DL-KDD: Dual-Lightness Knowledge Distillation for Action Recognition in the Dark

Chi-Jui Chang[1] , **Oscar Tai-Yuan Chen**[1] , **Vincent S. Tseng**[1,2*]
[1]Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University
[2]Department of Computer Science, National Yang Ming Chiao Tung University
{jerryyyyy708.cs12, oscarchen.cs10}@nycu.edu.tw, vtseng@cs.nycu.edu.tw

Abstract

Human action recognition in dark videos is a challenging task for computer vision due to the low quality of the videos filmed in the dark. Recent studies focused on applying dark enhancement methods to improve the visibility of the video. However, such video processing results in the loss of critical information in the original (un-enhanced) video. Conversely, traditional two-stream methods are capable of learning information from both original and enhanced videos, but it can lead to a significant increase in the computational cost. To address these challenges, we propose a novel knowledge-distillation-based framework, named *Dual-Lightness KnowleDge Distillation (DL-KDD)*, which simultaneously resolves the aforementioned issues by enabling a student model to obtain both original features and light-enhanced knowledge without additional complexity, thus improving the performance of the model and avoiding extra computational cost. Through comprehensive evaluations, the proposed *DL-KDD*, with only original video required as input during the inference phase, significantly outperforms state-of-the-art methods on the widely-used dark video datasets. The results highlight the excellence of our proposed knowledge-distillation-based framework for dark video human action recognition.

1 Introduction

Action Recognition is a popular task in computer vision that plays a key role in various applications, like autonomous vehicles [Xu *et al.*, 2022; Tammvee and Anbarjafari, 2021] and intelligent surveillance [Kardas and Cicekli, 2017], etc. Accurately recognizing human actions enables these technologies to function reliably in real-world scenarios. As a result, increasing research has focused on this task in recent years [Wasim *et al.*, 2023; Xian *et al.*, 2024b; Zhang *et al.*, 2024; Xian *et al.*, 2024a]. While action recognition under well-lighted conditions is relatively well-understood, recognizing actions in dark environments is more challenging due to the degradation of the information in the videos. However, many real-world applications, such as nighttime surveillance, rely on action recognition in low-light conditions. In these scenarios, ensuring reliable performance is essential for maintaining safety and functionality under poor visibility environments. Therefore, developing an effective dark action recognition method is crucial for enabling action recognition in real-world applications under low-light conditions.

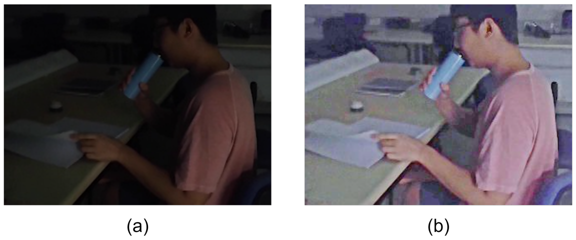

Figure 1: Comparison of original and enhanced video frame: (a) A sampled frame from the ARID dataset. The original frame suffers from low visibility. However, it preserves native features of the action. (b) Corresponding frame enhanced by ZeroDCE. The main subject is more precise due to the enhancement, while some blurring occurs as a side effect. Both frames contain crucial information for action recognition.

In response to this challenge, recent studies [Chen *et al.*, 2021; Singh *et al.*, 2023; Tu *et al.*, 2023] have proposed various frameworks to achieve better performance with dark video inputs. Common approaches include utilizing light enhancement methods such as ZeroDCE [Guo *et al.*, 2020] and Gamma Intensity Correction(GIC) to improve video features and visibility, followed by 3D convolutional networks like R(2+1)D [Tran *et al.*, 2018] or 3D-ResNext [Hara *et al.*, 2018] as the backbone classifiers. Two main architectures to incorporate these components are: 1) directly integrating two models [Singh *et al.*, 2023; Tu *et al.*, 2023], which takes the enhanced video frames generated by the enhancement module as input features as the backbone classification model. 2) using a two-stream method [Tran *et al.*, 2018] to improve the accuracy of action prediction from dark videos.

Recent researches [Singh *et al.*, 2023; Tu *et al.*, 2023] focused on applying enhancements and taking enhanced video as model inputs. While such approaches improve the features contained in videos, the enhancement process often leads to

[*]Corresponding author.

losing original content, which contains critical information for action recognition, such as motion cues and original texture. On the other hand, the work [Chen *et al.*, 2021] considering the importance of the original input in dark video human action recognition applies traditional two-stream [Simonyan and Zisserman, 2014] method, which takes both the original and enhanced video as the inputs to the model. This approach significantly increases the computational load, especially when the input data is video-based, making it slower to perform predictions during inference. In contrast, other methods using only the original video as input result in a performance gap compared to previously mentioned techniques since the model will only get less information from the original video without enhancement.

