HELLENIC REPUBLIC
MINISTRY OF DEVELOPMENT AND INVESTMENT
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
**HELLENIC FOUNDATION FOR RESEARCH AND INNOVATION**

This project is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union – NextGenerationEU (Implementation body: HFRI)

# Greece 2.0
# Basic Research Financing Action
# (Horizontal support of all Sciences)
# Sub-action 1
# Funding New Researchers

LEARNER

Project Title

**SLAM AND PATH PLANNING MIDDLEWARE PACKAGE FOR ROBOTS IN CHALLENGING ENVIRONMENTS**

Project Duration

**20 November 2023 – 19 November 2025**
**24 Months**

| Project Acronym | Project No |
|---|---|
| **LEARNER** | **015339** |

| Deliverable No | Deliverable Title |
|---|---|
| **D2.4** | **Journal paper on social-aware mapping and PP** |

Deliverable Completion Date

**19 September 2025**

## Table of Contents

*Table of Contents* .................... 2

*Document Revision History* .................... 2

*List of Acronyms* .................... 2

*1. Introduction* .................... 2

*2. Core Contributions of the Paper* .................... 2

*3. Integration with LEARNER Middleware* .................... 3

*4. Conclusion* .................... 3

*Appendix A* .................... 3

## Document Revision History

| Version | Date | Notes |
|---|---|---|
| 1.0 | 19/09/2025 | First version document the paper submitted regarding the developed social-aware mapping and path planning techniques. |

## List of Acronyms

| Acronym | Meaning |
|---|---|
| LEARNER | SLAM and Path Planning Middleware Package for Robots in Challenging Environments |
| PP | Path Planning |

## 1. Introduction

This deliverable summarizes the core contributions of the paper "Low-Light Adaptation for Action Recognition-Enabled Robot Navigation" (see Annex A) and clarifies its role within the LEARNER project.

Specifically, this work strengthens the social-awareness layer of the LEARNER middleware by improving the robot's ability to (i) perceive human presence and actions reliably, (ii) interpret social context, and (iii) adjust its navigation behavior in real-time, even under low-light conditions such as emergency scenarios or night-time operations.

## 2. Core Contributions of the Paper

The paper introduces a data-centric pipeline that adapts state-of-the-art human pose estimation and action recognition frameworks to operate under low-illumination conditions:

- **Synthetic Low-Light Dataset Generation:**
  A large-scale dataset was generated by transforming the COCO dataset into a low-light domain using CycleGAN-Turbo. This approach retains original geometric and pose annotations,

creating a paired dataset that simultaneously covers well-lit and dark scenarios without manual annotation.

- **Adaptation of Pose Estimation Model:**
  AlphaPose was re-trained on this combined dataset, yielding a robust keypoint detector capable of maintaining accuracy in low signal-to-noise conditions. This directly mitigates perception failures common in darkness or smoke-filled environments.
- **Confidence-Based Keypoint Filtering:**
  The system incorporates a keypoint confidence histogram-based filtering mechanism to discard unreliable detections, thus improving downstream action recognition accuracy.
- **Action Recognition Integration:**
  The filtered skeleton keypoints are fed into MMAction2 for temporal action classification, enabling the robot to recognize human behaviors such as walking, running, bending, or falling with improved robustness in low-light conditions.

Experimental results confirm statistically significant improvements in keypoint precision and action classification accuracy compared to models trained solely on well-lit datasets.

## 3. Integration with LEARNER Middleware

The work is directly aligned with WP2 objectives and complements Deliverables D2.2 and D2.3:

- **Enrichment of the Hybrid Map:**
  The recognized human keypoints and action labels are projected onto the hybrid map's semantic layer (D2.2), allowing the system to differentiate not only between humans and obstacles but also between static (standing) and dynamic (moving, running) agents.
- **Costmap Adaptation for Social-Aware PP:**
  The action recognition output modifies the cost function used for PP. For example, humans identified as running or engaged in critical tasks may result in widened safety margins, while stationary humans may be navigated around with minimal detour.
- **Robustness under Adverse Conditions:**
  The low-light adaptation ensures that these features remain operational in environments with reduced visibility, thus preserving both safety and mission continuity.

This integration supports adaptive safety bounds, dynamic path replanning, and enhanced human-robot interaction, as outlined in the LEARNER conceptual architectur.

## 4. Conclusion

This deliverable introduces a major advancement towards social-aware PP under challenging conditions. By shifting computationally heavy low-light adaptation to the training phase, the proposed approach ensures a lightweight inference pipeline, making it suitable for real-time deployment on the selected robotic platforms.

The results of this work will be used to (i) enrich the hybrid map with social and action-awareness, (ii) improve PP behavior in human-centric scenarios, and (iii) form the basis for evaluating LEARNER's performance in WP3 test cases. The full paper is provided in Appendix A for detailed methodology and experimental results.

