ELLAR: An Action Recognition Dataset for Extremely Low-Light Conditions with Dual Gamma Adaptive Modulation

Minse Ha*, Wan-Gi Bae*, Geunyoung Bae, and Jong Taek Lee†

School of Computer Science and Engineering
Kyungpook National University, Daegu, South Korea
{haminse,bwg7408,flora8207,jongtaeklee}@knu.ac.kr

Abstract. In this paper, we address the challenging problem of action recognition in extremely low-light environments. Currently, available datasets built under low-light settings are not truly representative of *extremely* dark conditions because they have a sufficient signal-to-noise ratio, making them visible with simple low-light image enhancement methods. Due to the lack of datasets captured under extremely low-light conditions, we present a new dataset with more than 12K video samples, named Extremely Low-Light condition Action Recognition (ELLAR). This dataset is constructed to reflect the characteristics of extremely low-light conditions where the visibility of videos is corrupted by overwhelming noise and blurs. ELLAR also covers a diverse range of dark settings within the scope of extremely low-light conditions. Furthermore, we propose a simple yet strong baseline method, leveraging a Mixture of Experts in gamma intensity correction, which enables models to be flexible and adaptive to a range of low illuminance levels. Our approach significantly surpasses state-of-the-art results by 3.39% top-1 accuracy on ELLAR dataset. The dataset and code are available at https://github.com/knu-vis/ELLAR.

Keywords: Extremely low-light conditions dataset · Action recognition

1 Introduction

Action recognition has seen remarkable advancements in recent years, driven by the growing interest in its real-world applications. Phenomenal advancements have been made possible by the development of state-of-the-art deep neural network models such as those based on Convolutional Neural Networks (CNNs) [5,17,47–49,52] and video Transformers [3,4,16,31], utilizing large-scale datasets [19, 27, 42, 43]. However, these architectures rely on datasets captured under normal lighting conditions such as daylight, indoor lighting, and studio settings where human actions are clearly perceived [13, 19, 42, 43]. As a result, when

⋆ Equal contributions
† Corresponding Author

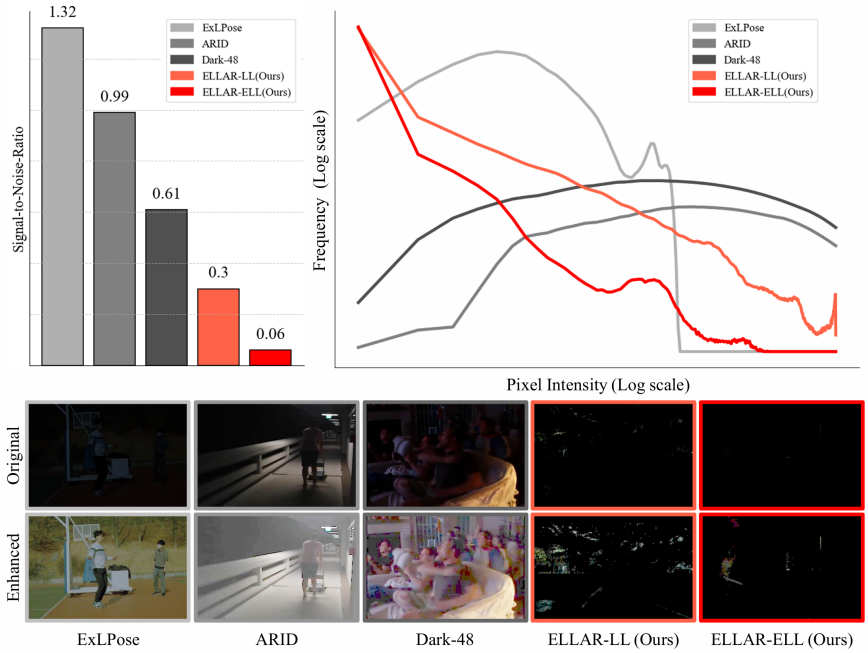

Fig. 1: Pixel distributions and signal-to-noise ratio among various low-light datasets: the line graph depicts pixel intensity distributions, with the x-axis representing pixel values from 0 to 255, plotted on a logarithmic scale to highlight values near 0. The y-axis represents the frequency of each pixel value, also on a logarithmic scale. The bar graph illustrates the signal-to-noise ratio for the aforementioned datasets, meaning that the higher the value, the more signals are present in an image. Two rows of 10 images show the original images from the datasets and images that Gamma Intensity Correction (GIC) is applied respectively.

employed in low-light environments in which the quality of videos is degraded harshly, existing methods perform poorly.

Several low-light datasets [29, 50, 58] have been released for specific downstream tasks. Dark-48 [50] is composed of videos from Kinetics-700 and Moments in Time [35, 42], selected simply based on their relatively lower average pixel intensity from the available pool, but they are not specifically designed for tasks under low-light condition. ExLPose [29] is a dataset of human pose estimation (HPE) in extremely low-light conditions, captured by a dedicated camera system that is capable of taking pairs of well-lit and low-light images. However, it is not possible to collect pairs of well-lit and low-light images *out in the wild* as it is challenging to replicate the camera system, and the dataset would lack distinct noise and blurs caused by real low-light conditions. ARID [58] has been released, maintaining that it is the first dataset of action recognition in the dark. One of the limitations of this dataset is that the vast majority of human actions take place in the center of the videos, which would mislead models to focus on the

videos' center. Furthermore, the major problem is that none of these datasets can be considered truly *extremely dark* as they do not accurately represent the challenges posed by low-light conditions such as increased noise, motion blur, and a low signal-to-noise ratio (SNR) [40, 45], making their visibility be significantly improved with simple image enhancement methods because of their sufficient SNR, as depicted in Fig. 1.