In summary, the three main challenges for current research on dark video human action recognition are: 1) Information Completeness: Ensuring the model to learn from enhanced video with essential information from the original video preserved. 2) Complexity Tradeoff: Making full use of original video and considering enhanced features (extracted from enhanced video) without additional model complexity for input features. 3) Consistent Performance: Improving the performance even if using only original video without enhancement as input during the inference phase.

To the best of our knowledge, no existing studies fully addressed these three challenges concurrently. To address these challenges, we propose a knowledge-distillation-based framework, *Dual-Lightness KnowleDge Distillation (DL-KDD)* for action recognition in the dark. Without additional computational cost required, the DL-KDD can achieve robust accuracy by taking only original video as input during inference, which simultaneously overcomes the three challenges. Specifically, a student model is used to exclusively process the original video, while a teacher model takes the role of enriching the learning of the student model. It provides augmented features extracted from the enhanced video for the student to learn from. Knowledge distillation [Hinton *et al.*, 2015], in this context, serves as an effective method, helping the model to learn from the information of the teacher model without increasing the input feature set. Thus, during the inference phase, we only need to use the original video as input without any additional enhancement or dual input to achieve state-of-the-art accuracy.

As this is the first work that addresses these three problems as mentioned above simultaneously, the main contributions of this work are threefold:

1. We propose a novel knowledge-distillation-based architecture named *Dual-Lightness KnowleDge Distillation (DL-KDD)*, specially designed for action recognition in the dark. By leveraging both the original video and the enhanced features, our approach facilitates more effective knowledge transfer. Unlike conventional knowledge distillation which primarily focuses on model compression, our method innovatively distills complementary knowledge for the student model, leading to a more robust learning paradigm for action recognition in the dark.
2. Our model can use only original video input without additional features or enhancement during inference, making it more efficient in real-world applications. This also avoids the need for additional decisions regarding the use of the enhancement module in real-world scenarios.
3. We achieve state-of-the-art performance in dark video human action recognition across widely-used datasets with the model complexity maintained. The performance improvement demonstrates the robustness of our proposed framework.

The remaining of this paper is organized as follows. Section 2 reviews action recognition methods and knowledge distillation strategies. Section 3 presents the proposed DL-KDD framework, including the teacher-student architecture, dual-lightness learning, and dynamical loss balancing with alpha decay. Section 4 provides detailed information on the datasets, implementation, experimental results, and comparison with recent dark action recognition methods. Finally, Section 5 concludes the paper and discusses future directions.

2 Related Work

2.1 Video-Based Action Recognition

In recent studies, various model architectures have been proposed for human action recognition. The most common approaches are those based on 3D Convolution Networks (3D-CNNs) [Tran *et al.*, 2015; Feichtenhofer *et al.*, 2019; Hara *et al.*, 2018; Hara *et al.*, 2017]. The success of 3D-ResNet in action recognition demonstrates the effectiveness of deeper networks in video classification. 3D-ResNext [Hara *et al.*, 2018] incorporates the concept of cardinality to further improve the learning process of the model. The split-transform-merge method enables the model to learn more efficiently with the same number of parameters and achieve high performance in action recognition. More recently, Tran *et al.* proposed an architecture known as R(2+1)D [2018], which differs from typical 3D-CNNs. The R(2+1)D decomposes 3D features into 2D and 1D components by employing both 2D and 1D convolutions for spatial and temporal features in the video. This method has improved the feature learning ability of the model and further enhanced the accuracy of action recognition. Transformer-based methods have also been explored recently [Bertasius *et al.*, 2021; Liu *et al.*, 2022; Yang *et al.*, 2022]. These technologies perform well under well-lit conditions for action recognition. However, their performance degrades while facing low-light videos. In our research, we build on the foundation of these action recognition models with training strategy specifically designed for such conditions to enhance the performance in dark/night conditions.

2.2 Action Recognition in the Dark

To overcome the challenge of performance degradation for action recognition in dark environments, innovative approaches have been introduced to address the problem, there are video-based approaches [Chen *et al.*, 2021; Singh *et al.*, 2023; Tu *et al.*, 2023], and cross-modality approaches such as infrared [Jiang *et al.*, 2017; Akula *et al.*, 2018] or skeleton based [Duan *et al.*, 2022; Yan *et al.*, 2018] data, where most of them selected CNN based model as the backbone due to their effectiveness. Although multi-modality data can provide diverse types of features, it also complicate data collection and processing. As a result, recent studies have focused on applying enhancement methods [Guo *et al.*, 2020] to improve model performance with video-based data. [Chen *et al.*, 2021] proposed DarkLight, which utilizes both original and frames enhanced by Gamma Intensity Correction (GIC) for action prediction. The method represents a significant advancement in this field, demonstrating the effectiveness of light enhancement for action recognition in the dark. The experiments also indicate that the dual-path approach, which utilizes both original and enhanced frames, captures more features than methods that estimate optical flow [Carreira and Zisserman, 2017], thereby achieving better performance. Building on these advancements, recent studies [Singh *et al.*, 2023; Tu *et al.*, 2023] have further improved the accuracy of action prediction by incorporating ZeroDCE [Guo *et al.*, 2020], a light enhancement module. These studies have integrated ZeroDCE with backbone classifiers and showed remarkable performance gains. Consequently, the architecture that directly utilizes enhanced video as input has emerged as the most prevalent method in the field.