## Appendix A

*Article*

# Low-Light Adaptation for Action Recognition-Enabled Robot Navigation

**Marilena Anastasiou [1],* and Loukas Bampis [2]**

[1] Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece; manastasi@ee.duth.gr
[2] Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece; lbampis@ee.duth.gr
* Correspondence: manastasi@ee.duth.gr

**Abstract:** Effective Human-Robot Interaction (HRI) requires robots to accurately perceive and understand human actions; yet, the performance of vision-based systems degrades significantly in non-ideal, low-light environments. This limitation poses a critical challenge for applications where a robotic agent needs to navigate close to human actors in darkness, such as in emergency response scenarios. Existing solutions often rely on computationally expensive real-time image enhancement or require large, manually annotated low-light datasets, which are difficult to acquire. This paper proposes a novel and efficient data-centric approach to overcome this challenge. Instead of enhancing images at inference time, we shift the adaptation to the training phase. We introduce a synthetic low-light dataset generated from the popular COCO collection, using the CycleGAN-Turbo model for unsupervised image-to-image transformation. This synthetic dataset and the original one are then fed to the state-of-the-art AlphaPose model for pose estimation. By specializing the pose estimator for dark conditions, our method improves the 2D human keypoints detection directly from low-light video feeds. These keypoints are then formatted and passed to the MMAction2 framework for final action classification. We demonstrate through extensive experiments, evaluated with confusion matrices and keypoint score analysis, that our method improves action recognition accuracy in low-light scenarios compared to the same model trained on standard datasets. Our findings present a practical and effective pipeline for developing robust HRI systems, capable of navigating among humans in challenging lighting conditions.

**Keywords:** action recognition; human-robot interaction; low-light conditions; pose estimation; deep learning; synthetic data; autonomous navigation

Received:
Revised:
Accepted:
Published:

**Citation:** Anastasiou, M.; Bampis, L. Low-Light Adaptation for Action Recognition-Enabled Robot Navigation. *Robotics* **2025**, *1*, 0. https://doi.org/

## 1. Introduction

The extensive use of autonomous systems, particularly robots, in human-centric environments such as homes, workplaces, and public spaces necessitates advanced perceptual capabilities. For safe and intuitive Human-Robot Interaction (HRI), a robot must not only navigate its environment but also understand human presence, intent, and actions. The ability to recognize human actions, from simple gestures to complex activities, is a cornerstone of this understanding, enabling applications ranging from collaborative manufacturing [1] and assistive healthcare [2], to search and rescue operations [3].

While many action recognition models demonstrate impressive performance in ideal, well-lit conditions, their accuracy deteriorates drastically when deployed in the real world. Vision-based systems are particularly vulnerable to adverse lighting, such as dusk, night-time, or poorly lit indoor areas. This performance gap is a significant barrier to the deployment of robots in 24/7 roles, including nighttime security surveillance [4], elder care monitoring [5], or emergency response scenarios [6], where power outages are common. Thus, extracting useful data from such inputs requires sophisticated cognitive systems that are capable of adapting their understanding of the environment to specific operational conditions.

Deep learning has become the standard approach for solving complex perception tasks, with action recognition typically framed as a spatio-temporal classification problem. To address the challenge of low-light conditions, two primary strategies have emerged in the literature. The first involves applying image enhancement or restoration algorithms as a pre-processing step before feeding the data to a standard action recognition model. However, these methods can introduce artifacts and are often too computationally expensive for real-time robotic applications [7]. A second strategy is to train models directly on large-scale, annotated low-light datasets. The primary obstacle here is the immense difficulty and cost associated with collecting and annotating such data, leading to a scarcity of suitable training resources [8].

To bridge this gap, this paper introduces a novel and efficient pipeline for low-light action recognition tailored for robotic navigation. Our core contribution is a data-centric adaptation strategy that improves the recognition results without the need for image enhancement during deployment. We propose the creation of a large-scale, synthetic low-light training dataset by transforming the existing COCO dataset [9]. Using this new resource, we train the AlphaPose model [10], a state-of-the-art pose estimator, to specialize in extracting accurate human skeleton keypoints directly from both well-lit and dark, noisy images. The keypoints generated by our adapted AlphaPose model are processed into the required format to serve as direct input for the action classification stage of the MMAction2 framework [11], decoupling pose estimation from action recognition. This approach concentrates the heavy computational work in the offline training phase, resulting in a lightweight and effective inference pipeline. The main contributions of this work are:

- A novel methodology for automating the formulation of a large-scale, synthetic low-light dataset with ground-truth annotations for training pose estimation models.
- The successful adaptation of the AlphaPose model for robust keypoint detection in darkness, shifting the computational load from inference to training.
- An integrated pipeline connecting the adapted AlphaPose output to the MMAction2 framework by dynamically filtering weak keypoint detections.

The remainder of this paper is organized as follows. Section 2 reviews the related work in action recognition and low-light adaptation. Section 3 details our proposed methodology, while Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper and discusses future work.

## 2. Related Work

This section reviews the foundational literature in human pose estimation and action recognition. A broad overview of the field can be found in the comprehensive survey by Kong and Fu [12]. However, emphasis is placed on the challenges and existing approaches in low-light conditions.