In this paper, we introduce a new dataset, Extremely Low-Light condition Action Recognition (ELLAR), in order to overcome the lack of dataset captured under extremely low-light conditions. It is noteworthy that ELLAR is carefully designed and captured to contain adversarial factors in real-world applications that seriously degrade the visibility and SNR of videos. This dataset is taken at night under extremely dark settings by three different cameras and it consists of two types: low-light (LL) and extremely low-light (ELL) videos, covering a diverse range of low illuminance scenarios in the boundary of extremely dark settings. Another important aspect is that human actions in our dataset occur not just in the center of the videos but also in the left, right, and other parts at different scales, which prevents models from being center-sighted.

Existing methods for addressing low-light action recognition [8,23] are trained with supplementary information such as enhanced images and optical flow. However, this preprocessing for producing additional information has two main problems: first, it is applied uniformly across the entire dataset, ignoring specific features of individual videos that require customized preprocessing. Second, it is isolated from the action recognition architecture, failing to deliver integrated and context-aware enhancements.

To solve these problems, we propose a simple yet strong baseline method named Dual Gamma Adaptive Modulation (DGAM). DGAM takes advantage of a Mixture of Experts approach [24, 36] to determine gamma intensity correction (GIC) that is optimal for each input video, selecting the reciprocal classification heads dedicated to the corrected inputs simultaneously. Our approach, which trains both the image enhancement and action recognition modules together, enables us to maintain high generalization performance in a variety of low-light environments.

We evaluate our proposed method on ELLAR dataset in terms of classification accuracy and cross-domain adaptation. DGAM tremendously outperforms state-of-the-art (Video Swin Transformer [31]), evaluated on ELLAR. Our key contributions are summarized as follows:

- We define the challenging problem of action recognition in *extremely* low-light conditions, which we contend is highly practical.
- We provide a new dataset called ELLAR that contains more than 12K videos captured under extremely dark settings and it covers a wide range of low illuminance scenarios within the scope of extremely low-light conditions.
- We propose a strong baseline method called DGAM that significantly surpasses state-of-the-art results by 3.39% top-1 accuracy on ELLAR and demonstrates its cross-domain adaptation, outperforming it by 1.47% (LL→ELL) and 8.29% (ELL→LL).

2 Related Work

2.1 Low-light Dataset

LOL [53] and SID [7] datasets were collected by manipulating camera parameters like exposure time and ISO, providing paired normal and dark images. These datasets are frequently used in learning-based low-light image enhancement tasks. ExLPose [29], the first low-light human pose estimation dataset, contains both well-lit and low-light images of the same content under different illumination conditions. Dark-48 [50] is a collection of dark videos selected from two large-scale action recognition datasets, Kinetics-700 [42] and MiT [35]. ARID [58] is the first dataset for action recognition in low-light environments, consisting of 11 classes and over 2k low-light video samples, each 2-3 seconds in duration. Nevertheless, the aforementioned datasets have an extensively higher number of non-zero pixels and SNR than those of our dataset, ELLAR, as shown in Fig. 1, indicating that they are not collected under extremely low-light conditions.

2.2 Action Recognition

CNN-based methods have been widely used for action recognition and have shown good performances before the advent of Transformers [51]. 2D CNN-based methods [25,41] employ a two-stream approach to separately handle spatial and temporal features. Those based on 3D CNNs [5, 17, 26, 47], on the other hand, perform convolutions over both spatial and temporal dimensions, allowing them to capture motion information across multiple frames. However, these models come at the price of demanding high computational complexity and resources. Transformer-based models [3, 4, 31] have recently gained attention for their ability to classify human actions, using spatio-temporal self-attention mechanisms that expand the receptive field more than previous methods with fewer parameters. TimeSformer [4] adapts the Transformer architecture to video by enabling spatiotemporal feature learning directly from a sequence of frame-level patches, based entirely on self-attention mechanisms over space and time. Unlike previous methods that use global self-attention, Video Swin Transformer [31] uses 3D shifted windows that leverage local self-attention, which is computationally more efficient.

2.3 Low-light Image Enhancement

Deterministic methods such as Gamma Intensity Correction (GIC) [22, 37] and Histogram Equalization (HE) [1, 6, 46] directly stretch out contrast and improve the visibility of low-light images. Conventional cognition methods [30, 38] based on Retinex theory combine color invariance with dynamic range compression to further improve the discernability. While these methods are computationally cost-effective and widely applicable, they lack the ability to optimize for individual samples. Learning-based methods [20, 21, 54, 55, 60] have shown remarkable

performances in recent years. These models are typically trained on pairs of light and dark images in a supervised manner, so their performance is often limited to cases similar to the training data, making them less effective in different illumination scenarios. Recent methods [12, 57, 61] address adaptive low-light image enhancement, based on SNR and illumination level of input images.

2.4 Action Recognition under Low-light Conditions

Existing methods for action recognition under low-light settings can be broadly categorized into two approaches: two-stage and two-stream [50, 59].

