Supplementary Materials of 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

1 About ELLAR Dataset

We developed a novel dataset titled Extremely Low-Light condition Action Recognition (ELLAR). This dataset is divided into two categories based on the illumination levels: low-light (LL) and extremely low-light (ELL). The LL section includes recordings from three outdoor locations under low-light conditions, while the ELL section features footage from two indoor settings with extremely low-light conditions. ELLAR was created using five actors who wore 17 different colors of clothing and filmed with five different scenes to capture a wide range of real-world scenarios. Recognizing the importance of color information in dark settings [2,6], we intentionally varied the colors of the actors' clothing. ELLAR includes 12 common daily atomic action classes such as running, turning, and sitting. The videos are approximately 3-4 seconds long and are in AVI format.

2 Dataset Collection and Specifications

In this section, we describe how the data was collected and what efforts were made to ensure that ELLAR was captured under extremely low-light conditions, covering a wide range of illuminance levels within the scope of extremely dark settings.

2.1 Data Collection Process

To create ELLAR dataset, we utilized the Robot Operating System (ROS) [7]. Specifically, three cameras were supposed to capture human actions simultaneously, so the ROS rosbag package [8] was used to run all three cameras. Three cameras were used for capturing ELLAR: Brio-4K (Brio-4K) [4], Logi-C920 (C920) [5], and Arducam-B0200 (Arducam) [1]. The Brio-4k and C920 are

⋆ Equal contributions

† Corresponding Author

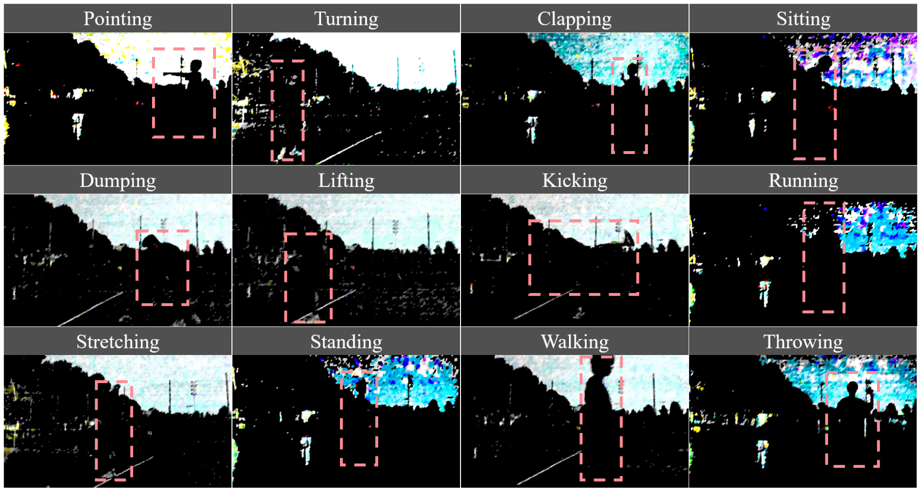

Fig. 1: The samples of each 12 classes in ELLAR. All images are applied Gamma Intensity Correction (GIC) and additional contrast enhancement for visualization.

commonly used webcams, representing standard consumer usage, while the Arducam is a widely used embedded camera module in robotics for machine vision applications. The configured parameters that were used to capture the dataset are detailed in Tab. 1.

Table 1: Configured camera parameters.

Camera	**Brightness**	**Contrast**	**Saturation**	**Gain**	**Sharpness**
Brio-4K	128 (0∼255)	128 (0∼255)	128 (0∼255)	1 (0∼255)	128 (0∼255)
C920	128 (0∼255)	128 (0∼255)	128 (0∼255)	0 (0∼255)	128 (0∼255)
Arducam	0 (-64∼64)	32 (0∼64)	64 (0∼128)	0 (0∼100)	2 (0∼6)

2.2 Action Classes

ELLAR is an extreme low-light action recognition dataset consisting of 12 atomic actions. Each action class was chosen to represent frequently occurring behaviors in the real world, and each action was carefully defined for consistency. Of the 12 actions, 12 samples were taken for classes where the action could be recognized even when turning around, and 9 samples for classes where it could not. For the classes Running and Walking, 10 non-overlapping paths were pre-selected and filmed. For each class, half of the actions were recorded strictly at the locations specified by the defined rules, and the other half were recorded with more free actions at random locations to capture both uniformity and diversity in the data. Details are shown in Tab. 2.