Human Pose Estimation (HPE) is a cornerstone of human-centric robotics, aiming to localize anatomical keypoints of the human body through the perception system of an autonomous platform. Modern approaches are dominated by deep learning and can be broadly categorized into top-down and bottom-up methods. Top-down approaches first detect human bounding boxes in an image and then perform single-person pose estimation within each box. This paradigm, employed by prominent models like AlphaPose [10], often yields high accuracy. AlphaPose is a comprehensive system that performs whole-body (including face, hands, and feet) multi-person pose estimation and tracking in real-time. A key architectural innovation in this area was the High-Resolution Network (HRNet) [13,14], which maintains high-resolution representations throughout the network by connecting multi-resolution subnetworks in parallel. This avoids the resolution loss and recovery seen in typical encoder-decoder frameworks, allowing more precise keypoint heatmap prediction. More recently, Vision Transformers have been successfully adapted for this task. ViTPose [15] demonstrated that a plain Vision Transformer backbone can achieve state-of-the-art results with a simple decoder, highlighting the power of transformer architectures for learning rich feature representations for pose. Bottom-up approaches, conversely, detect all keypoints in an image and then group them into individual human instances. While these methods can be faster in crowded scenes, our work aligns with the top-down paradigm due to its high accuracy on a per-person basis, which is critical for reliable action recognition and HRI tasks [16].

As models have grown in complexity, there has been a push towards more efficient and lightweight architectures. Lite-HRNet [17] and the subsequent Dite-HRNet [18] focused on creating efficient high-resolution networks for resource-constrained environments. Lite-HRNet replaced costly convolutions in HRNet with more efficient shuffle blocks, while Dite-HRNet introduced dynamic, input-dependent components to further optimize the trade-off between performance and complexity.

Once human poses are estimated, the resulting skeleton data provides a powerful and efficient representation for action recognition. This approach is robust to variations in appearance, background, and lighting. Early deep learning models for skeleton-based action recognition utilized Recurrent Neural Networks (RNNs) to model the temporal evolution of joints. However, recent advancements tend to architectures that can better capture the spatio-temporal structure of the skeleton. The Spatial Temporal Graph Convolutional Network (ST-GCN) [19] was a pioneering work that modeled the skeleton as a graph, where joints are nodes and natural bone connections are edges. By applying convolutions over both spatial and temporal dimensions, ST-GCN can effectively learn the complex dynamics of human actions. Another dominant approach involves processing video frames directly. SlowFast Networks [20] proposed a dual-pathway architecture: a "Slow" pathway operating at a low frame rate to capture spatial semantics, and a lightweight "Fast" pathway at a high frame rate to capture fine-grained motion. This design efficiently models both spatial and temporal aspects of an action. The MMAction2 toolbox [11] is a comprehensive, open-source toolkit that implements a wide variety of state-of-the-art action recognition models, providing a flexible platform for rapid development and evaluation.

Given the strong relationship between pose and action, some research has explored multi-task frameworks that handle both tasks simultaneously. Luvizon et al. [21] proposed a single, end-to-end trainable architecture for real-time 3D pose estimation and action recognition. By using a differentiable soft-argmax layer to regress joint coordinates from heatmaps, their model can backpropagate the action recognition loss through the entire network. In this manner, the two tasks can benefit from shared representations and joint optimization, demonstrating that their unification can lead to efficient and effective models.

The performance of standard vision-based models degrades significantly in low-light or nighttime conditions due to a low signal-to-noise ratio (SNR), motion blur, and a lack of visual detail. Addressing this challenge has become an active area of research, resulting in the development of several specialized datasets. For pose estimation, the ExLPose dataset [22] proposed a specifically designed camera system capable of capturing aligned, well-lit, and poorly exposed images to simulate dark environments. However, this approach is tailored to the development and explicit usage of custom equipment that cannot always be available. Furthermore, XPose [23] introduced a novel data collection of well-lit and low-light images specifically focusing on hand's poses. To address solely the action recognition task, the ARID dataset [24] was one of the first to focus on action recognition in dark videos, highlighting the importance of real-world recordings with distinct characteristics that cannot be fully replicated by synthetically darkening well-lit videos. The ELLAR dataset [25] further pushed the boundary by providing videos in "extremely low light" conditions. Hira et al. [26] proposed the Delta Sampling R-BERT model, specifically tackling the challenges of limited and low-light data in the ARID dataset by introducing a novel frame selection strategy and leveraging transfer learning. Nevertheless, the above action recognition datasets are not tailored to robotic applications due to the lack of pose estimation labels.

## 3. Methodology

This section details our proposed pipeline for robust human action recognition in low-light conditions, designed for deployment on an autonomous robotic platform. Our approach is centered on a data-centric adaptation strategy, which shifts the computational burden of low-light processing from online inference to the offline training phase. The pipeline, illustrated in Fig. 1, consists of three main stages: (i) synthesizing a large-scale pose estimation dataset through low-light adaptation, (ii) training a specialized pose estimator on this synthetic data, (iii) filtering noisy keypoint detections, and (iv) integrating the robust pose output with a state-of-the-art action recognition framework.