The two-stage method involves the sequential processes of applying low-light image enhancement methods and deploying action recognition models. This method benefits from using existing action recognition encoders, but it is highly influenced by the performance of the image enhancer and the training data. There are two detailed approaches to implementing a two-stage method for image enhancement: algorithm-based and model-based. Algorithm-based [1, 6, 22, 30, 37, 38, 46] two-stage approaches are low-cost and general-purpose, but they apply the same image enhancement to all samples, ignoring specific characteristics of inputs, which limits performance gains. In contrast, model-based enhancement [21,55,60] two-stage approaches show impressive performance and adaptive image intensity correction, but they heavily rely on training data. This dependency makes them challenging to apply to diverse action recognition scenarios as they require pairs of low-light and matching well-lit images for training. There are numerous pose-based two-stage models, such as those presented in [9–11]. However, these approaches are not optimal in extremely low-light conditions, as extracting pose information in the dark is challenging and can negatively impact action recognition.

The second approach is two-stream where low-light images and additional information (*e.g.*, optical flow, paired bright images, infrared images, and depth sensor images) are trained together [5, 8, 23, 41]. However, these methods have limitations since they require preprocessing, especially optical flow [15,44], which is extremely time-consuming and computationally intensive. DarkLight [8] employs a two-stream technique augmented with algorithmic-based enhancement to improve the visibility of low-light images, demonstrating commendable performance on existing action recognition datasets under low-light conditions. However, this approach necessitates extensive preprocessing, which is time-consuming and suffers from limited generalizability due to the uniform application of enhancement across all samples.

Recently, an end-to-end model approach has been developed that trains the image enhancement module and action recognition encoder together. An example of this is DTCM [50], which allows image enhancement considering action recognition and does not require additional computations and preprocessing. However, the performance of aforementioned methods drops significantly when they are employed in extremely low-light conditions.

3 ELLAR Dataset

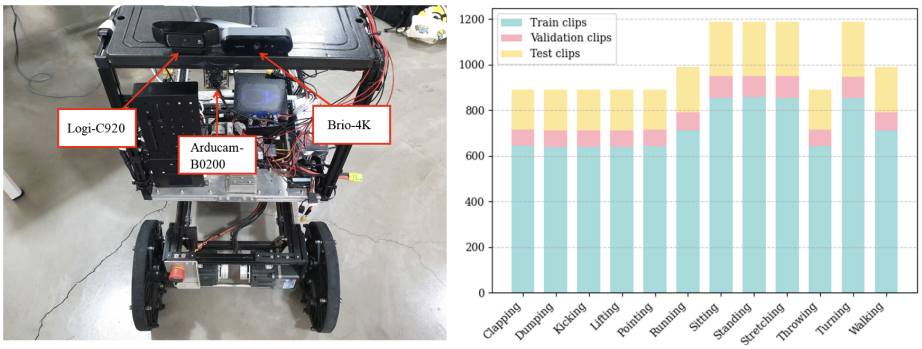

Fig. 2: Utilized cameras and class distribution of ELLAR.

Table 1: Statistics of ELLAR dataset. ELLAR is composed of two types, depending on illuminance level: low-light (LL) and extremely low-light (ELL).

Light Condition	#Videos	Time (s)	#Actions	#Outfits	#Actors	Camera	Resolution
LL	6,222	22,833	12	17	5	Brio-4k [32]	1280 x 720
						C920 [33]	1280 x 720
ELL	5,856	21,641				Arducam [2]	640 x 480

We collected a new dataset named Extremely Low-Light condition Action Recognition (ELLAR). This dataset is divided into two parts based on the illumination of the locations: low-light (LL) and extremely low-light (ELL). The LL part is captured at three outdoor locations under low-light conditions and the ELL part is recorded at two extremely low-light indoor settings. ELLAR was filmed with five actors wearing 17 different colors of clothing, using three different cameras to cover a broad spectrum of real-world scenarios. Motivated by the fact that color information is important in dark settings [14, 34], we purposely vary the colors of actors' clothing. For the dataset creation, we selected three diverse camera models: Brio-4K (Brio-4k) [32], Logi-C920 (C920) [33], and Arducam-B0200 (Arducam) [2], as shown in Fig. 2. The first two models are widely used as webcams, providing a standard for general consumer usage, while the third is an embedded camera module, utilized in robotics for machine vision applications. ELLAR consists of 12 daily atomic action classes such as running, turning, and sitting. Videos are approximately 3-4 seconds long in AVI format. Table 1 summarizes the statistics of ELLAR dataset.

Table 2: Statistical comparison between ARID, Dark-48, and ELLAR (Ours).

Dataset	Time (s)	Avg. Clips/class	Total Clips
ARID [58]	8,721	110	3,784
Dark-48 [50]	26,445	183	8,815
ELLAR (Ours)	**44,474**	**1,008**	**12,078**

As shown in Tab. 2, ELLAR stands out in several aspects. It includes 12,078 total clips, which is the highest among the datasets compared. This collection spans a total duration of 44,474 seconds, making it the largest dataset in terms of time. Additionally, each class in ELLAR contains an average of 1,008 clips, exceeding the average clips per class in both ARID and Dark-48.

