 are collected from various existing action recognition datasets by evaluating the video darkness, where the evaluation measure is described in the last paragraph of section III (see Eq. 14).

To verify that whether dark video action recognition can be benefited from the end-to-end training, we test different training settings. Experimental results show that the presented DTCM method not only enhances the quality of dark video enhancement but also significantly boosts the accuracy of dark video action recognition. For example, its action recognition accuracy achieves to Top-1 97.83% and Top-1 33.36% on the ARID dataset [7] and the UAVHuman-Fisheye dataset [32], respectively. In summary, our main contributions lie in three aspects:

- We propose an end-to-end learning model DTCM to recognize human actions in dark videos, where the dark video is jointly enhanced for optimizing the performance of action recognition in a one-stage pipeline.
- We exploit a spatio-temporal consistency loss to preserve the temporal smoothness of the enhanced dark video frames, which learns robust spatio-temporal representation from the dark action videos.
- A large-scale dataset termed Dark-48 with extremely dark videos is collected for dark video action recognition investigation, which contains larger action categories and richer semantics than the existing datasets, e.g. ARID [7].
- Extensive experiments demonstrated that the explored DTCM method achieves the state-of-the-art performance on both accuracy and efficiency for recognizing human actions in dark videos.

II. Related Work

A. Dark Image Enhancement

For dark image enhancement, the traditional methods, *e.g.* GIC [33] is often used to adjust the image luminance, whereas HE [34] produces higher contrast images. The multi-scale retinex (MSR) method [35] provides color constancy and dynamic range compression by using a retinex, which combines several SSR outputs, to produce a single output image. LIME [36] achieves the enhancement by estimating and refining a low-light image illumination map. For the recent deep learning methods, *e.g.* KinD [37], which is inspired by the retinex theory, decomposes image enhancement into light adjustment and degradation removal. Zero-DCE [28] formulates light enhancement as a task of image-specific curve estimation with a deep network. StableLLVE [26] enforces the temporal stability of low light video enhancement with only static images, and learns and infers motion field (*i.e.* optical flow, which is time-consuming to estimate) from a single image, then synthesizes the short-range video sequences. LL-Net [38] uses a deep autoencoder-based approach to identify signal features from low-light images and adaptively brightens images without over-amplifying/saturating the lighter parts in images with a high dynamic range. HLA-Face [39] boosts the dark images and degrades the normal-light images, making both domains move toward each other for face detection in

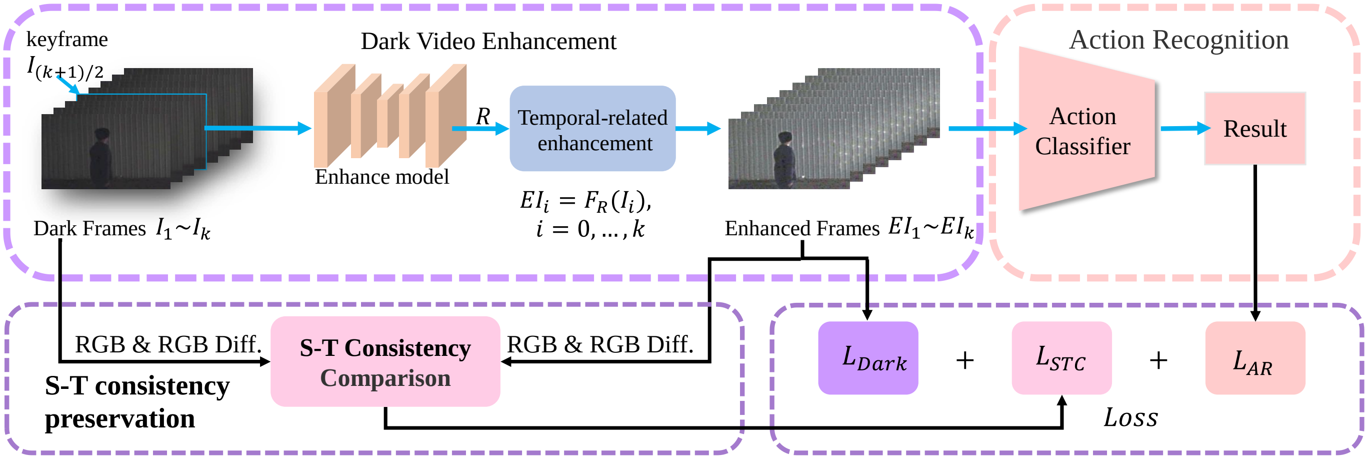
keyframe
$I_{(k+1)/2}$
Dark Video Enhancement
Enhance model
R
Temporal-related enhancement
$EI_i = F_R(I_i)$, $i = 0, \dots, k$
Dark Frames $I_1 \sim I_k$
Enhanced Frames $EI_1 \sim EI_k$
Action Recognition
Action Classifier
Result
RGB & RGB Diff.
S-T Consistency Comparison
RGB & RGB Diff.
S-T consistency preservation
L_{Dark}
+
L_{STC}
+
L_{AR}
$Loss$

Fig. 2. Overview the structure of the proposed Dark Temporal Consistency Model (DTCM), which consists primary components of dark video enhancement and action recognition. Action recognition is cascaded with dark video enhancement to form a single-stream model and being jointly optimized in a one-stage pipeline. Spatio-temporal (S-T) consistency of the enhanced video frames, which is beneficial to video action recognition in dark, can be preserved by comparing the RGB-difference before and after dark video enhancement. Dark enhancement is efficiently performed by temporal-related enhancement, which reusing the parameters R (estimated by only the *keyframe* as Eq. 11) to enhance all video frames for reducing parameter estimation's time-consumption.

dark. However, these methods process a low-light video frame-by-frame, which ignoring the temporal information to improve the enhancement performance. In contrast, our DTCM takes the correlation of neighboring video frames into account to preserve video temporal smoothness and reduce the time of parameter estimation effectively.

B. Action Recognition in Dark Videos

Although various action recognition methods [40], [41], [42], [10] and datasets [43], [44], [45] have been investigated and proposed, action recognition in dark videos is yet a challenging task. Due to the low quality of dark video, the existing optical flow estimation methods [46], [47], [48], [49], [50] are unable to obtain accurate motion information, causing the two-stream approaches [51], [52], [53], [54] perform poor. Skeleton-based action recognition methods [55], [56] are also commanly used, since the skeleton representation is much more efficient and robust than other modalities such as RGB frames [57]. Zhou et al. [58] propose a feature refinement module equipped with contrastive learning to solve the ambiguous actions for skeleton-based action recognition. TD-GCN [59] applies different adjacency matrices for skeletons from different frames to improve the flexibility of GCN. Mask-GCN [60] focuses on learning action-specific skeleton joints to handle motion patterns in a practical skeleton-based action recognition task. Although these skeleton-based action recognition methods improve the recognition robustness, they highly rely on the quality of the human pose estimation results, but the performance is unsatisfactory in extremely low-light images [61]. Ulhaq [14] presents an action recognition strategy with multiple video streams by using deep multi-view representation learning. Recently, Xu et al. [7] collect the first dataset named ARID focused on human actions in dark videos, and find that the current action recognition models and image enhancement methods are not effective for the task of dark video action recognition. DarkLight [1] utilizes both dark videos and their brightened counterpart to form a dual-pathway structure for learning video representation. Although performance is significantly improved, expensive computation is required. All of these methods conduct dark video action recognition in a two-stage pipeline, which breaks the connection between enhancement and recognition, and leads to the non-robust and unsatisfactory recognition performance. Our DTCM creates a correlation for enhancement and recognition by jointly optimizing dark enhancement and action classification interactively in dark videos.

C. Domain Adaptation

Image-to-image translation has been extensively studied before for domain adaptation, but few works have explicitly investigated visual domain adaptation for dark video action recognition. CyCADA [62] enforces cycle-consistency, while leverages a task loss and adapts representation learning at both the pixel-level and feature-level for image domain adaptation. [63] learns an invertable generator, which can transform the appearance condition of images, to improve visual place recognition and metric localization under appearance change. Despite these methods can be extended to videos by processing a video frame-by-frame, the temporal information will be missed and huge computation cost is needed since there are a large amount of video frames. DANorm [64] enables the model to learn features between source domain and target domain by constraining the vectors in feature subspace, but its performance is far from satisfactory. In the dark video action recognition task, these methods are either computational heavy and cannot be applied, or the performance is unsatisfactory. The proposed DTCM solves the above defects and achieves good performance on both computational efficiency and accuracy for dark video action recognition.

III. Proposed Method

In this section, we will describe the proposed DTCM, which combines dark video enhancement with dark video action recognition jointly in a one-stage framework, to learn the optimal spatio-temporal representation for action recognition in dark videos end-to-end.