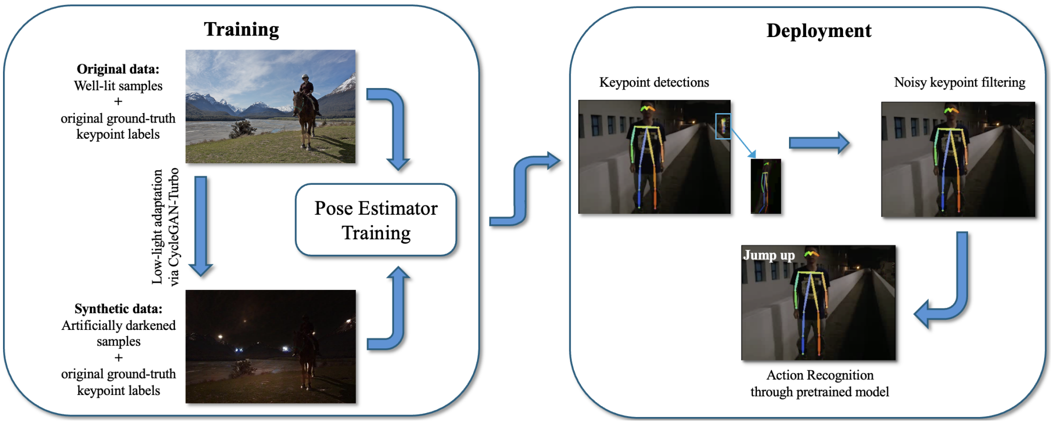
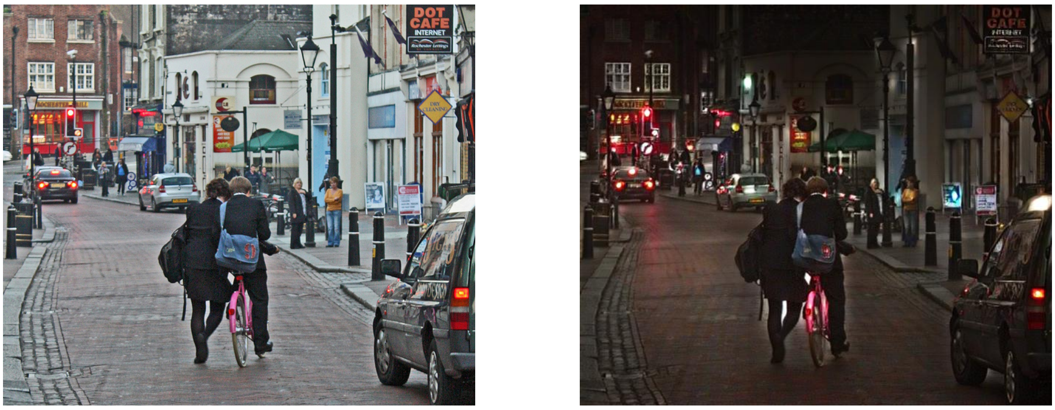

Training
Original data:
Well-lit samples
+
original ground-truth keypoint labels
Low-light adaptation via CycleGAN-Turbo
Pose Estimator Training
Synthetic data:
Artificially darkened samples
+
original ground-truth keypoint labels
Deployment
Keypoint detections
Noisy keypoint filtering
Jump up
Action Recognition through pretrained model

**Figure 1.** A high-level block diagram illustrating the complete proposed pipeline. The original labeled data are synthetically augmented to synthesize an artificial low-light dataset. Using both the original and the augmented data, a pose estimation architecture is trained. The detection results are then filtered to reduce noisy keyponts, and the remaining set is fed to a pretrained action recognition model.

### 3.1. Low-Light Adaptation

The foundation of our method is the creation of a specialized training dataset that enables a pose estimation model to learn features robust to low-light conditions. Instead of relying on real-world low-light data, which are scarce and difficult to annotate, we synthetically generate them. Our method requires an input set of image samples $I_l$ recorded under well-lit conditions. Those samples should contain humans and the corresponding ground-truth labels of their joint poses $P_l$, spanning a wide range of possible stances.

To create a low-light counterpart to $I_l$, we employ CycleGAN-Turbo [27], a highly efficient, unsupervised image-to-image translation model, based on CycleGAN [28] and the Stable Diffusion Turbo (SD-Turbo) architecture [29]. We utilized a pre-trained version of the model specialized for day-to-night conversion, which was trained using unpaired image collections and adapted for a single-step translation process. This approach eliminates the need for iterative denoising steps, which is common in other diffusion models. By applying the pre-trained day-to-night CycleGAN-Turbo generator to every image in $I_l$, we produced a visually plausible, low-light adaptation dataset, which we refer to as $I_d$. A critical advantage of this approach is that the geometric content and human poses within the images remain identical to their daytime counterparts. Therefore, the ground-truth pose annotations $P_l$ are directly transferable to their dark equivalents ($P_d$) without any manual effort. We then combined the original and synthetic datasets to create a comprehensive training set, $I_{ld} = I_l \cup I_d$ and its ground-truth $P_{ld} = P_l \cup P_d$, to expose our pose estimation model to a wide variety of lighting conditions.



(**a**) Original COCO Image

(**b**) Synthetic Low-Light Version

**Figure 2.** Example showing (a) an original image from the COCO dataset [9] and (b) its corresponding synthetically generated low-light version using our CycleGAN-Turbo-based [27] approach.