Table 2: The definitions of each of 12 classes in ELLAR. The "Fixed" and "Random" columns in the table represent the number of fixed locations of action and random locations of actions for each class in ELLAR.

Class	Definition	Fixed	Random
Walking	Walk on camera for 3 seconds	10	10
Running	Run in front of the camera for 3 seconds	10	10
Lifting	Squat down, pick up an object, and stand up	9	9
Kicking	Kick forward	9	9
Dumping	Squat down, drop an object, and stand up	9	9
Sitting	Squatting on a chair or floor	12	12
Standing	Rise from a chair or the floor	12	12
Turning	Turn 180 degrees to the left and right	12	12
Stretching	Stretch arms upward	12	12
Throwing	Throw a ball forward	9	9
Pointing	Point in front/left/right direction	9	9
Clapping	Clap from the center of the body	9	9

2.3 Scene & Variation

The ELLAR dataset ensures a comprehensive range of lighting conditions and unique data samples by meticulously diversifying action locations, as depicted in Fig. 2. We employed a two-stage filming process to guarantee sample variety and dataset consistency. In the first stage, termed the "fixed" session, each actor performed actions from predefined positions: left, right, and center relative to the camera, with their orientation varying between front, back, and side views. This setup aimed to maintain a minimum level of uniformity across the dataset.

The second stage, known as the "random" session, involved actors performing actions from random positions and angles. This approach allowed ELLAR to capture a wide array of samples while preserving essential dataset uniformity. Through this two-stage filming process, ELLAR successfully balances consistency and diversity, ensuring robustness in action recognition under varied lighting conditions. The ELLAR dataset includes a diverse range of lighting conditions and settings to ensure comprehensive coverage of real-world scenarios. The dataset comprises five distinct scenes, as illustrated in Fig. 3, labeled A, B, C, D, and E. Scenes A and B are recorded on different basketball courts, Scene C in an indoor building corridor, Scene D at a tennis court and its surrounding park, and Scene E within a classroom. These scenes are categorized based on their lighting conditions: Scenes A, B, and D are classified as low-light (LL) environments, while Scenes C and E are categorized as extremely low-light

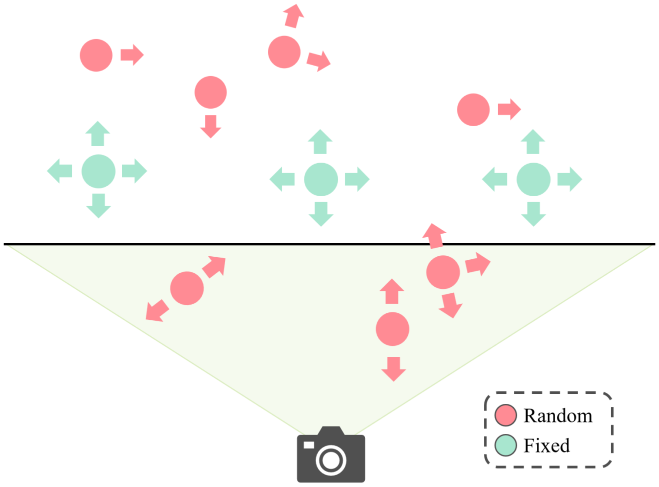

Fig. 2: Example of setting where actions occur when collecting data. Fixed, in green, is the same for all actions, while Random, in red, is taken from a different location for each shot. The arrows indicate the angle at which the actor is standing.