A. Motivation

There are three main difficulties that prevent the optimization of dark video enhancement for promoting human action recognition, instead of for boosting the dark video visual quality.

Spatio-temporal consistency preservation. The previous dark enhancement methods cause serious spatio-temporal instability, which harms the learning of meaningful spatio-temporal features from successive video frames for dark video based action recognition.

Suitable training strategy. The two-stage training strategy breaks the connection between enhancement and recognition. Besides, inappropriate end-to-end training settings fall into learn useless features easily, even causing gradient explosion.

Expensive hardware and consumption cost. The prior dark image enhancement methods [28], [36] usually require huge memory consumption and high computation cost, due to a mass of video frames and large image spatial size require to be processed.

As shown in Fig. 2, we design a one-stage framework – DTCM, which cascades recognition-oriented dark enhancement with an action classifier to solve the above problems uniformly. Specifically, our DTCM contains three main components: (1) Spatio-temporal consistency preservation, where a spatio-temporal consistency loss is exploited to encourage the temporal coherence of the enhanced dark video frames. (2) Interactive optimization for dark video enhancement and action classification, where an effective end-to-end training strategy in the one-stage manner is explored to optimize them jointly. (3) Dark enhancement model lightening and redundant computation reduction, where the illumination invariant assumption is applied to exploit the temporal correlation between video frames for improving the enhancement efficiency.

B. Spatio-temporal Consistency

Spatial consistency loss L_{SC}, which encourages spatial coherence of the enhanced dark video frames by preserving the luminance difference of spatial neighboring regions between the input video clip and its enhanced version:

$$L_{SCF}(DY, DI) = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\Omega(i)}(|Y_i - Y_j| - |P_i - P_j|)^2, \tag{1}$$

$$L_{SC} = \frac{1}{T}\sum_{t=1}^{T} L_{SCF}(DY_t, DI_t), \tag{2}$$

where DI denotes the input dark video frame and DY represents the enhanced version. We indicate Y and P as the average intensity value of the local spatial region in DY and DI, respectively. T represents the number of input video clip duration, K is the number of local regions, and $\Omega(i)$ denotes the eight neighboring regions (top, down, left, right, top-left, top-right, down-left, down-right) that are centered at the region i. The size of the local region is set to 4×4 as Zero-DCE [28]. Compared with Zero-DCE [28], we use more neighboring regions here (8 vs 4) to improve the spatial consistency before and after enhancement.

Temporal consistency loss L_{TC} encourages temporal coherence of the enhanced dark video frames via keeping the RGB-difference of adjacent frames between the input video clip and its enhanced version:

$$L_{TC} = \frac{1}{T-1}\sum_{t=1}^{T-1} L_{SCF}(|DY_{t+1} - DY_t|, |DI_{t+1} - DI_t|), \tag{3}$$

where T is the number of input video clip duration. We denote DY and DI as the enhanced frame and the input frame, respectively. It should be noticed that $|DY_{t+1} - DY_t|$ is the RGB-Difference of the enhanced frame, and $|DI_{t+1} - DI_t|$ is the RGB-Difference of the dark frame. Importantly, the temporal consistency loss L_{TC} ensures the trained model to generate enhanced video clips, which can mimic the temporal variation of the original dark video clips precisely.

Spatio-temporal consistency loss L_{STC}, which is a combination of the spatial consistency loss L_{SC} and the temporal consistency loss L_{TC}, used for training the dark video enhancement model, is defined as:

$$L_{STC} = L_{SC} + L_{TC} \tag{4}$$

It should be noted that the value of L_{STC} represents the spatio-temporal inconsistency of the enhanced video, and we learning to minimize the inconsistency to achieve spatio-temporal consistency preserving.

C. One-stage joint training

Dark enhancement loss L_{Dark}. The exposure, color constancy, and illumination smoothness should be considered for dark image enhancement together since the original semantics should also be preserved when enhancing the brightness. We use the loss functions that are well designed by Zero-DCE [28], which are expressed as:

$$L_{Dark} = L_{exp} + W_{col}L_{col} + W_{tv_A}L_{tv_A}, \tag{5}$$

where L_{exp} is the Exposure Control Loss, L_{col} denotes the Color Constancy Loss, and L_{tv_A} represents the Illumination Smoothness Loss. W_{col} and W_{tv_A} are the loss weights. In this work, we use the same settings as Zero-DCE [28] *i.e.* $W_{col} = 0.5, W_{tv_A} = 20$.

Action recognition loss L_{AR}, which used here is the standard cross entropy loss:

$$L_{AR} = CrossEntropy(y, \hat{y}), \tag{6}$$

where y is the action category, $\hat{y}$ is the predicted category.

Joint training loss L_{total}. To improve the performance of dark video action recognition with the help of dark video enhancement, we train the model jointly by back-propagating a linear combination of the spatio-temporal consistency loss, the dark enhancement loss, and the cross entropy loss of action recognition through the entire network. In other words, we train our DTCM with the usage of the following integrated loss function:

$$L_{total} = L_{AR} + \alpha(W_{STC}L_{STC} + L_{Dark}), \qquad (7)$$

where W_{STC} is a weight utilized to balance the scale of L_{STC}. The design of W_{STC} makes our DTCM enjoys great flexibility for spatio-temporal consistency learning, the detail will be shown in the last part of section IV-D. α is a scalar weight modulates the influence of dark enhancement. Both W_{STC} and α are determined experimentally.

D. Efficiency Promotion

Both the memory consumption and the computation cost become huge when enhancing a large number of dark video frames. Inspired by Zero-DCE [28], we present an efficient video enhancement method via applying the illumination invariant assumption and designing a lightweight network, which further utilizes the video temporal information for downstream dark video action recognition efficiently.

Illumination invariant assumption. A direct application of the existing low-light image enhancement methods to dark videos often leads to unsatisfactory results and requires high computation cost due to the neglect of the temporal information between neighboring video frames [27]. We try to exploit the correlation of neighboring video frames by proposing the illumination invariant assumption: the illumination of continuous k frames in ΔT time is constant. By applying the illumination invariant assumption, we can reduce redundant computation and speed up the enhancement speed effectively.

For a single low-light image I, we use a lightweight model [28] to learn a series of pre-defined enhancement parameters R.

$$R = F(I), R = \{r_1, r_2, ..., r_8\}, \qquad (8)$$

where r is utilized for iterative estimating the enhanced image, and we conduct 8 iterations here. The enhanced image EI can be estimated with R as follows:

$$EI = F_R(I) \qquad (9)$$

To promote the enhancement efficiency, based on the illumination invariant assumption, the estimated R is shared for the consecutive k video frames for enhancement as Eq. (10):

$$EI_i = F_R(I_i), i = 1, 2, ..., k \qquad (10)$$

Moreover, in our practice, to best maintain the illumination invariant assumption, the keyframe (*i.e.* the middle frame of the k frames' sequence) is selected for the estimation of R as Eq. (11):

$$R = F(I_{k//2}), \qquad (11)$$

In this way, compared to the common method, the computation and the memory occupation reduced to $1/k$ theoretically, and the temporal consistency is well guaranteed.

Lightweight enhanced network. We adopt the Zero-DCE [28] as the baseline for dark frame enhancement due to its simplicity and powerful performance. However, we find that Zero-DCE [28] causes huge memory consumption in the end-to-end dark action recognition training, because it only uses the simplest plain 3×3 CNN layer. To overcome this drawback, we lightweight the Zero-DCE [28] network, where the detailed network architecture is shown in Fig. 3. For the first convolutional layer, is a plain 3×3 CNN layer, which consists of 32 convolutional kernels with size 3×3 and stride 1 followed by the ReLU activation function. For Layers 2-4, we utilize the ShuffleNetV2 [65] basic module design. For Layers 5-6, we apply the ShuffleNetV2 [65] Spatial Down Sampling module design. The last convolutional layer is a plain 3×3 CNN layer followed by the Tanh activation function. Compared to the original Zero-DCE [28] network, our method uses only 28.92% (22.97K vs. 79.42K) trainable parameters and 28.87% (1.51G vs. 5.21G) Flops while can process an input image with size $256 \times 256 \times 3$.

E. Action Classifier.

In our DTCM, the action classifier is jointly trained with the dark video enhancement network in an end-to-end manner. Notably, our DTCM is flexible to be used for any action recognition network (e.g. TSM [66], 3D-ResNet [29], Video-Swin [31]). Following [29], we select 3D-ResNeXt-101 as the action classifier, due to its promising performance on various datasets [67]. The structure of 3D-ResNeXt-101 is analyzed in Table I for the reference.