### *3.2. Robust Pose Estimation*

With the training data prepared, the next stage is to train the human pose estimation model. We selected AlphaPose [10] as our pose estimation backbone. AlphaPose is a state-of-the-art, real-time, multi-person pose estimation and tracking system that follows a top-down paradigm, making it highly suitable for HRI, where per-person accuracy is critical. The pipeline begins by using an off-the-shelf person detector, such as YOLOv3 [30], to locate human bounding boxes. These cropped regions are then fed into a Single-Person Pose Estimator (SPPE) for keypoint regression.

The training objective is centered on the Symmetric Integral Keypoint Regression (SIKR) method. This approach calculates a direct regression loss based on the final predicted coordinates. The predicted keypoint coordinate ($\hat{\mu}$) is computed by taking the mathematical expectation over a normalized probability heatmap ($p_x$). This is also known as a soft-argmax operation:

$$\hat{\mu} = \sum_x x \cdot p_x \tag{1}$$

The training loss is then the $L_1$ norm between the predicted coordinate $\hat{\mu}$ and the ground-truth coordinate $\mu$:

$$\mathcal{L}_{\text{reg}} = \|\mu - \hat{\mu}\|_1 \tag{2}$$

Following the implementation in [10], a two-step process is used to obtain the normalized probability heatmap $p_x$ from the raw network logits $z_x$; thus, decoupling confidence prediction from localization:

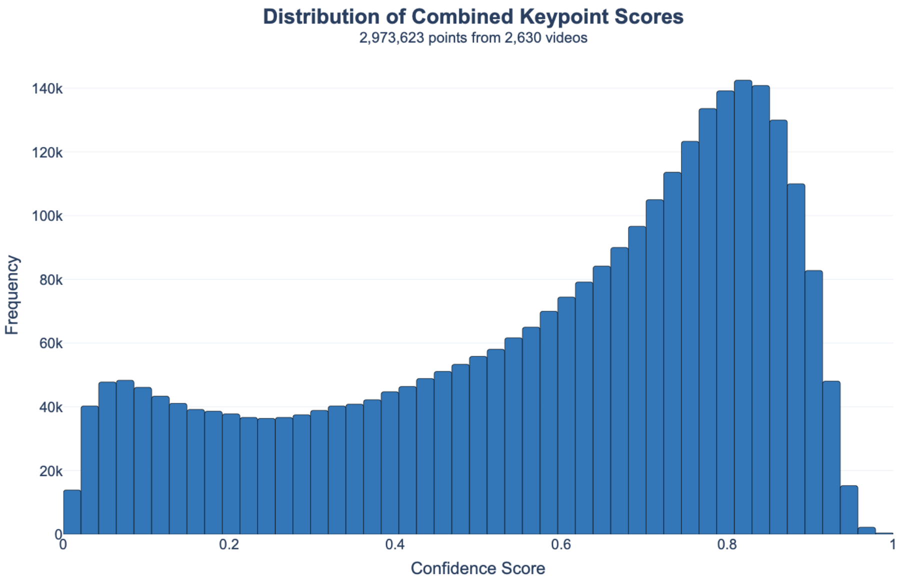


**Figure 3.** Histogram of keypoint confidence scores, aggregated from 2,973,623 detections across 2,630 videos. The distribution is clearly bimodal, confirming the presence of two underlying populations: a smaller mode of low-confidence, likely imprecise detections (scores < 0.3) and a dominant mode of high-confidence, reliable detections (scores > 0.7).

1. **Generate Confidence Heatmap (C):** An element-wise sigmoid function is applied to the logits $z_x$ to produce a confidence value $c_x$ for each pixel. The joint's overall confidence is the maximum value in this map.

$$c_x = \text{sigmoid}(z_x) \quad \text{and} \quad \text{conf} = \max_x(C) \tag{3}$$

2. **Generate Probability Heatmap (P):** The confidence heatmap $C$ is then globally normalized by its sum to produce the final probability heatmap $_x$, which is used in the integral regression.

$$p_x = \frac{c_x}{\sum_x c_x} \tag{4}$$

For our implementation, we utilize the version of AlphaPose built upon a ResNet-50 backbone. To evaluate the effectiveness of our data-centric strategy, we trained two separate instances of the above model. The **Baseline Model** ($M_b$) was trained exclusively on the original well-lit images $I_l$ using the $P_l$ labels, and it represents the standard performance of a pose estimation model without the inclusion of our proposal low-light adaptation. Furthermore, the **Adapted Model** ($M_a$) corresponds to our proposed model, which was trained on the combined dataset $I_{ld}$. The objective is for this model to learn features that are invariant to lighting conditions, enabling it to perform robustly on both bright and dark inputs. Both models were trained until convergence using the standard loss functions and optimization parameters specified in the official AlphaPose implementation.

### *3.3. Filtering Noisy Keypoints*

The confidence scores produced by AlphaPose are critical indicators of keypoint quality. In low-light conditions, many keypoints may be detected with low confidence, introducing noise into the action recognition stage. To mitigate this, we developed a systematic process to find an optimal threshold for filtering out unreliable keypoints. Such a threshold should be tailored to the specific environmental conditions and characteristics of each targeted use case. Therefore, we implement an approach for dynamically computing the filtering threshold based on the detections' statistics.