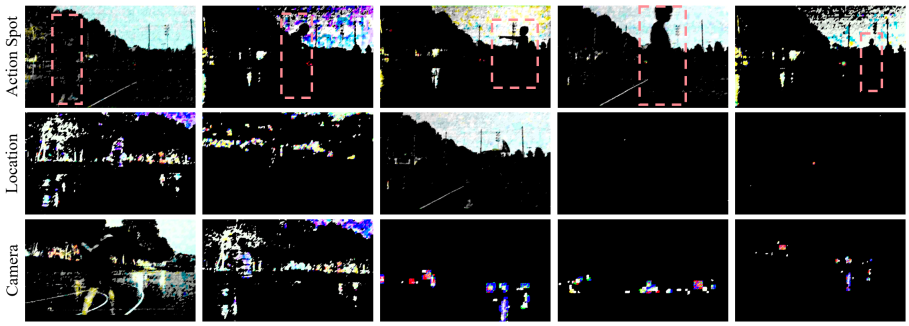

Fig. 3: Samples of ELLAR dataset. All samples are subjected to GIC for display purposes. The red dotted boxes depicted in the images represent the action spot in each sample.

The dataset introduced in this study was partitioned into training, validation, and testing subsets, as presented in Fig 2. The LL part was divided into 4,479 samples for training, 498 samples for validation, and 1,245 samples for testing. Similarly, for the ELL part configuration, the dataset comprised 4,212 samples for training, 471 samples for validation, and 1,173 samples for testing.

The highlight of ELLAR, illustrated in Fig. 3, is that it is strategically designed to include three challenging factors that would affect the quality of videos: locations, action spots, and types of cameras. ELLAR was recorded in five different locations, each introducing a range of environmental conditions, such as varying lighting conditions and spatial layouts. In ELLAR, human actions occur at various spots, including left, center, and right at different scales (*i.e.*, large and small proportions of camera views). In real-world applications, actions do not always occur in the center of the frame. By varying action spots, ELLAR better reflects real-life situations where actions can happen anywhere within the field of view. Finally, ELLAR is recorded by three different cameras, as different cameras have unique characteristics such as resolution, frame rate, and sensitivity to light, which significantly influence the quality of the videos.

4 Proposed Method

ELLAR contains videos captured under a wide range of extremely dark settings. Consequently, it is crucial to do image enhancement adaptively based on each video's illumination levels. Applying the same level of enhancement to the entire dataset may amplify undesirable noise in inputs that do not need enhancement; on the other hand, it would underperform in visibility correction for inputs that need more enhancement. In addition, the adaptive enhancement should collaborate with the action recognition architecture, allowing the architecture to flexibly concentrate on different features of the enhanced inputs. Therefore, we propose a method called DGAM that performs GIC tailored to the lighting conditions of each sample while cooperating with the action recognition architecture. The overall architecture is illustrated in Fig. 4.

4.1 Dual Gamma Adaptive Modulation

The core idea of DGAM is its dual Mixture of Experts structure. This structure first identifies the characteristics of each sample and performs adaptive image enhancement that is optimal for action recognition. To this end, we designed a gating network named Adaptive Gamma Correction (AGC) that selects and applies the optimal GIC based on the illuminance information of the samples. Moreover, the matching classification head called Adaptive Head Selection (AHS) is selected to recognize features specific to the enhanced inputs. This dual mixture of expert systems allows the action recognition model to dynamically respond to inputs from diverse dark settings.

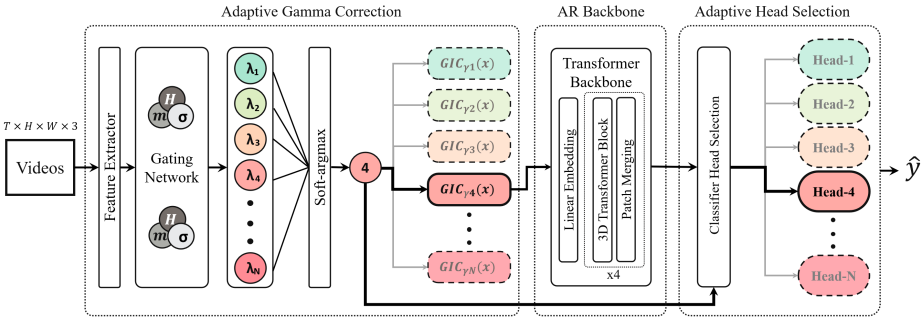

Fig. 4: The proposed model architecture: Dual Gamma Adaptive Modulation (DGAM).

4.2 Adaptive Gamma Correction (AGC)

Our method is inspired by the Mixture of Experts (MoE) methodology [24, 36], a machine learning paradigm that employs multiple specialized models and selects the best result from these experts. Building on this idea, we introduce AGC to flexibly choose a gamma value and apply GIC based on the features of each input.

As shown in Fig. 4, the AGC is divided into a feature extractor, a gating network, and an image enhancement part with optional GIC modules. For each input clip $\boldsymbol{I} \in \mathbb{R}^{T\times H\times W\times C}$ that comprises T frames, H height, and W width and C channels, feature extractor extracts features derived from the mean μ, standard deviation σ, and Shannon entropy S [39] of every frame. Since we compute the three values μ, σ, and S per frame, the features extracted from a single clip are a vector of size $(3 \times T,)$. The extracted feature $\mathbf{x_1}$ is then fed into a gating network to determine the optimal gamma value, which is used for GIC. Once gating network receives an input $\mathbf{x_1}$, it returns a λ vector of dimension $(N,)$, where N represents the number of GIC modules. Each element of the λ vector represents a different GIC module with a corresponding index. The following is the forward process of AGC given input $\boldsymbol{I}$.