TABLE I
THE ILLUSTRATION OF OUR USED ACTION CLASSIFIER 3D-RESNEXT101 [29]. NOTE THAT BOTH THE KERNEL SIZE AND THE OUTPUT SIZE ARE $T \times W \times H$

Stage	Layer		Output size
raw	–		$64 \times 112 \times 112$
conv$_1$	$7 \times 7 \times 7, 64$, stride 1,2,2		$64 \times 56 \times 56$
maxpool$_1$	$3 \times 3 \times 3$, stride 2,2,2		$32 \times 28 \times 28$
res$_2$	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 3 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 256 \end{bmatrix}$	$\times$ 3	$32 \times 28 \times 28$
res$_3$	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 3 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 512 \end{bmatrix}$	$\times$ 4	$16 \times 14 \times 14$
res$_4$	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ 3 \times 3 \times 3, 512 \\ 1 \times 1 \times 1, 1024 \end{bmatrix}$	$\times$ 23	$8 \times 7 \times 7$
res$_5$	$\begin{bmatrix} 1 \times 1 \times 1, 1024 \\ 3 \times 3 \times 3, 1024 \\ 1 \times 1 \times 1, 2048 \end{bmatrix}$	$\times$ 3	$4 \times 4 \times 4$
adaptive average pool, fc			$1 \times 1 \times 1$

F. Video darkness evaluation.

We explore a simple and effective video darkness evaluation method to assess the darkness of a video. Given a video with T_N frames $V(T) = I_1, I_2, ..., I_{T_N}$. The video frame is denoted as $S_t(x, y), x = 1, 2, ..., M, y = 1, 2, ..., N.$, where S_t denotes the pixel value and M, N represent the height and width of the video frame, respectively. The darkness of a video is first identified by thresholding the statistical quantity:

$$I_t^m = \frac{\sum_x \sum_y S_t(x, y)}{M \times N} \qquad (12)$$

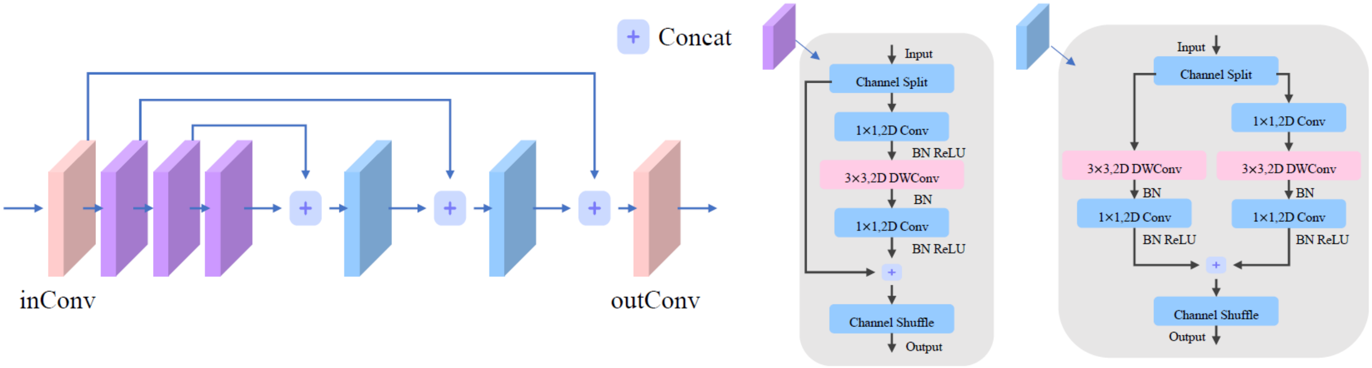

Fig. 3. The architecture of the dark video enhancement model. The inConv and outConv denote the first and last convolutional layers, respectively. DWConv represents the depthwise convolution.

$$Dv = \frac{1}{T_N}\sum_{t=1}^{T_N}\frac{S_t^m - S_c}{S_c}, \tag{13}$$

S_c is a constant, which is defined as the expectation value of the global average brightness value of the normal bright frames. Following [68], we set $S_c = 112$ for 8-bit images. In practice, we do not use all video frames but adopt a segment-based sampling strategy. Specifically, we first divide a video evenly into 8 segments and then randomly select a frame from each segment for evaluation.

The input video is judged as dark if $Dv < -\tau$, and bright if $Dv >= -\tau$. Where τ is the threshold, and τ is set experimentally in consideration of the average darkness of the ARID dataset.

$$F(Dv) = \begin{cases} 1 & Dv < -\tau \\ 0 & Otherwise \end{cases} \tag{14}$$

The proposed video darkness evaluation method, which helps selecting the extremely dark videos to form our dataset Dark-48 by assessing the video darkness of the existing action recognition datasets.

IV. Experiments

In this section, we test the performance of the proposed DTCM method by conducting extensive experiments. First, we introduce the video datasets and the implementation details. Importantly, the dataset Dark-48, which is collected by us from the existing video action recognition datasets with extremely dark videos cover more action classes, is presented. Second, we compare our DTCM with the state-of-the-arts. Third, the ablation studies are performed on the ARID [7] dataset split-1 for analyzation. Finally, some visual results are shown to further illustrate our DTCM method.

Dark video datasets. ARID [7] includes 3784 videos with 11 action categories, and all of them are collected in low-light environments. The ARID dataset is used for the 4th UG2+ Workshop and attracted a lot of attention. The dataset UAVHuman-Fisheye [32] contains 22,476 videos and nearly half of them are dark. Accordingly, for our experiments, we prefer the ARID dataset to the UAVHuman-Fisheye dataset, since ARID is specially designed for recognizing human actions in dark videos.

TABLE II
STATISTICS COMPARISON OF THE DARK ACTION RECOGNITION DATASETS ARID AND OUR DARK-48. 'ACTIONS', SPECIFIES THE NUMBER OF ACTION CLASSES; 'SCENES', THE NUMBER OF SCENES THAT THE VIDEOS ARE COLLECTED IN; 'TOTAL', IS THE TOTAL NUMBER OF CLIPS.

Dataset	Actions	Scenes	Total
ARID	11	18	3784
Dark-48	48	≥ 40	8815

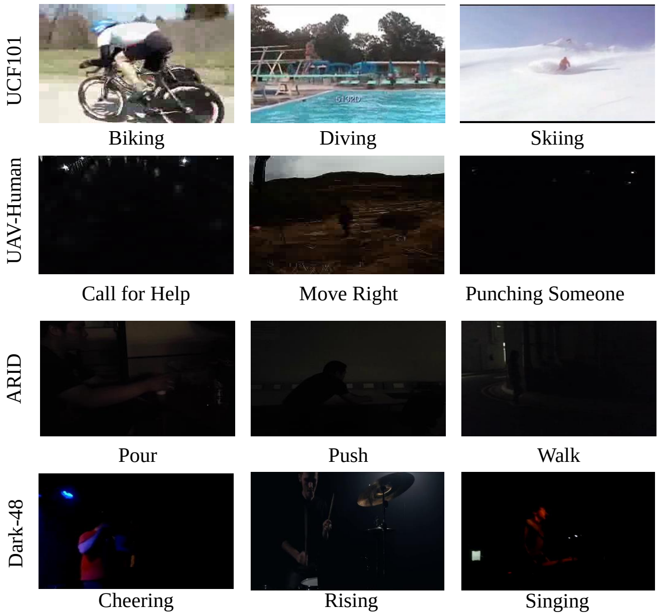

Fig. 4. Example classes from the UCF101, UAV-Human, ARID and our Dark-48 datasets. Note that the examples in UCF101 are normal bright videos.