For the datasets, while there are numerous datasets for action recognition in normal conditions [Carreira and Zisserman, 2017; Soomro *et al.*, 2012; Kuehne *et al.*, 2011; Goyal *et al.*, 2017; Karpathy *et al.*, 2014], there are less datasets for action recognition in dark conditions. [Xu *et al.*, 2021] introduced the ARID dataset, the first dataset for action recognition in the dark. As the first dataset of the task, ARID has become a foundational dataset for several studies in this field [Chen *et al.*, 2021; Singh *et al.*, 2023; Tu *et al.*, 2023] . To enrich the scene of the data, an expanded version called ARID V1.5 with more data collected was later released. More recently, [Tu *et al.*, 2023] highlighted the problem, which is the scarcity of datasets for this topic. They gathered video with extremely low light from multiple sources to propose the Dark-48 dataset, offers a larger number of action videos and more diverse data classes, providing a more challenging benchmark for action recognition in the dark.

Our DL-KDD addresses the issue of overlooking the importance of original video in recent studies, and the increased computational cost brought by two-stream methods by applying knowledge distillation with an enhancement module. We evaluated our model performance with all the three datasets — ARID [Xu *et al.*, 2021] and its updated version ARID V1.5, and Dark-48 [Tu *et al.*, 2023] — in dark video human action recognition to demonstrate the effectiveness of our approach.

2.3 Knowledge Distillation

Knowledge distillation [Hinton *et al.*, 2015] has been applied across various sub-tasks within human action recognition, including cross-modality knowledge distillation [Thoker and Gall, 2019; Liu *et al.*, 2021], multi-view knowledge distillation [Lin and Tseng, 2023], and low-resolution action recognition [Purwanto *et al.*, 2019]. Such methodologies improve model performance, particularly when limited input is available during inference. [Lin and Tseng, 2023] proposed a Multi-view knowledge distillation framework that enables the model to efficiently learn from a single view while effectively capturing knowledge from all views. This demonstrates the capability of knowledge distillation to enable models to learn and integrate information from diverse sources. For light enhancement learning in dark environments, several studies [Park *et al.*, 2022; Li *et al.*, 2023] have also utilized knowledge distillation and achieved notable success in the upstream enhancement task. Finally, for the specific task of human action recognition in the dark, [Jin *et al.*, 2023] highlighted the critical role of knowledge distillation, due to the high computational cost of video input. Their experiments showed that integrating knowledge distillation with optical flow and RGB features effectively supports model training. In our research, we aim to explore a novel approach by applying knowledge distillation to the downstream classification task using the enhanced features. This strategy allows the student model to benefit from light enhancement while only requiring the original video input, which optimizes both performance and computational cost.

3 Proposed Methods

3.1 Problem Setup

Action recognition aims to predict action labels from given input videos. Formally, let $D = \{(x_i, y_i)\}_{i=0}^{n}$ be a video-based dataset, where x_i is a sample input, y_i is the corresponding action label, and n is the number of samples. The goal is to train a model that takes each video inputs x_i to predict its action label $\hat{y_i}$ and minimizing the difference between each y_i and $\hat{y_i}$.

In our method using knowledge distillation, we train a teacher model to transfer its knowledge to a separate student model. During training phase, we train a teacher model T, which consists of an enhancement module T_e and a backbone classifier T_c , and a student model S separately. For teacher model T, given training samples $X = \{x_1, x_2, ..., x_j\}$ from D_{train} as the input of T_e, the module would generate enhanced samples X'. After that, X' will be served as the input of T_c, which predict on X' to the probability of each class, denote as $\hat{y}^t$:

$$\hat{y}^t = T_c(x') = T_c(T_e(x)) = T(x), \tag{1}$$

The loss function is then applied on $\hat{y}^t$ and ground truth y.

For the student model S, given training samples $X = \{x_1, x_2, ..., x_j\}$ from D_{train} as the input of S, the model generates the probability of each class, denoted as $\hat{y}^s$. To train the student model with both ground truth and the knowledge provided by the teacher model, the loss function is then applied on $\hat{y}^s$, $\hat{y}^t$ and the ground truth y. Detailed loss function will be explained in section 3.4. The process of the student model can be denoted as:

$$\hat{y}^s = S(x). \tag{2}$$

During the inference phase, only the student model S will be in use for prediction. The student model takes the video inputs x_i and generates their prediction labels $\hat{y_i}$, this completes the action recognition process.