$$\mathbf{x_1} = \phi(\boldsymbol{I}) = [\mu_1, \sigma_1, S_1, \mu_2, \sigma_2, S_2, \ldots, \mu_T, \sigma_T, S_T], \tag{1}$$

$$\mathbf{x}_{k+1} = \alpha(\mathrm{BN}(\mathrm{FC}(\mathbf{x}_k))), \quad \text{for } k = 1, 2, \ldots, n, \tag{2}$$

$$L_{\mathrm{gate}}(\boldsymbol{I}) = \mathrm{SoftArgmax}(\boldsymbol{\lambda}), \quad \text{for } \boldsymbol{\lambda} = \mathrm{Softmax}(\mathbf{x_n}) \tag{3}$$

where $\mathbf{x_1}$ denotes extracted features from feature extractor ϕ ; α, BN, and FC denote activation function, batch normalization, and feed-forward network, respectively. The gating network iterates the Eq. 2 n times.

To enable the model to learn through gradient-based optimization, the SoftArgmax function [28] is applied to the output of the gate network. Unlike the argmax function, which is non-differentiable, SoftArgmax provides a smooth and differentiable approximation, allowing gradients to flow through the network during backpropagation. This is essential for the training process, as it ensures that the gate network can be optimized alongside other components of the model. The equation for SoftArgmax is displayed below:

$$\mathrm{SoftArgmax}(\boldsymbol{\lambda}) = \sum_{p\in\mathbb{Z}^2} \mathrm{Softmax}(\boldsymbol{\lambda})(p)p \tag{4}$$

$$= \sum_{p\in\mathbb{Z}^2} \frac{e^{\boldsymbol{\lambda}(p)}}{\sum_{q\in\mathbb{Z}^2} e^{\boldsymbol{\lambda}(q)}} p. \tag{5}$$

In these equations, p represents the index or position of the elements in the input, while q serves as a dummy variable for summation in the denominator to ensure normalization. As a result, the gate network returns $L_{\mathrm{gate}}(\boldsymbol{I})$, the index of the maximum value in the lambda vector.

For enhancement, we introduce N GIC modules as experts, each with a different gamma hyperparameter value. The GIC algorithm is as follows:

$$GIC_\gamma(\boldsymbol{I}) = \left[(\max(\boldsymbol{I}) - \min(\boldsymbol{I})) \cdot \left(\frac{\boldsymbol{I} - \min(\boldsymbol{I})}{\max(\boldsymbol{I}) - \min(\boldsymbol{I})} \right)^{\frac{1}{\gamma}} \right] + \min(\boldsymbol{I}) \tag{6}$$

γ is the value of gamma hyperparameter. The higher γ applies a stronger intensity correction to the original pixels, but it also introduces more noise.

Based on the gating network result $L_{\mathrm{gate}}(\mathbf{I})$, GIC is applied to each clip and then fed into the backbone model. The gate function is trained end-to-end with the backbone network using the final classification loss function, ensuring that the gate function

finds the optimal gamma correction function while considering action recognition. This approach reduces the discrepancy between the enhancer and the recognizer, a common issue in traditional two-stage methods.

4.3 Adaptive Head Selection (AHS)

Adaptive Head Selection (AHS) is a method to dynamically choose the optimal classification head based on the index of max value in λ from the Adaptive Gamma Correction (AGC). Each head follows the structure of an I3D head. The AHS method integrates several layers for feature processing:

$$\mathbf{x} = \text{AdaptiveAvgPool3d}(\mathbf{x}), \tag{7}$$

$$\mathbf{x} = \text{Dropout}(\mathbf{x}), \tag{8}$$

$$\mathbf{y} = \mathbf{W}_i\mathbf{x} + \mathbf{b}_i, \quad \text{if } \lambda = \lambda_i, \quad \text{for } i = 1, 2, \ldots, N. \tag{9}$$

First, the input tensor $\mathbf{x}$ from the transformer backbone undergoes adaptive average pooling to reduce spatial dimensions. Then, dropout is applied to prevent overfitting. Finally, the tensor is passed through a set of fully connected layers. The last fully connected layer is selected based on the index of max value in λ value derived from the AGC, which identifies the optimal classification head for action recognition. This dynamic selection process ensures that each sample is processed by the head best suited for its specific GIC module.

5 Experiments

5.1 Setup

Pretraining Datasets. In our experiments, we employed several pre-training datasets crucial for high-quality action recognition models. The primary dataset used was Kinetics-400 (K400) [27], consisting of approximately 240K training videos and 20K validation videos spanning 400 human action categories. This dataset's diversity helps in training robust models. Additionally, we used Kinetics-700 (K700) [42], an extension of K400, which includes about 650K video clips across 700 categories, providing further complexity and variety. These datasets are crucial for pre-training action recognition models, enabling full utilization of the data's diversity by not freezing the backbone and classification head during training.