Dark-48. We find that there are some limitations of the ARID dataset, such as the videos amount is not large (3784 videos) and the action classes is not abundant (11 classes). Besides, the dark semantic information is not rich enough since the videos are collected in only 18 scenes [7]. However, it is difficult to build a completely new dark video dataset with much larger number of videos and action categories. As a compromise, we collect the useful dark videos from current benchmark datasets in an economical way. To align with the darkness of the ARID dataset, first, we evaluate the

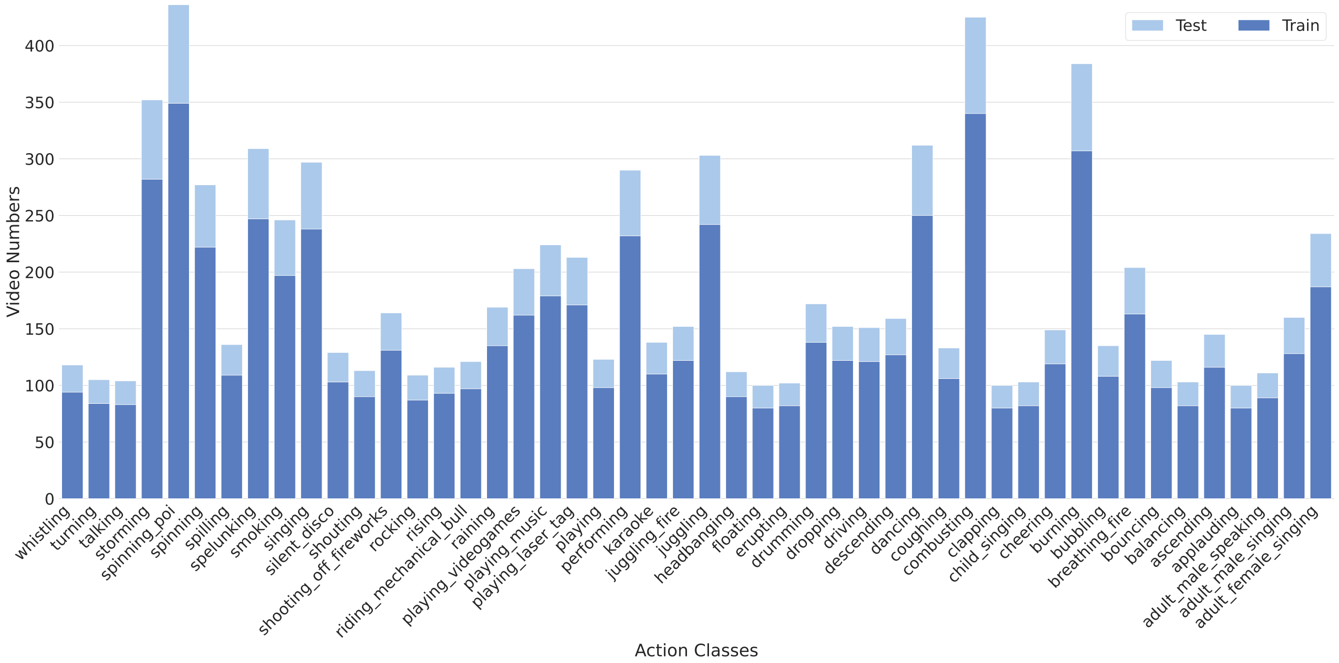

Fig. 5. The video distribution for all action classes in the proposed dataset Dark-48. The pastel blue and muted blue bars indicate the number of videos in the training and testing sets, respectively. The ratio of the training set to the testing set is 4:1. Each action class in our Dark-48 contains at least 100 dark videos. There are three splits and we show split 1 here.

darkness of every ARID video by using our designed video darkness evaluation measure E.q 13, and set τ to be the average darkness of all ARID videos ($\tau = 0.877$). Second, we access the darkness of the videos from Kinetics700 [45] and MiT [44], and select the dark videos via E.q 14. Third, we count the categories of the selected dark videos and keep the categories that have more than 100 dark videos. Fourth, the training and testing sets are partitioned by splitting the videos with 80% to the training set, and the remaining 20% to the testing set. Following UCF101 [69], we also make three training/testing splits. In this way, the dark video dataset Dark-48 is constructed, which contains 8815 dark videos belong to 48 action classes. Some example classes in extremely dark conditions from the UCF101, UAV-Human, ARID and our Dark-48 datasets are shown in Figure 4. The statistic comparison between our Dark-48 and ARID is reported in Table II, and the distribution of videos among all action classes in Dark-48 is displayed in Fig. 5.

Training. In our experiments, the dense sampling strategy [29] is adopted to sample $T = 16$ or $T = 64$ frames from each dark video. For ARID [7], each video frame is resized to 320×240; For UAVHuman [32], each fisheye video frame is center cropped to 128×128 to degrade the effect of image distortion. For all datasets, a crop of 112×112 is randomly performed for the training inputs. The model is fine-tuned from the Kinetics400 pre-trained weights. The batch size is set to 4, the initial learning rate is 1e-5, and the total training epoch is set as 50. AdamW [80] is applied to optimize the parameters.

Testing. Following the common practise [66], [81], we choose the efficient testing scheme named **1-clip and center-crop**, where only a single clip with frames center-cropped to 112×112 is used for evaluation. We do not utilize the computation expensive protocol [81], due to the testing results from the efficient protocol are good enough. For ARID [7], the average accuracy of Top-1 and Top-5 on the three splits are reported. For UAVHuman-Fisheye [32], following the common practise [32], where the accuracy of top-1 and top-5 on the split-1 are reported.

A. *Comparison with the state-of-the-arts*

We extensively compare our DTCM with the state-of-the-art methods on the datasets of ARID [7], UAVHuman-Fisheye [32] and our Dark-48. We mainly report the performance of the proposed DTCM with 64-frame inputs.

In Table III, the comparison and analysis on the ARID dataset is presented. In the upper part, the results are copied directly from the ARID [7] benchmark. The results of DarkLight [1], Delta Sampling Resnet-BERT [21] and MRAN [21] are directly copied from the corresponding papers. Other results are our reproduction. Compared to the previous best results of DarkLight-R(2+1)D, we only use 6.4% GFLOPs and 71.3% parameters, while obtaining 2.32% improvement on the Top-1 accuracy (96.36% vs 94.04%). In addition, we finetune the transfomer-based models [30], [31], [77] and compare them with our DTCM. Although the transformer-based methods show good performance, for the Top-1 accuracy, our DTCM surpasses TimeSformer-L [30] by 14.97% (96.36% vs. 81.39%), Video-Swin-B by 6.57% (96.36% vs. 89.79%), and MViT-B by 4.93% (96.36% vs. 91.43%), respectively. The DTCM model is less computationally intensive (43.24 vs. 2380) and has fewer parameters (47.47 vs. 121.41) compared to the TimeSformer-L model. It also requires less computation

TABLE III
COMPARISON WITH THE STATE-OF-THE-ARTS ON THE ARID DATASET. "INPUTS" INDICATES # TEMPORAL SIZE(T) × # SPATIAL SIZE(H × W,H=W). MIT DENOTES MOMENTS IN THE TIME DATASET [44]. K400 REPRESENTS THE KINETICS400 DATASET [4], AND K700 IS THE KINETICS700 DATASET [45].

Models	Inputs	Pretrained	GFLOPs	Params(M)	Top-1(%)	Top-5(%)
VGG(Two-stream) [70]	16×224^2	ImageNet	-	138	32.08	90.76
TSN(Two-stream) [71]	16×224^2	ImageNet	-	-	57.96	94.17
I3D(Two-stream) [4]	16×224^2	ImageNet	-	12.29	72.78	99.39
C3D [72]	16×224^2	Scratch	132.21	181.12	40.34	94.17
Separable-3D [73]	16×224^2	-	-	-	42.16	93.44
3D-ShuffleNetV1 [74]	16×224^2	K400	0.77	0.96	50.18	94.17
I3D-RGB [4]	16×224^2	ImageNet	27.83	12.29	68.29	97.69
3D-ResNet-18 [29]	16×224^2	K400	32.90	1.84	54.68	96.60
3D-ResNet-50 [29]	16×224^2	K400	39.98	46.22	71.08	99.39
3D-ResNet-101 [29]	16×224^2	K400	55.29	85.27	71.57	99.03
Pseudo-3D-199 [75]	16×224^2	K400	-	-	71.93	98.66
3D-ResNext-101 [29]	16×224^2	K400	38.24	47.54	74.73	98.54
3D-MobilenetV1 [74]	64×112^2	K400	0.95	3.31	40.37	87.13
3D-MobilenetV2 [74]	64×112^2	K400	2.19	2.37	62.92	96.19
3D-ShufflenetV1 [74]	64×112^2	K400	0.78	0.96	46.07	91.14
3D-ShufflenetV2 [74]	64×112^2	K400	0.76	1.31	53.84	95.49
3D-Squeezenet [74]	64×112^2	K400	3.66	1.84	47.47	92.01
R(2+1)D [76]	64×112^2	K700	305.51	63.50	62.87	96.64
I3D-RGB [4]	64×112^2	ImageNet	27.98	12.29	76.25	98.85
3D-ResNet-18 [29]	64×112^2	K700+MiT	33.28	33.21	64.02	96.18
3D-ResNet-50 [29]	64×112^2	K700+MiT	40.41	46.22	78.26	97.98
3D-ResNet-101 [29]	64×112^2	K700+MiT	55.72	85.27	81.11	98.74
3D-ResNext-101 [29]	64×112^2	K400	38.47	47.54	86.36	99.52
DarkLight-ResNext101 [1]	64×112^2	K400	141.14	97.97	87.27	99.47
DarkLight-R(2+1)D [1]	64×112^2	IG-65M	674.84	66.73	94.04	99.87
Delta Sampling Resnet-BERT [21]	64×112^2	K400	-	-	90.46	-
ACAN [19]	64×112^2	K400	-	-	58.00	-
MRAN [21]	64×112^2	K400	-	-	93.73	-
TimeSformer-L [31]	8×224^2	K400	2380	121.41	81.39	98.26
Video-Swin-B [31]	32×224^2	K400	282	88.13	89.79	99.53
MViT-B, 64×3 [77]	64×224^2	K400	455	36.65	91.43	99.72
DTCM	16×224^2	K400	43.01	47.57	82.38	98.93
DTCM	64×112^2	K400	43.24	47.57	**96.36**	**99.92**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ARTS ON THE UAVHUMAN-FISHEYE DATASET.