First, we analyze the distribution of all keypoint confidence scores generated by the detection model. The distribution of such a set follows the bimodal model since there are two underlying populations of scores: one corresponding to correctly localized keypoints and another to noisy, low-confidence detections. An instance of such a distribution, plotted as a histogram, is shown in Fig. 3. We evaluate three distinct statistical methods to separate these two populations:

- **Gaussian Mixture Model (GMM):** We model the score distribution $p(s)$ as a mixture of two Gaussian distributions:

$$p(s) = \pi_1 \mathcal{N}(s|\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(s|\mu_2, \sigma_2^2). \tag{5}$$

  The threshold $T_{GMM}$ is then calculated as the midpoint between the means ($\mu_1, \mu_2$) of the two fitted components:

$$T_{GMM} = \frac{\mu_1 + \mu_2}{2}. \tag{6}$$

- **Weibull Mixture Model (WMM):** This method represents the score distribution using a two-component Weibull mixture. The threshold $T_{WMM}$ is determined by finding the intersection point of the two weighted Probability Density Functions (PDFs), which is solved numerically using:

$$T_{WMM} = \arg\min_{s \in [0,1]} |p_1 \cdot f_1(s; \alpha_1, \beta_1) - (1 - p_1) \cdot f_2(s; \alpha_2, \beta_2)|, \tag{7}$$

  where $f_k(s; \alpha_k, \beta_k)$ is the Weibull PDF for component $k$ and $p_1$ is its mixture proportion.
- **Otsu's Method:** Borrowed from image processing, this method finds a threshold $T_{Otsu}$ that maximizes the inter-class variance $\sigma_B^2(t)$ between the two groups of scores it separates:

$$T_{Otsu} = \arg\max_t \sigma_B^2(t), \tag{8}$$

  where the inter-class variance $\sigma_B^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2$ is a function of the class probabilities ($\omega_k$) and class means ($\mu_k$) determined by the threshold $t$.

### *3.4. Integrated Action Recognition*

The final stage of our pipeline is to classify human actions based on the skeleton sequences generated by our trained pose estimators. We utilize the MMAction2 [11] open-source toolbox, a comprehensive and modular framework for video understanding. Within this framework, we chose the Spatial-Temporal Graph Convolutional Network (ST-GCN) [19] for skeleton-based action recognition. ST-GCN models the human skeleton as a graph, where joints are nodes and bones are edges, and applies convolutions across both spatial and temporal dimensions to effectively learn action dynamics. This approach is computationally efficient and robust to variations in camera viewpoint and subject appearance.

A core contribution of our work is the decoupling of pose estimation from the action recognition framework. We bypass the integrated human detection and pose estimation modules within MMAction2 and instead feed it the keypoints produced from our externally trained AlphaPose models. This required the development of a custom data processing script to bridge the two frameworks, addressing incompatibilities in both data structure and coordinate systems.

To achieve this, we developed a custom script to bridge AlphaPose's output per frame (bounding boxes, keypoint coordinates $(x_i, y_i)$, and confidence scores) with the ST-GCN action recognition framework, adapting the confidence scores $(c_i)$ to also serve as the required keypoint visibility data. ST-GCN requires all the above information, plus the keypoints' visibility, which AlphaPose does not provide. Following a standard approach [11] we adopt the keypoint scores for this input, as well. As a final note, the input frames are also scaled before being processed by the ST-GCN network, such that the shorter dimension matches the target length. The same scaling is performed on the keypoints coordinates to ensure proper alignment.

## 4. Experimental

This section details the experimental protocol used to validate our proposed pipeline. We first describe the datasets and implementation details. Next, we evaluate the selection of an optimal keypoint filtering threshold. We finally provide a direct comparative analysis of the end-to-end action recognition performance in low-light conditions.

### *4.1. Datasets and Evaluation Metrics*

Our experimental setup leverages three distinct datasets. We made use of the COCO dataset [9] for training the pose estimation models. Specifically, we utilized both the train2017 part, which contains over 118,000 images with rich annotations for 17 keypoints per person and the val2017 part, which contains over 5,000 images also with annotations. For the final action recognition evaluation, we used the ARID (Action Recognition in the Dark) dataset [24], which is specifically designed for low-light scenarios and serves as our primary testing ground. The action recognition model, ST-GCN, was pre-trained on the NTU-RGB+D 60 dataset [31]. To create a consistent evaluation framework, we identified the common action classes shared between the ARID and NTU-60 datasets, viz: i) "jump", ii) "drink", iii) "pickup", iv) "sit down", and v) "stand up".

To quantify performance, we employed two primary metrics. Overall accuracy was used to measure the percentage of correctly classified actions across the entire test set. Additionally, we generated a confusion matrix to provide a more granular, class-by-class analysis, revealing specific strengths and weaknesses of each approach.