Implementation Details. DGAM chose the Video Swin Transformer [31] as a backbone due to its superior performance on our proposed ELLAR dataset among existing video action recognition models, and its use of the shifted window approach, which is computationally more efficient than other transformer-based models. DGAM was conducted with five pairs of GIC experts and a multi-head structure. The gamma values of each expert were set to 1.0, 1.5, 2.0, 2.5, and 3.0. The number of GIC module-head pairs was determined experimentally to optimize performance. The classification head adopted an I3D head structure, and each head was trained with a one-to-one match with its corresponding GIC module. DGAM used the same data preprocessing strategy as Video Swin Transformer [31], with AdamW as the optimizer, an initial learning rate of 3×10^{-4}, and a weight decay of 0.05. We also multiplied 0.1 by the

learning rate in order for the backbone architecture to be stabilized during training. The input videos were resized to a spatial dimension of 224×244 and sampled in a temporal dimension of 32 total lengths using stride 2. The batch size for training was set to 128.

Training Strategy. For effective learning and performance enhancement of the DGAM, we used the following initializing strategy: First, we used the Video-Swin-B backbone architecture pre-trained with K400 for initialization of the DGAM model. Then, we changed the classification head structure to fit the ELLAR dataset and fine-tuned it with all layers. These trained backbone network parameters were used to initialize the action recognition backbone of the DGAM model. The five multi-classification heads of DGAM were equally initialized with the parameters of the head obtained after training the Video Swin Transformer with ELLAR.

DGAM was trained using a switching training strategy. First, we froze the gating network and trained the transformer backbone and multiple classification heads. Since the gate network is frozen, the input clips entering the backbone architecture are randomly applied with different levels of GIC, facilitating the training of each head corresponding to its respective GIC. This process allows each head to more optimally fit the output of the GIC module it is paired with. The second step is to train the gating network. After the backbone and heads have been trained in the previous step, the gating network is trained by freezing the heads and the entire backbone network. This way, the gating network learns which heads are most beneficial to use on a sample-by-sample basis. The reason for not training the gate, backbone, and heads together is that the gating network takes a long time to train due to the large number of parameters in the backbone that prevent sufficient gradients from flowing down toward it. We also empirically found learning rate multiplier in order to stabilize the training process of the gating network by setting it to 10, 100, and 1000. Moreover, since it has to be trained together with the heads, there is a high probability that it will not be sufficiently trained on the data. With the proposed simple switching training strategy, the training of DGAM is stable.

5.2 Comparison on ELLAR Dataset

Table 3: Comparison to state-of-the-art on ELLAR dataset.

	Pretrained	**Input Size**	**Top-1**	**Top-5**
ResNet101	K700	$3\times16\times112^2$	10.46	45.69
ResNeXt101	K400	$3\times16\times112^2$	9.63	39.37
DarkLight	IG-65M	$3\times64\times112^2$	28.58	64.31
TimeSformer	K400	$3\times96\times224^2$	15.51	55.96
Video-Swin-B	K400	$3\times32\times224^2$	35.03	68.87
DGAM (Ours)	K400	$3\times32\times224^2$	**38.42**	**74.44**

To evaluate the performance of the proposed baseline method on ELLAR, we compare it with some well-known models in the field of action recognition. As mentioned

in Sec. 2, there are two types of existing action recognition models in terms of architectures: CNN-based and Transformer-based. We adopted 3DResNet101 [26] and 3DResNeXt101 [56], which are frequently used in CNN-based action recognition, and selected TimeSformer [4] and Video Swin Transformer [31], which are frequently used as video classification task backbones in Transformer-based models. We also compared with DarkLight [8], a model that uses both CNN and Transformer self-attention structures specialized for action recognition in low light. DarkLight's CNN encoder utilizes a (2+1)D-34 [49] structure trained on the IG-65M dataset [18]. The IG-65M dataset contains 65 million video clips sourced from Instagram, designed to improve video understanding and action recognition models. The training was validated every 2 epochs, and an early stopping technique was applied to prevent overfitting by stopping training if the performance did not improve by more than 5 epochs based on the validation score. We used top-1 accuracy and top-5 accuracy as evaluation metrics.

The results in Table 3 show that the proposed DGAM model outperforms state-of-the-art models on the ELLAR dataset, achieving the highest performance with a top-1 accuracy of 38.42% and a top-5 accuracy of 74.44%. These results highlight the effectiveness of the DGAM approach, which leverages dynamic adaptive GIC to better handle the variations and challenges of low-light action recognition in the ELLAR dataset, thereby demonstrating its superior capability in action recognition under dynamic and extremely low-light domains.

5.3 Comparison on Dynamic Light Conditions

Table 4: Comparison between 5 state-of-the-art models and our method DGAM on dynamic light conditions and cameras.

	LL				**ELL**			
	Brio-4k	Arducam	C920	Avg.	Brio-4k	Arducam	C920	Avg.
ResNet101	12.77	9.87	10.12	10.92	9.97	9.97	9.97	9.97
ResNeXt101	8.91	9.15	10.84	28.9	13.29	8.18	7.41	9.62
DarkLight	69.88	33.25	26.02	43.05	19.18	9.46	11.00	13.21
TimeSformer	23.86	20.72	16.63	20.40	11.51	9.67	9.72	10.98
Video-Swin-B	75.90	46.72	40.48	54.37	22.25	9.72	11.51	14.49
DGAM (Ours)	**78.07**	**52.53**	**44.58**	**58.39**	**24.4**	9.72	9.72	**14.61**

Table 4 presents a comparison of performances of various models under dynamic illuminance conditions, with LL and ELL conditions. The second row indicates the various cameras used to record videos, each working within a dynamic luminance range. In the LL condition, our proposed DGAM model achieves state-of-the-art performance. Notably, DGAM achieves the highest top-1 accuracy of 78.07% with the Brio-4k camera. This superior performance can be attributed to the brighter output of the Brio-4k camera, which provides better illumination compared to Arducam and C920. On the ELL data, the performance of all models, including ours, is considerably lower. We attribute this to the extreme darkness of the ELL data, characterized by a high number

of zero-value pixels and very low signal-to-ratio, posing a significant challenge for conventional models as well as our own. Despite adversarial factors caused by extremely low-light conditions, our DGAM model still achieves the best top-1 accuracies with the Brio-4k camera in ELL conditions, demonstrating its robustness and effective performance.