Models	Inputs	Pretrained	GFLOPs	Params(M)	Top-1(%)	Top-5(%)
3D-MobilenetV1 [74]	64×112^2	K400	0.95	3.31	2.97	10.89
3D-MobilenetV2 [74]	64×112^2	K400	2.19	2.37	5.93	18.04
3D-ShufflenetV1 [74]	64×112^2	K400	0.78	0.96	5.49	16.48
3D-ShufflenetV2 [74]	64×112^2	K400	0.76	1.31	-	-
3D-Squeezenet [74]	64×112^2	K400	3.66	1.84	6.95	19.53
3D-ResNet-18 [29]	64×112^2	K400	33.28	33.21	18.25	36.28
3D-ResNet-50 [29]	64×112^2	K400	40.41	46.22	20.16	36.94
3D-ResNet-101 [29]	64×112^2	K400	55.72	85.27	21.78	38.74
3D-ResNext-101 [29]	64×112^2	K400	38.47	47.54	22.30	42.27
GT-I3D [32]	64×112^2	K400	-	-	23.24	-
DTCM	64×112^2	K400	43.24	47.57	**27.43**	**45.28**

than the Video-Swin-B and MviT-B models (43.24 vs. 282 and 43.23 vs. 455) and achieves better results. Noticeably, ACAN [19] and MRAN [21] are cross-domain adaptation methods and they require additional data from the dataset HMDB-51 [82] for training. Even with no additional data used, our DTCM outperforms MRAN [21] by 2.63% (96.36% vs. 93.73%) in accuracy.

In Table IV, the comparison and analysis on the UAVHuman-Fisheye dataset is presented. Remarkably, our DTCM also achieves the best performance, where its Top-1 accuracy is 27.43% and Top-5 accuracy is 45.28%, which significantly exceeds the current advanced method GT-I3D [32] by 4.19% (27.43% vs 23.24%) on the Top-1 accuracy.

In Table V, the comparison and analysis on our constructed Dark-48 dataset is presented. Since the dark videos in Dark-48 are collected from the previous benchmarks MiVT [44] and Kinetics700 [45], the models, which are pretrained on MiVT and Kinetics700, may have remembered their ground truth. To make a fair comparison, we retrain these models on the Kinetics400 dataset with the validation dark videos removed. As we introduced in Section IV that Dark-48 is very challenging, the 2D CNN-based and 3D CNN-based action recognition methods have poor performance. Specifically, the representative methods TSM [66] has only 36.46% Top-1 accuracy, and SlowFast [79] has only 39.58% Top-1 accuracy. In contrast, the transformer-based methods obtain

TABLE V
COMPARISON WITH THE STATE-OF-THE-ARTS ON OUR CONSTRUCTED DARK-48 DATASET. "K400$*$" DENOTES THAT THE MODEL IS PRE-TRAINED ON THE FILTERED K400 VERSION.

Models	Inputs	Pretrained	GFLOPs	Params(M)	Top-1(%)	Top-5(%)
3D-MobilenetV1 [74]	64×112^2	K400$*$	0.95	3.31	23.37	52.34
3D-MobilenetV2 [74]	64×112^2	K400$*$	2.19	2.37	24.92	56.19
3D-ShufflenetV1 [74]	64×112^2	K400$*$	0.78	0.96	26.71	58.17
3D-ShufflenetV2 [74]	64×112^2	K400$*$	0.76	1.31	27.42	62.43
3D-Squeezenet [74]	64×112^2	K400$*$	3.66	1.84	28.45	62.11
R(2+1)D [76]	64×112^2	K400$*$	305.51	63.50	31.74	64.42
I3D-RGB [4]	64×112^2	ImageNet	27.98	12.29	32.25	65.35
3D-ResNet-50 [29]	64×112^2	K400$*$	40.41	46.22	34.26	66.82
3D-ResNet-101 [29]	64×112^2	K400$*$	55.72	85.27	36.11	68.74
3D-ResNext-101 [29]	64×112^2	K400$*$	38.47	47.54	37.23	68.86
TSN [71]	8×224^2	K400$*$	33.43	24.3	26.04	58.82
TSM [66]	8×224^2	K400$*$	33.47	24.3	36.46	67.26
TIN [78]	8×224^2	K400$*$	32.96	23.90	33.38	64.97
SlowFast16+64 [79]	64×224^2	K400$*$	234.32	32.93	39.58	69.19
TimeSformer-L [30]	8×224^2	K400$*$	2380	121.41	43.27	74.62
Video-Swin-B [31]	32×224^2	K400$*$	282	88.13	41.92	72.51
MViT-B, 64×3 [77]	64×224^2	K400$*$	455	36.65	40.37	70.91
DarkLight-ResNext101 [1]	64×112^2	K400$*$	141.14	97.97	42.27	70.47
DTCM	64×112^2	K400$*$	43.24	47.57	**46.68**	**75.92**

higher accuracy, e.g., TimeSformer-L [30] gets 43.27% Top-1 accuracy, Video-Swin-B [31] gets 41.92% Top-1 accuracy, and MViT-B [77] gets 40.37% Top-1 accuracy. However, these methods require much higher computation cost. Importantly, our DTCM outperforms all of them by a large margin, which achieves 46.68% Top-1 accuracy that surpasses the previous best method TimeSformer-L [30] by 3.41%.

B. Ablation studies

One-stage joint training. To illustrate the effectiveness of our one-stage joint training strategy, we compare it with the prior two-stage setting. In particular, the two-stage training is generally conducted as follows: pre-enhance the dark video frames first, then use the enhanced video frames for action recognition training. In contrast, the one-stage joint training of us can be implemented in two types of settings: one is to optimize the entire network with the usage of only the action recognition loss L_{AR} (see Eq. 6), and the other utilizes the joint loss L_{total} (see Eq. 7). As shown in Table VI, compared to the two-stage training setting, our one-stage joint training strategy boosts the performance by at least 1.37% (86.20% vs 84.83%) in the Top-1 accuracy. Remarkably, when using the well-designed L_{total} loss, the Top-1 accuracy is largely improved by 10.43% (95.26% vs 84.83%).

TABLE VI
PERFORMANCE OF DIFFERENT TRAINING SETTINGS ON ARID.

Settings		Top-1(%)	Top-5(%)
two-stage		84.83	92.76
one-stage	L_{AR}	86.20	94.64
	L_{total}	95.26	98.53

Selection of α. α determines the impact of the dark video enhancement component in our DTCM. Increasing the value of α means to increase the impact of dark enhancement. Different α (e.g. $\alpha = 10^n, n = -2, -1, 0, 1$) are tried for finding the most suitable one. As shown in Table VII, when α is too small, the increased useful information contained in the enhanced frames is limited, leading to the action recognition accuracy improvement is low. When α is too large, the action recognition accuracy improvement is also low, because of the model enhances too much unnecessary information. Eventually, we choose $\alpha = 0.1$ for our method DTCM.

TABLE VII
PERFORMANCE OF DIFFERENT α ON THE ARID DATASET. WE INCREASE THE VALUE OF α EXPONENTIALLY.

α	0.01	0.1	1	10
Top-1(%)	95.6	**96.8**	96.3	95.3

Selection of W_{STC}. W_{STC}, which is the weight of the spatio-temporal consistency loss, determines the spatio-temporal smoothness of the enhanced video frames. We empirically test different W_{STC} values *i.e.* $W_{STC} = 2n, n = 0, 1, 2, 3, 4, 5$. As shown in Table VIII, the spatio-temporal smoothness is insufficient when W_{STC} is too small, resulting in low action recognition accuracy improvement. When W_{STC} is too large, the action recognition accuracy improvement is also low, due to the disturbance of the enhancement process. Accordingly, we select $W_{STC} = 4$. Furthermore, we compare the results of with/without ($W_{STC} \neq 0/W_{STC} = 0$) the spatio-temporal consistency learning, and find that learning the spatio-temporal consistency boosts the action recognition accuracy significantly by 9.1% (96.2% vs. 87.1%) in Top-1 accuracy. The experimental results verify that our designed spatio-temporal consistency loss is crucial and useful for action recognition in dark videos.