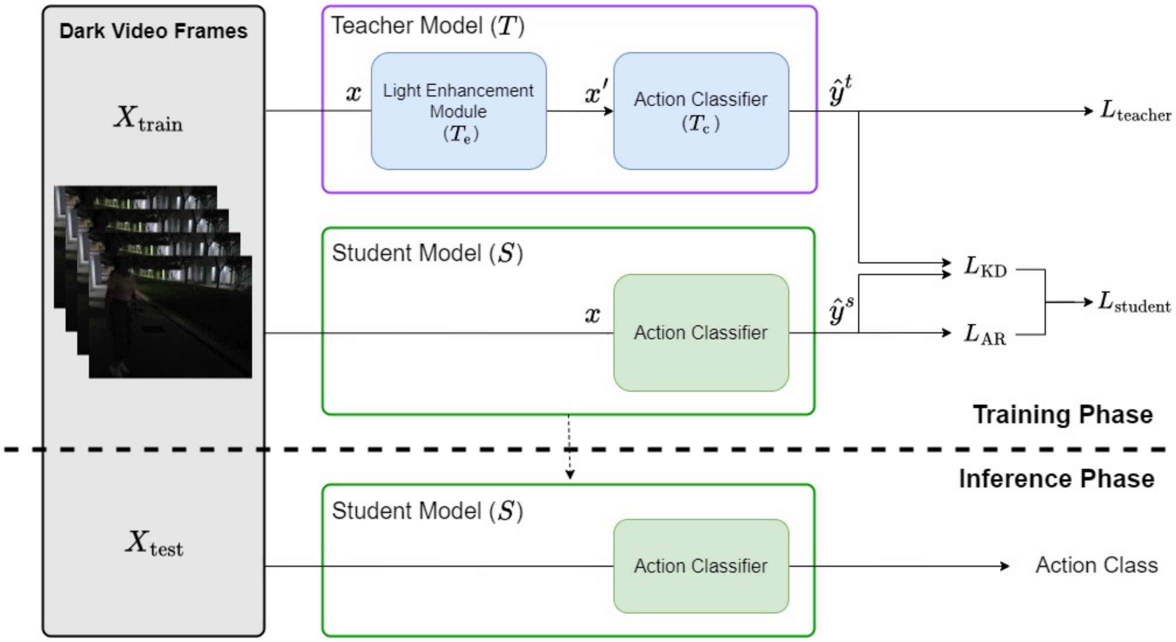

Figure 2: The overall architecture of DL-KDD. The teacher model includes a light enhancement module and an action classifier. The student model includes only an action classifier. Knowledge distillation is applied to train the student model from the representation generated by the teacher model.

3.2 Overview of the Proposed Framework

The overall architecture of the proposed DL-KDD is shown in Figure 2. It consists of two main components: a teacher model and a student model. The two models are trained with a dual-lightness knowledge distillation process, which is the core novel part of our research. Our framework offers an effective learning process for the model to learn diverse knowledge efficiently from the input source. Specifically, the components and workflow of the entire workflow are as follows. The teacher model includes a light enhancement module followed by an action classifier. The module first enhances the original video frames to improve visibility in order to extract features that may be loss in the dark. These enhanced features are then fed into the action classifier to generate feature representations.

After training the teacher model, we train the student model by taking original video frames without enhancement as input. This approach allows the student model to learn directly from the original data, ensuring that the model is not dependent on enhanced inputs during inference, which results to a reduction of computational cost. The student model is an action classifier, and the training of the student model involves a dual learning process:

1. Direct Learning: The student model learns directly from the ground truth labels.
2. Distillation Learning: The student model is trained using the outputs from the teacher model as soft targets, which allows the student model to learn from the feature representation that has been extracted from enhanced frames by teacher model.

Due to the relatively high computational cost of the video input compared to regular image-based input, we train the two models separately to minimize the computational load during the training phase. The teacher model's parameters are frozen while training the student model. Since the teacher model is fully trained at this point, it provides high-quality soft targets for the student model. This ensures the student model can obtain more accurate knowledge and avoids the potential negative impact of incorrect information during the early stages of training if the two models are trained simultaneously. This architecture make use of both enhanced and original video, which optimizes the model's performance without additional computational cost of using both data or processing enhancement during inference.