5.4 Domain Adaptation and Generalization Performances

Table 5 presents the domain adaptation performance of our proposed method, comparing it with the Video Swin Transformer, which serves as our baseline model. The arrows in the table indicate the training data to test data transition, with LL representing low-light conditions and ELL representing extremely low-light conditions. Both models were trained separately with only LL and ELL conditions data to evaluate their performance in cross-domain scenarios. Our DGAM model demonstrates superior performance across all scenarios. Specifically, in the cross-domain tests, where the training and testing conditions differ from each other (LL⇒ELL and ELL⇒LL), DGAM shows notable improvements, achieving 14.24% and 40.56% of top-1 accuracies compared to Video-Swin-B's 12.77% and 32.27%. These results highlight the effective cross-domain adaptability of our DGAM model to varying illumination of extremely dark conditions.

Table 5: Domain Adaptation Performance under Different Light Conditions.

	In Domain		Cross Domain	
	LL⇒LL	ELL⇒ELL	LL⇒ELL	ELL⇒LL
Video-Swin-B	56.71	15.77	12.77	32.27
DGAM (Ours)	**63.78**	**16.71**	**14.24**	**40.56**

In addition, our method not only achieves SOTA performance on the ELLAR dataset but also demonstrates strong performance on ARID [58] dataset, as shown in Tab. 6. This result highlights the generalization capability of DGAM.

Table 6: Comparison of model performance on ARID dataset.

Model	Pretrained	Top-1	Top-5
ResNet101	K700	71.57	99.03
ResNext101	K400	74.73	98.54
DarkLight	IG-65M	<u>94.04</u>	<u>99.87</u>
Timesformer	K400	81.39	98.26
Video-Swin-B	K400	89.79	99.53
DGAM (Ours)	**K400**	**93.76**	**99.85**

5.5 Ablation Study

Table 7 presents the results of the ablation study conducted to evaluate the contributions of the gating network and multi-head components in our proposed method. The baseline model without both the gating network and multi-head components achieves top-1 accuracy of 35.03% and top-5 accuracy of 66.87%. Adding the multi-head component improves performance, yielding a top-1 accuracy of 38.25% and a top-5 accuracy of 74.36%. Finally, DGAM, incorporating both the gating network and multi-head components, achieves the highest performance with top-1 accuracy of 38.42% and top-5 accuracy of 74.44%, demonstrating the effectiveness of integrating both elements. This also indicates that the significant performance increase is mainly attributed to the inclusion of the gating network, while the multi-head structure further improves performance slightly, indicating that both components play a role in improving the model's accuracy.

Table 7: Ablation study on gating network and multi-head structure.

	Top-1	Top-5
w/o gate + multi-head	35.03	66.87
w/o multi-head	38.25	74.36
Dual	**38.42**	**74.44**

6 Conclusion

This paper tackles the challenging problem of action recognition in extremely low-light environments by introducing the ELLAR dataset and a simple yet strong baseline method that utilizes a Mixture of Experts, named Dual Gamma Adaptive Modulation (DGAM). Existing datasets have struggled to accurately reflect real-world scenarios due to their insufficient low-light representation and inability to handle dynamic illuminance conditions. To address these issues, we developed the ELLAR dataset, which includes a wide variety of scenes, outfits, and multiple cameras to capture extremely low-light environments more realistically. Additionally, we proposed DGAM, a simple yet effective baseline model designed to enhance images adaptively for better action recognition by jointly training image enhancement models and action recognition encoders. Experimental results demonstrate that DGAM outperforms existing methods in extremely low-light environments. Furthermore, DGAM exhibits robust performance in cross-domain settings, showcasing its ability to adapt to dynamic low-light conditions.

Acknowledgments. This work was supported by the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korean Government (Ministery of Science and ICT) (RS-2023-00304695) and by the Daegu Mechatronics and Materials Institute (DMI) grant funded by Daegu Metropolitan City (2024 Project for Expanding the Value Chain of the Robotics Industry and Establishing a Win-Win Collaboration System).