TABLE VIII
PERFORMANCE OF DIFFERENT W_{STC} ON THE ARID DATASET.

W_{STC}	0	2	4	6	8	10
Top-1(%)	87.1	96.5	**97.1**	96.5	96.3	96.2

Effect of different k. We share the dark enhancement parameters in k neighboring video frames, and test $k =$

$2^n, n = 0, 1, ..., 4$ experimentally. The computation efficiency and the accuracy, which are compared with the baseline Zero-DCE [28], are shown in Table IX. In contrast to the baseline, our method reduces the number of parameters by about 70% (23.77k vs 79.42k), while boosts the dark action recognition accuracy larger, where the Top-1 accuracy is modified by 3.3% (+6.7% vs +3.4%). The variance of the action recognition accuracy improvement is not large when $k <= 8$. However, the performance drops sharply when $k > 8$, which could attribute to the destruction of the illumination invariant assumption. Particularly, compared to the baseline, when $k = 4$, the accuracy of our DTCM method is boosted significantly by 3.1% (+6.5% vs +3.4%) while with needing only 7.5% GFLOPs. Consequently, we set $k = 4$ for accuracy and computation efficiency trade-off to our DTCM method.

TABLE IX
COMPUTATION CONSUMPTION AND ACCURACY COMPARISON WITH THE USAGE OF DIFFERENT k ON THE ARID DATASET. REMARKABLY, OUR DTCM METHOD DECREASES GFLOPS, PARAMS, AND MEMORY FOOTPRINT LARGELY.

Settings	GFLOPs	Params(k)	Mem(M)	Δ Top-1(%)
Baseline	63.77	79.42	5269	+3.4
Ours(k=1)	19.08	23.77	3037	+6.7
Ours(k=2)	9.55	23.77	2432	+6.4
Ours(k=4)	4.77	23.77	2073	+6.5
Ours(k=8)	2.38	23.77	1737	+6.3
Ours(k=16)	1.19	23.77	1663	+3.7

C. Generalization

The proposed DTCM has a strong generalization capacity, which can use various action recognition network as the action classifier. Table X shows the Top-1 accuracy with the action classifiers of TSM [66], 3D-ResNet50 [29], I3D [4], and Video-Swin-B [31] on the ARID dataset. Compared to these baseline backbones, our DTCM improves the accuracy by at least 2.08% (78.33% vs 76.25%) consistently.

TABLE X
GENERALIZATION CAPACITY OF THE PROPOSED DTCM METHOD ON DIFFERENT 3D-CNN BACKBONES.

	TSM	MobileV2	Res50	I3D	Swin-B
Baseline	61.34	47.92	74.26	76.25	89.79
DTCM	63.75	53.70	80.05	78.33	92.35
$\Delta Top-1$	+2.41	+5.78	+5.79	+2.08	+2.56

D. Empirical Analysis

To validate whether our DTCM method can maintain spatio-temporal consistency to benefit dark video based action recognition, several empirical analyses are conducted.

TABLE XI
PERFORMANCE OF DIFFERENT ENHANCEMENT METHODS USED FOR DARK VIDEO ACTION RECOGNITION ON THE ARID DATASET. 3D-RESNEXT-101 IS USED HERE AS THE ACTION CLASSIFIER.

Methods	Spatio-temporal inconsistency	Top-1(%)
No Enhancement	-	75.21
BIMEF [23]	0.45	72.49(-2.72)
LIME [36]	0.62	73.67(-1.54)
KinD [37]	0.53	70.53(-4.68)
GIC [33]	0.08	78.46(+3.25)
StableLLVE [26]	0.10	79.62(+4.41)
DTCM(Ours)	0.15	**82.74(+7.53)**

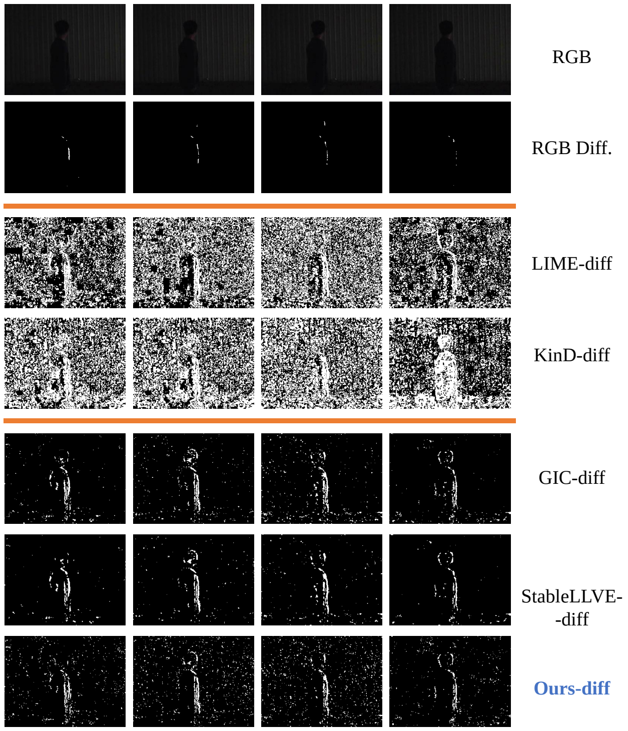

Fig. 6. RGB-Difference comparison of different dark enhancement methods. (1) The upper part shows the dark video frames on the ARID dataset and their corresponding RGB-Difference without conducting dark enhancement. (2) The middle part shows the RGB-Difference of methods LIME [36] and KinD [37], which are failed to improve the action recognition performance in dark videos. (3) The lower part shows the RGB-Difference of the methods GIC [33], StableLLVE [26] and our DTCM, which significantly improve the performance of action recognition in dark videos.

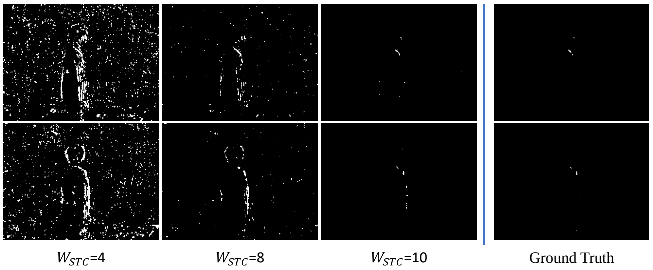

Fig. 7. DTCM enjoys great flexibility in learning spatio-temporal consistency by adjusting W_{STC}.

Effect of spatio-temporal inconsistency. To reveal the relationship between spatio-temporal inconsistency of dark video enhancement and action recognition performance, we calculate the spatio-temporal inconsistency value of the videos that are enhanced by different dark enhancement strategies in the ARID dataset by E.q 4. The results are shown in Table XI. It can be seen that the methods *e.g.* BIMEF, LIME, and KinD,