3.3 Feature Enhancement and Backbone Classifier

In this section, we will discuss the selection of the enhancement module and backbone classifier, which are foundational to our proposed framework. Our architecture can be generalized across different enhancement methods and backbone classifiers. However, a suitable choice of these features can further amplify the benefits brought by our framework. Therefore, we selected these components based on the following reasons:

Light Enhancement Module. For the light enhancement module, we selected ZeroDCE [Guo *et al.*, 2020] to generate enhanced video frames. Compared to other light enhancement methods [Fu *et al.*, 2016; Guo *et al.*, 2016] , ZeroDCE performs well in enhancing low-light samples by dynamically adjusting parameters with different inputs. It preserves the relationships between pixels in the frames, which has proven effective in supporting action classification in dark video human action recognition. This functionality makes it an ideal choice for the enhancement module in the teacher model of our framework.

Action Classifier. Inspired by [Chen *et al.*, 2021; Singh *et al.*, 2023], we have adopted the R(2+1)D [Tran *et al.*, 2018] combined with a self-attention block as our action classifier. The R(2+1)D can effectively capture critical information from the input video by decomposing the input into 2D and 1D components, which form spatial and temporal information in the video. The self-attention block further enables the model to recognize long-term dependencies between the extracted features, allowing the model to interpret complex activities in the videos. In our framework, we integrate the enhancement module and classifier to form the teacher model and a separated classifier as the student model to build a Knowledge Distillation framework. Detailed discussions on the KD learning process will be presented in the following section.

3.4 Dual-Lightness Knowledge Learning

Enhanced Frame Extraction. The teacher model's training begins with enhancing the original video frames. The enhancement module transforms the input X into X'. This enhancement improves the feature visibility and information for action recognition in dark videos. After enhancement, the action classifier processed the enhanced frames into a set of feature representations $\hat{y}^t$ that capture the crucial information of the action label. A standard Cross-Entropy Loss is applied here for training the teacher model:

$$L_{\text{teacher}}(y, \hat{y}^t) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i^t), \tag{3}$$

where y is the class label of the input video, and $\hat{y}^t$ is the feature representation extracted by the teacher model, and C is the number of class labels.

Original Representation Learning. Unlike the teacher model, the student model directly takes the original video frames X as input. The action classifier is the same as that of the teacher models, but there is no enhancement module for the student model. The inputs are processed by the student model to generate the results $\hat{y}^s$, and a cross-entropy loss is applied here to align the student model's prediction with the ground truth:

$$L_{\text{AR}}(y, \hat{y^s}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i^s). \tag{4}$$

Dual Knowledge Learning. In addition to learning from the ground truth, a knowledge distillation process is applied to the student model, where it learns from the feature representation $\hat{y}^t$ generated by the teacher model. This is achieved by minimizing the Kullback-Leibler divergence between the student model's output $\hat{y}^s$ and the soft targets provided by the teacher model, which indirectly transfers the enhanced features knowledge from the teacher to the student model:

$$L_{\text{KD}}(\hat{y}^t, \hat{y}^s, \tau) = \sum_{i=1}^{C} \hat{y}_i^t(\tau) \log\left(\frac{\hat{y}_i^t(\tau)}{\hat{y}_i^s(\tau)}\right). \tag{5}$$

In this case, τ served as the temperature used to adjust the smoothness of the softmax function. The temperature parameter helps soften the probability distribution of the model output, which makes it more suitable for transferring the light-enhanced knowledge from the teacher model to the student model. Specifically, the logits are scaled by τ before applying the softmax function. The use of τ can be formalized as:

$$\hat{y}_i(\tau) = \frac{\exp(z_i/\tau)}{\sum_{j=1}^{C} \exp(z_j/\tau)}, \tag{6}$$

where z_i represent the logits of the models.

The overall loss function for training the student model is a combination of these two loss functions — Cross-Entropy loss for learning from the ground truth and KL divergence for knowledge transfer from the teacher model, with a hyperparameter α to control the weight of the knowledge distillation loss. The total loss function for the student model is denoted as:

$$L_{\text{student}} = L_{\text{AR}} + \alpha L_{\text{KD}}, \tag{7}$$

Alpha Decay. Unlike conventional knowledge distillation frameworks where the student model remains inferior to the teacher, our student model surpasses the teacher model after integrating dual-lightness knowledge. As a result, over-relying on the teacher's distribution might limit the student's potential during later training stages. To address this, we introduce an alpha decay mechanism that dynamically reduces the importance of the knowledge distillation loss:

$$\alpha_t = \alpha_0 \cdot \gamma^t \tag{8}$$

where α_0 is the initial weight, γ is the decay rate, and t denotes the epoch. This adjustment ensures a smooth transition from teacher-guided learning to independent optimization by ground truth, which allows the student model to adapt its learning focus at different training periods.