### *4.2. Training and Hyperparameter Tuning*

For our experiments, we trained two pose estimation models to directly compare the proposed low-light adaptation technique to the original model. Both models were trained using the official AlphaPose implementation with a ResNet-50 backbone on a single NVIDIA GeForce RTX 4070 GPU, using a batch size of 76. The **Baseline Model** $(M_b)$, was trained for 120 epochs on the original COCO dataset, while the **Adapted Model** $(M_a)$ was trained for 240 epochs on our combined dataset of original and synthetic low-light images. An initial learning rate of $10^{-4}$ was used with the Adam optimizer, decaying by a factor of 0.1 at epochs 90 and 120. The total training time for the $M_a$ model was approximately 91 hours, while the $M_b$ model was trained for 23 hours.

### *4.3. Optimal Keypoint Threshold Selection*

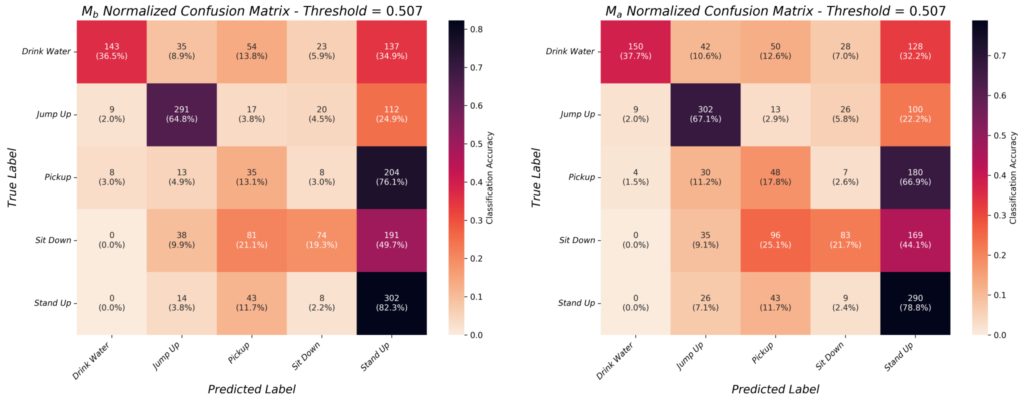
As described in Section 3.3, filtering low confidence keypoints is critical for robust action recognition. To determine the optimal filtering threshold for both the baseline and our adapted model, we performed a preliminary evaluation on a validation subset of the ARID dataset. We generated candidate thresholds using the GMM, WMM, and Otsu's method and evaluated the downstream action recognition accuracy for each. The results are summarized in Table 1, showing that for both models the Otsu's method provided the threshold that yielded the highest accuracy.

**Table 1.** Action recognition accuracy on the ARID [24] validation set for different keypoint confidence thresholding methods. The highest accuracy for each model is shown in **bold**.

| Model | Threshold Method | Threshold Value | ARID Accuracy |
|---|---|---|---|
| $M_b$ | **Otsu** | **0.507** | **42.63%** |
| | GMM | 0.514 | 42.28% |
| | WMM | 0.528 | 41.62% |
| $M_a$ | **Otsu** | **0.507** | **44.05%** |
| | GMM | 0.521 | 43.59% |
| | WMM | 0.532 | 42.63% |

### 4.4. Action Recognition Performance

For the final evaluation, we compared the end-to-end action recognition performance of the two optimized pipelines on the ARID test set. The pipeline using our $M_a$ pose estimator achieved an overall accuracy of 44.05%, an absolute improvement of 1.42% over the baseline pipeline accuracy of 42.63%.



(**a**) Baseline Model ($M_b$)

(**b**) Adapted Model ($M_a$)

**Figure 4.** Confusion matrices for action recognition on the ARID test set [24]. (a) Results from the baseline pipeline. (b) Results from our proposed adapted pipeline. The diagonal shows higher values for our method, indicating improved accuracy across most classes.

The confusion matrices in Fig. 4 provide a detailed class-by-class breakdown of performance. A detailed analysis of the confusion matrix for $M_b$ (Fig. 4(**a**)) reveals that it frequently misclassifies actions involving rapid motion. For example, the action "jump up" was misclassified as "standing up" in 32% of its instances. This is likely due to motion blur causing the baseline pose estimator to fail. Our adapted model's confusion matrix (Fig. 4(**b**)) shows a marked improvement in this area, reducing the misclassification rate for "pickup" to just to 28%. This indicates that by training on synthetic low-light data, the pose estimator learns to generate robust poses even from visually degraded inputs, leading to a more accurate and reliable action recognition system.

## 5. Conclusions

This paper presented a data-centric pipeline for robust human action recognition in low-lighting environments, designed for HRI applications. By synthetically generating a large-scale low-light dataset and training a specialized pose estimator, we successfully shifted the computational burden of low-light adaptation from inference to an offline training phase. Our experiments on the ARID dataset demonstrate that this approach improves the accuracy of action recognition compared to a baseline model trained only on well-lit data. The proposed method offers a practical and efficient solution for developing robots capable of safely and effectively navigating among humans in challenging lighting conditions. Future work will explore more advanced image-to-image translation models and the integration of temporal information directly into the pose estimation stage.