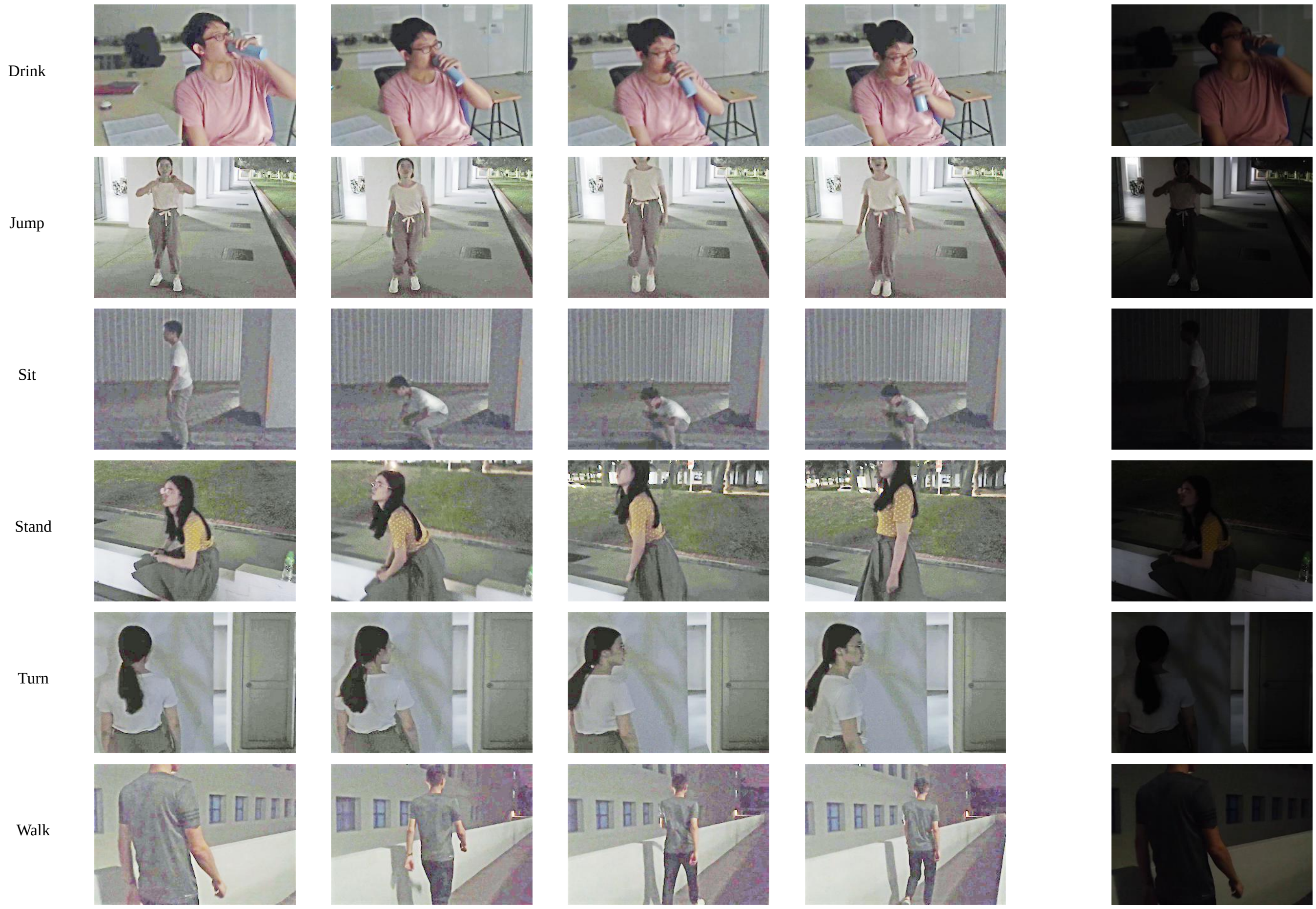

Fig. 8. Examples of the proposed DTCM method for dark video enhancement. In each row, the left four video frames, which are recovered by our dark video enhancement strategy, show one human action that performs from start to end, and the right one is the original dark video frame (start) that is used for comparison. These videos are selected from the ARID [7] dataset.

which have high spatio-temporal inconsistency, are failed to improve the Top-1 accuracy of dark video based action recognition. In contrast, the methods *e.g.* GIC, StableLLVE, and our DTCM, which have maintained the spatio-temporal consistency well, are able to improve the Top-1 accuracy of dark video based action recognition successfully.

Besides, we also find that too strong spatio-temporal consistency constraint would lead to performance degradation, *i.e.* GIC has lower spatio-temporal inconsistency than StableLLVE (GIC:0.08 vs StableLLVE:0.10), but StableLLVE gets higher performance modification (StableLLVE:+4.41 vs. GIC:+3.25). Remarkably, the proposed DTCM method enjoys great flexibility in controlling the spatio-temporal consistency constraint by adjusting W_{STC}. Benefiting from the maintenance and flexible adjustment capacity on spatio-temporal consistency, our DTCM achieves the best performance, *e.g.* its Top-1 accuracy is increased by 7.53% for the upmost.

Spatio-temporal consistency visualization. We compare the RGB-Difference of several dark image enhancement methods in Fig. 6 to further explore the importance of spatio-temporal consistency for video action recognition in dark. LIME and GIC are the traditional methods, and KinD and StableLLVE are the deep learning-based methods. Methods in the middle part, which produce confused RGB-difference (chaotic and unclear), fail to improve the action recognition performance in dark videos. Methods in the bottom part, *i.e.* GIC, StableLLVE, and our DTCM, which generate clear RGB-Difference, success to improve the action recognition performance in dark videos. The clean static background and clear motion boundaries indicating that the spatio-temporal consistency is well preserved in these methods. In summary, only the dark enhancement methods, which can maintain the spatio-temporal consistency of the dark video, are beneficial for dark video action recognition.

Flexibility of DTCM in preserving spatio-temporal consistency. Fig. 7 shows the RGB-Difference of the proposed DTCM with different W_{STC}. The rightmost side is the RGB-Difference of the raw dark video frames (*i.e.* the ground truth). DTCM can gain different spatio-temporal smoothness by adjusting the value of W_{STC} to select the most suitable one, this is consistent with our analysis in Table XI.

E. Visual quality analysis

Fig. 8 reveals that the proposed DTCM can not only boost the accuracy of action recognition but also greatly enhance the visibility of the low-light video. In particular, DTCM generates admirable visual results, where the enhanced dark video frames are clear and bright, the color is uniform and the border is obvious. These advantages promote the usage range of our DTCM for real-world application, since there is a great demand to improve both the action recognition accuracy and the low-light video's visual quality, *e.g.* video surveillance.

V. Conclusion

In this work, we have presented a unified, single-stage training framework, termed DTCM, specifically designed for effective action recognition in low-light video scenarios. The core contributions of our DTCM are that 1) optimizing dark video enhancement and human action classification interactively in a one-stage pipeline with fast speed, 2) developing a process that ensures the preservation of spatio-temporal consistency in the enhanced frames of dark videos, and 3) empowering a video-based action recognition model, which is originally pre-trained on bright videos, to extract and learn valuable spatio-temporal features directly from low-light video scenarios. Benefiting from the newly designed network architecture and the particularly formulated loss functions, our DTCM enjoys high flexibility, efficiency, and accuracy. Extensive experiments on benchmark datasets demonstrate that the proposed DTCM method outperforms the state-of-the-arts on recognizing human actions in dark videos for a large margin.

Acknowledgment

This work was supported by the Natural Science Fund for Distinguished Young Scholars of Hubei Province under Grant 2022CFA075, and the National Natural Science Foundation of China under Grant 62106177. The numerical calculation was supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] R. Chen, J. Chen, Z. Liang, H. Gao, and S. Lin, "Darklight networks for action recognition in the dark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 846–852.

[2] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3633–3645, 2014.

[3] J. Zheng, Z. Jiang, and R. Chellappa, "Cross-view action recognition via transferable dictionary learning," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2542–2556, 2016.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4724–4733.

[5] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] H. Yang, C. Yuan, L. Zhang, Y. Sun, W. Hu, and S. J. Maybank, "Sta-cnn: Convolutional spatial-temporal attention learning for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 5783–5793, 2020.

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

[8] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, 2016.

[9] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 588–597.

[10] Y. Liu, J. Yuan, and Z. Tu, "Motion-driven visual tempo learning for video-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 4104–4116, 2022.

[11] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, and Y.-D. Zhang, "Multi-domain and multi-task learning for human action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 853–867, 2019.

[12] J. Cai, J. Hu, X. Tang, T.-Y. Hung, and Y.-P. Tan, "Deep historical long short-term memory network for action recognition," *Neurocomputing*, vol. 407, pp. 428–438, 2020.

[13] Z. Tu, X. Liu, and X. Xiao, "A general dynamic knowledge distillation method for visual analytics," *IEEE Transactions on Image Processing*, vol. 31, pp. 6517–6531, 2022.

[14] A. Ulhaq, "Action recognition in the dark via deep representation learning," in *IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 2018, pp. 131–136.

[15] Y. Zheng, M. Zhang, and F. Lu, "Optical flow in the dark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6748–6756.

[16] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[17] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 325–333.

[18] Y. Sasagawa and H. Nagahara, "Yolo in the dark-domain adaptation method for merging multiple models," in *European Conference on Computer Vision*. Springer, 2020, pp. 345–359.

[19] Y. Xu, H. Cao, K. Mao, Z. Chen, L. Xie, and J. Yang, "Aligning correlation information for domain adaptation in action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[20] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *IEEE International Conference on Computer Vision*, 2017, pp. 5843–5851.

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

[22] H. Zhang, D. Liu, and Z. Xiong, "Two-stream action recognition-oriented video super-resolution," in *IEEE International Conference on Computer Vision*, 2019, pp. 8798–8807.

[23] Z. Ying, G. Li, and W. Gao, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," *ArXiv preprint*, 2017.

[24] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1013–1037, 2021.