4 Experimental Evaluations

4.1 Experiment Settings

Dataset. We evaluated our method on three datasets: ARID [Xu *et al.*, 2021] and its updated version ARID V1.5, and Dark-48 [Tu *et al.*, 2023]. These datasets provide various scenes and lightness conditions to validate the generalizability of the model. For the experiments of all three datasets used in this work, the training and testing settings are the same as the original work [Xu *et al.*, 2021; Tu *et al.*, 2023]. The splits and folds are fixed to ensure fair comparisons.

The ARID [Xu *et al.*, 2021] dataset has been a primary benchmark of dark video human action recognition. It contains over 3780 video clips collected with 11 action classes, and the training and testing set ratio is 7:3. The videos are collected in 9 outdoor and 9 indoor scenes.

To further enhance the complexity of the dataset, ARID V1.5 was introduced. The class number remains the same, but the video count is expanded to 6207, and the videos are collected from 24 scenes, with 12 indoor and 12 outdoor scenes. There are more than 320 clips for each action class in this version.

Model	**Method**	**Input Feature**		**Top-1 Accuracy (%)**
		Training	**Inference**	
I3D-RGB	-	Original	Original	68.29
I3D Two-stream	Two-Stream	Original + Optical Flow	Original + Optical Flow	72.78
3D-ResNet-101	-	Original	Original	71.57
3D-ResNext-101	-	Original	Original	74.73
Video-Swin-B	-	Original	Original	89.79
DarkLight	Two-Stream	Original + GIC	Original + GIC	94.04
DTCM	-	ZeroDCE	ZeroDCE	96.36
R(2+1)D-GCN+BERT	-	ZeroDCE	ZeroDCE	96.60
DL-KDD (Ours)	KD	Original + ZeroDCE	**Original**	**97.64**

Table 1: Comparison with state-of-the-art (SOTA) methods on the ARID dataset with detailed method descriptions. The ‘Method‘ column specifies the approach used to integrate multiple features, where a dash (-) indicates models that do not utilize multiple features. The ‘Input Feature‘ section details the data used for training and inference, specifying any enhancement techniques. Finally, the ‘Top-1 Accuracy‘ column presents the classification performance of each model.

Model	**Top-1 Accuracy (%)**
I3D-RGB	48.75
I3D Two-stream	51.24
DarkLight	84.13
R(2+1)D-GCN+BERT	86.93
DL-KDD (Ours)	**88.12**

Table 2: Comparison with state-of-the-art (SOTA) methods on the ARID V1.5 dataset. The ‘Top-1 Accuracy‘ column presents the classification performance of each model.

Model	**Top-1 (%)**	**Top-5 (%)**
I3D-RGB	32.25	65.35
3D-ResNext-101	37.23	68.86
DarkLight	42.27	70.47
DTCM	46.68	75.92
DL-KDD (Ours)	**52.26**	**80.18**

Table 3: Comparison with state-of-the-art (SOTA) models on the Dark-48 dataset. The ‘Top-1 Accuracy‘ and ‘Top-5 Accuracy‘ columns represent the classification performance of each model.

The Dark-48 [Tu *et al.*, 2023] dataset comprises 8815 dark videos from more than 40 scenes, featuring 48 classes with over 100 videos each. The dataset is split into training and testing sets in a ratio of 8:2.

Implementation Details. This work is implemented with PyTorch [Paszke *et al.*, 2019]. Inspired by [Chen *et al.*, 2021] , we selected R(2+1)D [Tran *et al.*, 2018] followed by BERT [Devlin *et al.*, 2018] in replace of the conventional temporal global average pooling layer as our backbone classifier. The backbone classifier was pretrained on IG65M [Ghadiyaram *et al.*, 2019] . For the enhancement module of the teacher model, we selected ZeroDCE [Guo *et al.*, 2020] with original pretrained weight. The input sequences were resized to 112 x 112 pixels, and the final input shape was 3 x 64 x 112 x 112 with batch size 8. We trained the teacher and the student model with AdamW [Loshchilov and Hutter, 2017] optimizer with a 0.0001 learning rate and a 0.00001 decay rate. The parameter α for the loss function was set to 1.5, the alpha decay rate γ is set to 0.95, and the temperature τ was set to 5.0. We trained 30 epochs for the ARID and ARID V1.5 datasets, and for Dark-48, we trained 50 epochs to optimize the parameters of the model. The experiments are conducted on a server with a Tesla V100 GPU.

Metrics. In this task, we recorded both top-1 and top-5 accuracy to evaluate the performance of the model. Since the ARID and ARID V1.5 [Xu *et al.*, 2021] datasets contain only 11 classes and most previous work has almost reached 100% top-5 accuracy, the comparison of this metric becomes less meaningful in these datasets. To provide more insight within the limited space, we primarily present the top-1 accuracy for ARID and ARID V1.5.