References

1. Abdullah-Al-Wadud, M., Kabir, M.H., Dewan, M.A.A., Chae, O.: A dynamic histogram equalization for image contrast enhancement. IEEE transactions on consumer electronics **53**(2), 593–600 (2007)
2. ArduCam: 1080p low light wide angle usb camera module with microphone for computer, `https://bit.ly/arducam-1080p-low-light`
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: ICCV. pp. 6836–6846 (2021)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? (2021)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
6. Celik, T., Tjahjadi, T.: Contextual and variational contrast enhancement. TIP **20**(12), 3431–3441 (2011)
7. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark (2018)
8. Chen, R., Chen, J., Liang, Z., Gao, H., Lin, S.: Darklight networks for action recognition in the dark. In: CVPR. pp. 846–852 (2021)
9. Cheron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: ICCV (December 2015)
10. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: Potion: Pose motion representation for action recognition. In: CVPR (June 2018)
11. Crescitelli, V., Kosuge, A., Oshima, T.: Poison: Human pose estimation in insufficient lighting conditions using sensor fusion. IEEE Transactions on Instrumentation and Measurement **70**, 1–8 (2021)
12. Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., Harada, T.: You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press (2022), `https://bmvc2022.mpi-inf.mpg.de/0238.pdf`
13. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV **130**, 33–55 (2022)
14. Deng, S., Tian, Y., Hu, X., Wei, P., Qin, M.: Application of new advanced cnn structure with adaptive thresholds to color edge detection. Communications in Nonlinear Science and Numerical Simulation **17**(4), 1637–1648 (2012)
15. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV (December 2015)
16. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: ICCV. pp. 6824–6835 (2021)
17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
18. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: CVPR. pp. 12046–12055 (2019)
19. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV. pp. 5842–5850 (2017)

20. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR (June 2020)
21. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. TIP **26**(2), 982–993 (2016)
22. Huang, S.C., Cheng, F.C., Chiu, Y.S.: Efficient contrast enhancement using adaptive gamma correction with weighting distribution. TIP **22**(3), 1032–1041 (2012)
23. Hussain, A., Khan, S.U., Khan, N., Rida, I., Alharbi, M., Baik, S.W.: Low-light aware framework for human activity recognition via optimized dual stream parallel network. Alexandria Engineering Journal **74**, 569–583 (2023)
24. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the em algorithm. Neural computation **6**(2), 181–214 (1994)
25. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. pp. 1725–1732 (2014)
26. Kataoka, H., Wakamiya, T., Hara, K., Satoh, Y.: Would mega-scale datasets further enhance spatiotemporal 3d cnns? arXiv preprint arXiv:2004.04968 (2020)
27. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
28. Kwon, H., Kim, M., Kwak, S., Cho, M.: Learning self-similarity in space and time as generalized motion for video action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13065–13075 (October 2021)
29. Lee, S., Rim, J., Jeong, B., Kim, G., Woo, B., Lee, H., Cho, S., Kwak, S.: Human pose estimation in extremely low-light conditions. In: CVPR. pp. 704–714 (2023)
30. Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. TIP **27**(6), 2828–2841 (2018)
31. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: CVPR. pp. 3202–3211 (2022)
32. Logitech: Logitech brio webcam with 4k hdr webcam, `https://bit.ly/brio4k-hdr`
33. Logitech: Logitech c920 pro hd webcam, 1080p video with stereo audio, `https://bit.ly/logitech-brio-c920`
34. Maitlo, N., Noonari, N., Ghanghro, S.A., Duraisamy, S., Ahmed, F.: Color recognition in challenging lighting environments: Cnn approach. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). pp. 1–7. IEEE (2024)
35. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. PAMI **42**(2), 502–508 (2019)
36. Na, T., Lee, M., Mudassar, B.A., Saha, P., Ko, J.H., Mukhopadhyay, S.: Mixture of pre-processing experts model for noise robust deep learning on resource constrained platforms. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2019)
37. Poynton, C.: Digital video and HD: Algorithms and Interfaces. Elsevier (2012)
38. Rahman, Z.u., Jobson, D.J., Woodell, G.A.: Multi-scale retinex for color image enhancement. In: ICIP. vol. 3, pp. 1003–1006. IEEE (1996)
39. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

40. Sijbers, J., Scheunders, P., Bonnet, N., Van Dyck, D., Raman, E.: Quantification and improvement of the signal-to-noise ratio in a magnetic resonance image acquisition procedure. Magnetic resonance imaging **14**(10), 1157–1163 (1996)
41. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. NeurIPS **27** (2014)
42. Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., Zisserman, A.: A short note on the kinetics-700-2020 human action dataset. arXiv preprint arXiv:2010.10864 (2020)
43. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
44. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
45. Thong, J., Sim, K., Phang, J.: Single-image signal-to-noise ratio estimation. Scanning **23**(5), 328–336 (2001)
46. Trahanias, P.E., Venetsanopoulos, A.N.: Color image enhancement through 3-d histogram equalization. In: 11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis,. vol. 1, pp. 545–548. IEEE Computer Society (1992)
47. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
48. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: ICCV. pp. 5552–5561 (2019)
49. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018)
50. Tu, Z., Liu, Y., Zhang, Y., Mu, Q., Yuan, J.: Dtcm: Joint optimization of dark enhancement and action recognition in videos. TIP (2023)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
52. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
53. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement (2018)
54. Wei, K., Fu, Y., Yang, J., Huang, H.: A physics-based noise formation model for extreme low-light raw denoising. In: CVPR (June 2020)
55. Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In: CVPR. pp. 5901–5910 (June 2022)
56. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (July 2017)
57. Xu, X., Wang, R., Fu, C.W., Jia, J.: Snr-aware low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17714–17724 (June 2022)
58. Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: 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. pp. 70–84. Springer (2021)

59. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X., Chen, D.S.: A comprehensive survey of vision-based human action recognition methods. Sensors **19**(5), 1005 (2019)
60. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: ACMMM. pp. 1632–1640 (2019)
61. Zhou, H., Dong, W., Liu, X., Liu, S., Min, X., Zhai, G., Chen, J.: Glare: Low light image enhancement via generative latent feature based codebook retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)