**Funding:** This research is implemented in the framework of H.F.R.I call "Basic research Financing (Horizontal support of all Sciences)" under the National Recovery and Resilience Plan "Greece 2.0" funded by the European Union – NextGenerationEU (H.F.R.I. Project Number: 15339).

**Data Availability Statement:** Any data generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Baptista, J.; Castro, A.; Gomes, M.; Amaral, P.; Santos, V.; Silva, F.; Oliveira, M. Human–Robot Collaborative Manufacturing Cell with Learning-Based Interaction Abilities. *Robotics* **2024**, *13*, 107.
2. Masala, G.L.; Giorgi, I. Artificial Intelligence and Assistive Robotics in Healthcare Services: Applications in Silver Care. *International Journal of Environmental Research and Public Health* **2025**, *22*, 781.
3. Chitikena, H.; Sanfilippo, F.; Ma, S. Robotics in Search and Rescue (SAR) Operations: An Ethical and Design Perspective Framework for Response Phase. *Applied Sciences* **2023**, *13*, 1800.
4. Abdalla, G.O.E.; Veeramanikandasamy, T. Implementation of SPY Robot for a Surveillance System Using Internet Protocol of Raspberry Pi. In Proceedings of the Proceedings of the IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, 2017.
5. Sawik, B.; Tobis, S.; Baum, E.; Suwalska, A.; Kropińska, S.; Stachnik, K.; Pérez-Bernabeu, E.; Cildoz, M.; Agustin, A.; Wieczorowska-Tobis, K. Robots for Elderly Care: Review, Multi-Criteria Optimization Model and Qualitative Case Study. *Healthcare* **2023**, *11*, 1286.
6. Cani, J.; Koletsis, P.; Foteinos, K.; Kefaloukos, I.; Argyriou, L.; Falelakis, M.; Del Pino, I.; Santamaria-Navarro, A.; Čech, M.; Severa, O.; et al. TRIFFID: Autonomous Robotic Aid For Increasing First Responders Efficiency. *arXiv preprint arXiv:2502.09379* **2025**.
7. Zaki, A.A.; Fathy, A.M.; Carnevale, M.; Giberti, H. Application of Realtime Robotics Platform to Execute Unstructured Industrial Tasks Involving Industrial Robots, Cobots, and Human Operators. *Procedia Computer Science* **2022**, *200*, 1359–1367.
8. Heimbach, M.P.; Weber, J.; Schmidt, M. Training a Robot with Limited Computing Resources to Crawl Using Reinforcement Learning. In Proceedings of the Proceedings of the IEEE International Conference on Robotic Computing, 2022.
9. Lin, T.Y.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Proceedings of the European Conference on Computer Vision, 2014.
10. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.L.; Lu, C. AlphaPose: Whole-body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 7157–7173.
11. Contributors, M. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2, 2020.
12. Kong, Y.; Fu, Y. Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision* **2022**, *130*, 1366–1401.
13. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *43*, 3349–3364.
14. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv preprint arXiv:1904.04514* **2019**.

15. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *Advances in Neural Information Processing Systems* **2022**, *35*, 38571–38584.
16. Wang, Z.; Li, P.; Liu, H.; Deng, Z.; Wang, C.; Liu, J.; Yuan, J.; Liu, M. Recognizing Actions from Robotic View for Natural Human-Robot Interaction. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025.
17. Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J. Lite-HRNet: A Lightweight High-Resolution Network. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
18. Li, Q.; Zhang, Z.; Xiao, F.; Zhang, F.; Bhanu, B. Dite-HRNet: Dynamic Lightweight High-Resolution Network for Human Pose Estimation. In Proceedings of the Proceedings of the International Joint Conference on Artificial Intelligence, 2022.
19. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
20. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
21. Luvizon, D.C.; Picard, D.; Tabia, H. Multi-task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *43*, 2752–2764.
22. 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 Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
23. Rosh, G.; Shankar, M.; Kukreja, P.; Namdev, A.; Prasad, B.P. XPose: Towards Extreme Low Light Hand Pose Estimation. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2025.
24. Xu, Y.; Yang, J.; Cao, H.; Mao, K.; Yin, J.; See, S. ARID: A New Dataset for Recognizing Action in the Dark. In Proceedings of the Proceedings of the International Workshop on Deep Learning for Human Activity Recognition, 2021.
25. Ha, M.; Bae, W.G.; Bae, G.; Lee, J.T. ELLAR: An Action Recognition Dataset for Extremely Low-Light Conditions with Dual Gamma Adaptive Modulation. In Proceedings of the Proceedings of the Asian Conference on Computer Vision, 2024.
26. Hira, S.; Das, R.; Modi, A.; Pakhomov, D. Delta Sampling R-BERT for Limited Data and Low-Light Action Recognition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
27. Parmar, G.; Park, T.; Narasimhan, S.; Zhu, J.Y. One-Step Image Translation with Text-to-Image Models. *arXiv preprint arXiv:2403.12036* **2024**.
28. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017.
29. Sauer, A.; Lorenz, D.; Blattmann, A.; Rombach, R. Adversarial Diffusion Distillation. In Proceedings of the Proceedings of the European Conference on Computer Vision, 2024.
30. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767* **2018**.
31. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.