(ELL) environments. The classification of lighting conditions into LL and ELL is based on Signal-to-Noise-Ratio (SNR) values. ELLAR also takes into account the impact of colors in dark settings, supported by studies such as [2, 6]. To ensure a comprehensive range of pixel values, actors wore a mix of light and dark outfits, as illustrated in Fig. 4. Light outfits are marked with a red border, while dark outfits have a gray border. The classification criterion was based on whether more than 50% of the upper and lower clothing was in dark shades. This approach enables ELLAR to encompass a broad spectrum of color variations, contributing to the dataset's richness and practical applicability in real-world scenarios.

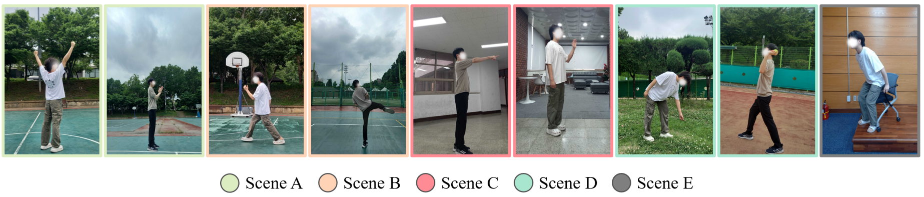

Fig. 3: Example of five scenes from the ELLAR dataset. Scenes A, B, and D are categorized as low light environments (LL), while scenes C and E are categorized as extremely low light environments (ELL). In each scene, the action was captured from different angles and in different locations.

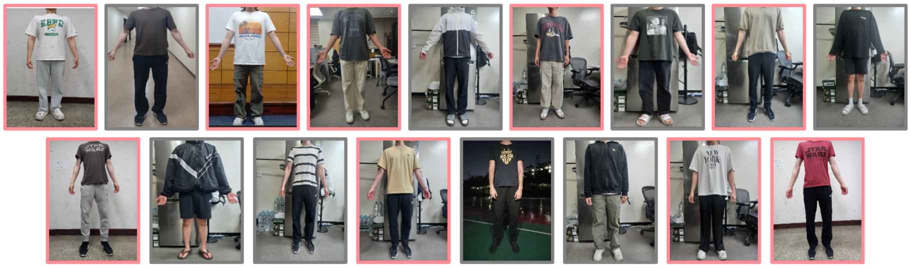

Fig. 4: Example outfits worn by Actors during the filming of the ELLAR dataset. To represent the diversity of clothing colors, the ratio of light to dark clothing is halved. The picture with the red borderline is the light clothing and the picture with the gray borderline is the dark clothing.

3 Data Analysis

3.1 Signal-to-Noise Ratio (SNR) analysis

The Signal-to-Noise Ratio (SNR) is a measure used to compare the level of a desired signal to the level of background noise. It is typically expressed in decibels (dB). When calculating SNR for a video, each frame of the video is considered as an individual image. The SNR can be computed using the following equation:

$$\mathrm{SNR}_{\mathrm{video}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mu_{\mathrm{signal},i}}{\sigma_{\mathrm{noise},i}} \tag{1}$$

where:

- μ_{signal} is the mean pixel intensity of the signal.
- σ_{noise} is the standard deviation of the pixel intensities representing the noise.
- N is the total number of frames in the video.
- $\mu_{\mathrm{signal},i}$ and $\sigma_{\mathrm{noise},i}$ are the mean and standard deviation of the i-th frame, respectively.

Table 3 provides a comparative analysis of the 0-pixel Ratio and Signal-to-Noise Ratio (SNR) for Low Light (LL) and Extremely Low Light (ELL) conditions within the ELLAR dataset. It highlights the significant differences in pixel information between the two lighting conditions. Specifically, the 0-Pixel Ratio is extremely high in both LL and ELL conditions, with LL showing 96.29% and ELL nearly reaching 99.98% indicating that most pixels are zero-valued and correspond to very dark regions. Furthermore, the SNR values are notably low, with LL at 0.060 and ELL at a mere 0.003, emphasizing the challenging nature of extracting meaningful information from these frames. This stark contrast underscores the difficulty of action recognition tasks under such low-light environments and highlights the necessity for advanced enhancement techniques to extract usable information from these challenging conditions.