4.2 Comparison with State-of-the-Art Methods

We conduct extensive experiments to compare our work with the recent state-of-the-art methods in dark video human action recognition, including DarkLight [Chen *et al.*, 2021], DTCM [Tu *et al.*, 2023], and R(2+1)D-GCN+BERT [Singh *et al.*, 2023] across the ARID, ARID V1.5, and Dark48 datasets. Partial results from previous works are collected from [Chen *et al.*, 2021; Singh *et al.*, 2023; Tu *et al.*, 2023].

Table 1 presents detailed results of the ARID Dataset. Despite high baseline performances on the dataset, our proposed method outperforms existing methods and reached 97.64% accuracy, achieving the State-of-the-Art result in the dataset.

Table 2 presents detailed results for the ARID V1.5 dataset, which is a more complicated version of the ARID dataset. Our model obtains the best performance in this dataset with an accuracy of 88.12%, which demonstrates the robustness of our model in diverse data conditions.

Table 3 indicates that our model reached a Top-1 accuracy of 52.26% on the Dark-48 dataset. This demonstrates a significant improvement over the best previously reported result on Dark-48 by 5.58%. Compared to the ARID datasets, the Dark-48 dataset is more complex in the video scene and has a

Model	Top-1 Accuracy (%)
Backbone Only	92.97
DL-KDD-Teacher	95.73
DL-KDD-Student (Ours)	**97.64**

Table 4: Comparative performance of DL-KDD on the ARID dataset. The student and backbone models share the same architecture, while our method shows a 4.67% improvement.

Method	GPU Memory (MB)	GFLOPs
Enhancement-based	2133	310.28
Two-stream	1599	674.84
DL-KDD (Ours)	**1597**	**305.51**

Table 5: Computational cost comparison of different methods in terms of GPU memory and GFLOPs during inference. DL-KDD achieves the lowest computational cost.

more significant number of action classes. The performance improvement shows the effectiveness of our proposed strategy to learn features more efficiently in complex scenarios.

These results illustrate that our proposed knowledge distillation framework successfully enhances the information learned by the model, which enables our model to achieve best performance while using only the original video input during testing.

4.3 Ablation Study

In this section, we focus on an ablative comparison on the ARID dataset to demonstrate the effectiveness of our proposed framework. To illustrate the efficacy of our training method, we present results comparing the backbone model trained with and without our method. Additionally, comparisons between the teacher and student models are displayed to show that the student network can achieve better results even without enhancement after the knowledge distillation training. Table 4 provides a detailed display of the final performance of the teacher, student model of DL-KDD, and the performance of similar architecture without knowledge distillation training method. To assess the computational efficiency of our method, Table 5 reports the computational cost of different training methods using the same backbone model.

Effectiveness of Knowledge Distillation. As shown in Table 4, our knowledge distillation training method improved the performance of the same architecture by 4.67%, which shows the effectiveness of learning from the knowledge distilled from the enhanced features by the teacher model. With the additional knowledge provided by the teacher model, the student model can take advantage of enhanced representation even without enhanced feature inputs during testing. Furthermore, the comparison between the student and teacher models shows that even though the student model uses a simpler architecture, it achieves an improvement of 1.91% over the teacher model, which indicates that the original video also contains critical information that improves model performance. Our knowledge distillation approach does not simply align student model's performance to the teacher model. Instead, it focuses on integrating the features extracted from the teacher model into the student model. Besides the knowledge provided by the teacher model, learning from original inputs allows the student model to access additional information from enhanced features, resulting in better performance than the teacher model.

Computational Cost Comparison. Table 5 presents a comparison of computational cost among our method, enhancement-based method (with ZeroDCE), and traditional two-stream methods (with rule-based enhancement). To ensure a fair comparison, all methods share the same backbone model architecture. Compared to enhancement-based methods, which require frame-wise enhancement due to the video input, our approach eliminates the need for an enhancement module during inference, leading to a 25% reduction in GPU memory requirement. On the other hand, in comparison to traditional two-stream methods, our method shows advantage in its computational efficiency. Although two-stream methods leverage weight-sharing to maintain the model size, they have to process the input video twice for each stream, resulting in doubled GFLOPs compared to our method. These results demonstrate that our method incorporating knowledge distillation optimizes both memory usage and computational efficiency, making it more suitable for real-world applications.

5 Conclusion

In this work, we propose a novel knowledge-distillation-based framework named *DL-KDD* for dark video human action recognition, emphasizing the importance of utilizing both the original video and the enhanced features to prevent the loss of original information. Moreover, the proposed framework avoids the additional cost brought by two-stream methods. We effectively distill the knowledge of light enhancement to the student model, enabling the student model to use only original videos as input during inference and achieve better results. The excellent performance on the ARID, ARID V1.5 and Dark-48 datasets proves the effectiveness of our method.