[25] A. S. Karadeniz, E. Erdem, and A. Erdem, "Burst photography for learning to enhance extremely dark images," *IEEE Transactions on Image Processing*, vol. 30, pp. 9372–9385, 2021.

[26] F. Zhang, Y. Li, S. You, and Y. Fu, "Learning temporal consistency for low light video enhancement from single images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4967–4976.

[27] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1777–1786.

[29] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[30] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *International Conference on Machine Learning*, vol. 139, 2021, pp. 813–824.

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

[32] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 266–16 275.

[33] C. Poynton, *Digital video and HD: Algorithms and Interfaces*. Elsevier, 2012.

[34] P. E. Trahanias and A. N. Venetsanopoulos, "Color image enhancement through 3-d histogram equalization," in *IAPR International Conference on Pattern Recognition*, vol. 1, 1992, pp. 545–548.

[35] U. R. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *IEEE International Conference on Image Processing*, vol. 3, 1996, pp. 1003–1006.

[36] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, 2017.

[37] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *ACM international conference on multimedia*, 2019, pp. 1632–1640.

[38] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[39] W. Wang, X. Wang, W. Yang, and J. Liu, "Unsupervised face detection in the dark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1250–1266, 2023.

[40] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, pp. 32–43, 2018.

[41] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal vlad for video action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2799–2812, 2019.

[42] A. J. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9945–9953.

[43] Y. Yoshikawa, J. Lin, and A. Takeuchi, "Stair actions: A video dataset of everyday home actions," *ArXiv preprint*, 2018.

[44] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 502–508, 2019.

[45] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the kinetics-700-2020 human action dataset," *ArXiv preprint*, 2020.

[46] F. Steinbrücker, T. Pock, and D. Cremers, "Large displacement optical flow computation withoutwarping," in *IEEE International Conference on Computer Vision*, 2009, pp. 1609–1614.

[47] Z. Tu, W. Xie, D. Zhang, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "A survey of variational and cnn-based optical flow techniques," *Signal Processing: Image Communication*, vol. 72, pp. 9–24, 2019.

[48] L. Tian, Z. Tu, D. Zhang, J. Liu, B. Li, and J. Yuan, "Unsupervised learning of optical flow with cnn-based non-local filtering," *IEEE Transactions on Image Processing*, vol. 29, pp. 8429–8442, 2020.

[49] S. I. Young, B. Girod, and D. Taubman, "Fast optical flow extraction from compressed video," *IEEE Transactions on Image Processing*, vol. 29, pp. 6409–6421, 2020.

[50] Z. Tu, H. Li, W. Xie, Y. Liu, S. Zhang, B. Li, and J. Yuan, "Optical flow for video super-resolution: a survey," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6505–6546, 2022.

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

[52] Z. Tu, W. Xie, J. Dauwels, B. Li, and J. Yuan, "Semantic cues enhanced multimodality multistream cnn for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1423–1437, 2018.

[53] S. Liu, K. Luo, N. Ye, C. Wang, J. Wang, and B. Zeng, "Oiflow: Occlusion-inpainting optical flow estimation by unsupervised learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 6420–6433, 2021.

[54] R. R. A. Pramono, W.-H. Fang, and Y.-T. Chen, "Relational reasoning for group activity recognition via self-attention augmented conditional random field," *IEEE Transactions on Image Processing*, vol. 30, pp. 8184–8199, 2021.

[55] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[56] Z. Li, X. Gong, R. Song, P. Duan, J. Liu, and W. Zhang, "Smam: Self and mutual adaptive matching for skeleton-based few-shot action recognition," *IEEE Transactions on Image Processing*, vol. 32, pp. 392–402, 2022.

[57] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[58] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 608–10 617.

[59] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Transactions on Multimedia*, 2023.

[60] M. Liu, F. Meng, C. Chen, and S. Wu, "Novel motion patterns matter for practical skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[61] S. Lee, J. Rim, B. Jeong, G. Kim, B. Woo, H. Lee, S. Cho, and S. Kwak, "Human pose estimation in extremely low-light conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 704–714.

[62] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning*, vol. 80, 2018, pp. 1994–2003.

[63] H. Porav, W. Maddern, and P. Newman, "Adversarial training for adverse conditions: Robust metric localisation using appearance transfer," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 1011–1018.

[64] 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," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 4251–4258.

[65] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *European conference on computer vision*, 2018, pp. 116–131.

[66] J. Lin, C. Gan, and S. Han, "TSM: temporal shift module for efficient video understanding," in *IEEE International Conference on Computer Vision*, 2019, pp. 7082–7092.

[67] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: motion-augmented RGB stream for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7882–7891.

[68] G. Cao, L. Huang, H. Tian, X. Huang, Y. Wang, and R. Zhi, "Contrast enhancement of brightness-distorted images by improved adaptive gamma correction," *Computers & Electrical Engineering*, vol. 66, pp. 569–582, 2018.

[69] 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.

[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[71] 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.

[72] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[73] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *European conference on computer vision*, 2018, pp. 305–321.

[74] O. Köpüklü, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3d convolutional neural networks," in *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1910–1919.

[75] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 5534–5542.

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

[77] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *IEEE International Conference on Computer Vision*, 2021, pp. 6804–6815.

[78] H. Shao, S. Qian, and Y. Liu, "Temporal interlacing network," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 11 966–11 973.

[79] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *IEEE International Conference on Computer Vision*, 2019, pp. 6201–6210.

[80] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[81] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: temporal excitation and aggregation for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 906–915.

[82] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *IEEE International Conference on Computer Vision*, 2011, pp. 2556–2563.

Zhigang Tu received the Ph.D. degree respectively from Wuhan University (China), 2013, and Utrecht University (Netherlands), 2015. From 2015 to 2016, he was a postdoctoral researcher at Arizona State University, US. Then from 2016 to 2018, he was a research fellow at Nanyang Technological University, Singapore. He is currently a professor at Wuhan University. His research interests include computer vision, image processing, video analytics, and machine learning. Special for motion estimation, human action and gesture recognition, and anomaly event detection.

He has co-/authored more than 70 articles on international SCI-indexed journals and conferences. He is an Associate Editor of the SCI-indexed journal *The Visual Computer* (IF=2.835) and a Guest Editor of *Journal of Visual Communications and Image Representation* (IF=2.887), the Area Chair of AAAI2023/2024 and VCIP2022. He is the first organizer of the ACCV2020 Workshop on MMHAU (Japan). He received the "Best Student Paper" Award in the 4^{th} Asian Conference on Artificial Intelligence Technology, and IEEE TCSVT "One of the Three Best Reviewers" Award 2022.

Yuanzhong Liu is currently a postgraduate student at Wuhan University. He received the B. Eng. degree from the school of resources and environment from University of Electronic Science and Technology of China in 2020. His research interests mainly include computer vision and machine learning.

Yan Zhang received the M.D. degree from the Huazhong University of Science and Technology in 2017. She is currently the deputy chief technician of the Department of Clinical Laboratory, Renmin Hospital of Wuhan University. She has co-/authored more than 20 academic journal papers. Her research interests mainly include clinical laboratory medicine, reproductive immunity and artificial intelligence in medicine.

Qizi Mu was born in Datong, Shanxi, China 1996. She received her M.E. degree in computer science and technology from Northeast Electric Power University in 2022. She has been working at CHN ENERGY DIGITAL INTELLIGENCE TECHNOLOGY DEVELOPMENT (BEIJING) CO., LTD since 2022. Her research interests include data mining and artificial intelligence.

Junsong Yuan is Professor and Director of Visual Computing Lab at Department of Computer Science and Engineering, State University of New York at Buffalo (UB), USA. Before that he was Associate Professor (2015-2018) and Nanyang Assistant Professor (2009-2015) at Nanyang Technological University (NTU), Singapore. He obtained his Ph.D. from Northwestern University in 2009, M. Eng. from National University of Singapore in 2005, and B. Eng. from Huazhong University of Science Technology in 2002. His research interests include computer vision, pattern recognition, video analytics, human action and gesture analysis, large-scale visual search and mining. He received Best Paper Award from IEEE Trans. on Multimedia, Nanyang Assistant Professorship from NTU, and Outstanding EECS Ph.D. Thesis award from Northwestern University.

He served as Associate Editor of IEEE Trans. on Image Processing (T-IP), IEEE Trans. on Circuits and Systems for Video Technology (T-CSVT), Machine Vision and Applications, and Senior Area Editor of Journal of Visual Communications and Image Representation. He was Program Co-Chair of IEEE Conf. on Multimedia Expo (ICME'18), and Area Chair for CVPR, ICCV, ECCV, and ACM MM. He was elected senator at both NTU and UB. He is a Fellow of IEEE and IAPR.