Table 3: Analysis of average 0-Pixel Ratio and SNR (Signal-to-Noise Ratio) for each lighting condition in the ELLAR dataset.

	LL	**ELL**
0-Pixel-Ratio(%)	96.29	99.98
SNR	0.060	0.003

3.2 Intensity Distribution of ELLAR

As shown in Tab. 4, extremely low-light conditions (ELL) have very limited signals, which degrade the performance of models and present a particularly challenging problem. While the proposed method outperforms the baseline method on ELL, as indicated in Tab. 4 of the main paper, recognition accuracy remains low, suggesting limited risk of overfitting.

Table 4: Intensity distribution of ELLAR by light condition.

Light	**0**	**1–2**	**3–8**	**9–32**	**33–128**	**129–255**
LL	95.75%	1.89%	1.12%	1.09%	0.12%	0.03%
ELL	99.99%	6.5e-3%	1.9e-3%	3.4e-4%	1.2e-4%	1e-6%

4 Details about Methods & Experiment

4.1 Cross Domain Experiment

Table 5: Cross domain adaptability by light conditions and cameras.

	Cross Domain					
	LL ⇒ ELL			**ELL ⇒ LL**		
	Brio-4K	C920	Arducam	Brio-4K	C920	Arducam
Video-Swin-B	15.86	12.73	9.72	55.42	10.31	31.08
DGAM (Ours)	**19.95**	**13.04**	9.72	**63.13**	**24.82**	**33.73**

Table 5 presents the domain adaptation performance of our proposed method in terms of light conditions and cameras, comparing it with the Video Swin Transformer [3]. The terms before the arrows denote the source domains and the ones after the arrows indicate destination domains, 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. DGAM shows notable improvements, achieving 19.95% and 13.04% of top-1 accuracies compared to Video-Swin-B's results of 15.86% and 12.73% on Brio-4K and C920 respectively. In the case where the source is ELL and the destination is LL, DGAM obtains 63.13%, 24.82%, and 33.73%, outperforming state-of-the-art that gain 55.42%, 10.31%, and 31.08% on Brio-4K, C920, and Arducam.

4.2 Impact of clothing color on performance

As shown in Table 6, in low-light environments (LL), the model performs significantly better when individuals wear short sleeves or bright-colored clothing, likely due to greater contrast and exposure of skin. Specifically, the top-1 accuracy for short sleeves or shorts is 59.4%, compared to 42.68% for those wearing clothes that cover more skin. Similarly, bright-colored clothing achieves a top-1 accuracy of 57.54%, while dark-colored clothing yields a lower accuracy of 35.01%.

However, in extremely low-light conditions (ELL), these advantages diminish, and the model struggles to maintain accuracy. The top-1 accuracy for short sleeves or shorts drops to 14.88%, only slightly better than 14.35% for individuals wearing long sleeves or pants. Likewise, bright-colored clothing results in a top-1 accuracy of 13.47%, while dark-colored clothing performs slightly better at 15.98%. This indicates that while clothing style and color play an important role in low-light recognition, their impact becomes less significant as lighting conditions worsen.

Table 6: Top-1 accuracy comparison by clothing on ELLAR.

		LL	**ELL**
Short-Sleeves/Shorts	**True**	59.4	14.88
	False	42.68	14.35
Clothes color	**Bright**	57.54	13.47
	Dark	35.01	15.98

References

1. ArduCam: 1080p low light wide angle usb camera module with microphone for computer, https://bit.ly/arducam-1080p-low-light 1
2. 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) 1, 4
3. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: CVPR. pp. 3202–3211 (2022) 6
4. Logitech: Logitech brio webcam with 4k hdr webcam, https://bit.ly/brio4k-hdr 1
5. Logitech: Logitech c920 pro hd webcam, 1080p video with stereo audio, https://bit.ly/logitech-brio-c920 1
6. 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) 1, 4
7. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., et al.: Ros: an open-source robot operating system. In: ICRA workshop on open source software. vol. 3, p. 5. Kobe, Japan (2009) 1
8. Staples, G.: rosbag, https://wiki.ros.org/rosbag 1