One limitation of the proposed architecture is that the teacher-student framework requires training an additional model compared to regular single-model approaches. However, this is a necessary trade-off that brought the advantage of higher performance with efficient use of resources during inference, which is crucial for real-world applications.

For future work, the proposed *DL-KDD* can be extended with a more complex knowledge distillation process or provide diverse features such as cross-modality data to further improve performance. Additionally, this work is flexible for modifications as a more lightweight architecture for various real-time applications.

Acknowledgements

This research was supported in part by National Science and Technology Council Taiwan under grant no. NSTC 113-2223-E-A49-008 and 113-2634-F-A49-005.

References

[Akula *et al.*, 2018] Aparna Akula, Anuj K Shah, and Ripul Ghosh. Deep learning approach for human action recognition in infrared images. *Cognitive Systems Research*, 50:146–154, 2018.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021.

[Carreira and Zisserman, 2017] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[Chen *et al.*, 2021] Rui Chen, Jiajun Chen, Zixi Liang, Huaien Gao, and Shan Lin. Darklight networks for action recognition in the dark. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 846–852, 2021.

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

[Duan *et al.*, 2022] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978, June 2022.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[Fu *et al.*, 2016] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2782–2790, 2016.

[Ghadiyaram *et al.*, 2019] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019.

[Goyal *et al.*, 2017] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

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

[Guo *et al.*, 2020] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1777–1786, 2020.

[Hara *et al.*, 2017] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[Hara *et al.*, 2018] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Jiang *et al.*, 2017] Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[Jin *et al.*, 2023] Ruibing Jin, Guosheng Lin, Min Wu, Jie Lin, Zhengguo Li, Xiaoli Li, and Zhenghua Chen. Unlimited knowledge distillation for action recognition in the dark. *arXiv preprint arXiv:2308.09327*, 2023.

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

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[Kuehne *et al.*, 2011] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[Li *et al.*, 2023] Ziwen Li, Yuehuan Wang, and Jinpu Zhang. Low-light image enhancement with knowledge distillation. *Neurocomputing*, 518:332–343, 2023.

[Lin and Tseng, 2023] Ying-Chen Lin and Vincent S. Tseng. Multi-view knowledge distillation transformer for human action recognition. *arXiv preprint arXiv:2303.14358*, 2023.

[Liu *et al.*, 2021] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Transactions on Image Processing*, 30:5573–5588, 2021.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

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

[Park *et al.*, 2022] Jeong-Hyeok Park, Tae-Hyeon Kim, and Jong-Ok Kim. Dual-teacher distillation for low-light image enhancement. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1351–1355, 2022.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Purwanto *et al.*, 2019] Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Extreme low resolution action recognition with spatial-temporal multi-head self-attention and knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[Singh *et al.*, 2023] Himanshu Singh, Saurabh Suman, Badri Narayan Subudhi, Vinit Jakhetiya, and Ashish Ghosh. Action recognition in dark videos using spatio-temporal features and bidirectional encoder representations from transformers. *IEEE Transactions on Artificial Intelligence*, 4(6):1461–1471, 2023.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Tammvee and Anbarjafari, 2021] Martin Tammvee and Gholamreza Anbarjafari. Human activity recognition-based path planning for autonomous vehicles. *Signal, Image and Video Processing*, 15, 2021.

[Thoker and Gall, 2019] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10, 2019.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

[Tu *et al.*, 2023] Zhigang Tu, Yuanzhong Liu, Yan Zhang, Qizi Mu, and Junsong Yuan. Dtcm: Joint optimization of dark enhancement and action recognition in videos. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, PP, 2023.

[Wasim *et al.*, 2023] Syed Talal Wasim, Muhammad Uzair Khattak, Muzammal Naseer, Salman Khan, Mubarak Shah, and Fahad Shahbaz Khan. Video-focalnets: Spatio-temporal focal modulation for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13778–13789, October 2023.

[Xian *et al.*, 2024a] Ruiqi Xian, Xijun Wang, Divya Kothandaraman, and Dinesh Manocha. Pmi sampler: Patch similarity guided frame selection for aerial action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6982–6991, January 2024.

[Xian *et al.*, 2024b] Ruiqi Xian, Xijun Wang, and Dinesh Manocha. Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6625–6634, January 2024.

[Xu *et al.*, 2021] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2*, pages 70–84. Springer, 2021.

[Xu *et al.*, 2022] Feiyi Xu, Feng Xu, Jiucheng Xie, Chi-Man Pun, Huimin Lu, and Hao Gao. Action recognition framework in traffic scene for autonomous driving system. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):22301–22311, 2022.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Yang *et al.*, 2022] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022.

[Zhang *et al.*, 2024] Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. Pgvt: Pose-guided video transformer for fine-grained action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6645–6656, January 2024.