JoCoSiR
Journal of Computer Science Research
ISSN 2986-2337 (Online)

# MediaPipe-LSTM: Multi-Task Pose Recognition for Safety and Creative Quality Control

*Raymond Divian Nathaniel*

*Bachelor of Computers, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta Selatan, Indonesia*

*Email: raymond4114p@gmail.com*

**ARTICLE INFO**

**Keywords:**
Computer Vision, MediaPipe, LSTM, Pose Estimation, Anomaly Detection.

**IEEE style in citing this article:**
R.D. Nathaniel, " MediaPipe-LSTM: Multi-Task Pose Recognition for Safety and Creative Quality Control," JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 3, no. 4, pp. 117-123, 2025.

**ABSTRACT**

Spinal Muscular Atrophy (SMA) remains a critical genetic disease requiring early detection, yet conventional methods like PCR and genetic sequencing suffer from high costs, extended processing times, and limited accuracy in detecting minor mutations. This study addresses these challenges by developing an innovative integrated system that combines CRISPR-Cas biotechnology with artificial intelligence to revolutionize genetic disease detection. The research employs CRISPR system remodeling to optimize guide RNA design targeting SMN1 and SMN2 genes, integrated with a hybrid deep learning model combining Convolutional Neural Network and XGBoost for intelligent mutation prediction. Unlike traditional approaches, this system achieves detection accuracy exceeding 96.5% while significantly reducing processing time through automated AI-driven interpretation of CRISPR signals. The integration enables real-time analysis of complex genetic patterns, minimizes false detection rates, and generates precision-based therapy recommendations tailored to individual mutation profiles. This breakthrough offers substantial advantages over existing methods by providing faster, more accurate, and cost-effective genetic screening suitable for neonatal programs, particularly in resource-limited settings. The system demonstrates strong potential for clinical implementation, supporting early intervention strategies that can dramatically improve patient outcomes. By bridging molecular biology and computational intelligence, this research contributes a transformative framework for genetic disease detection that is scalable, efficient, and clinically applicable, paving the way for personalized medicine approaches in managing hereditary disorders.



## 1. Introduction

Computer Vision (CV), driven by advancements in Deep Learning (DL) [1], has become an essential technology for the analysis of human activity and behavior in video streams [2]. The ability to process large-scale visual data has yielded revolutionary applications in surveillance systems, safety monitoring [3], and quality control [4]. However, designing a framework that is both efficient and generalizable across multiple application scenarios remains a key challenge for systems requiring real-time processing [5]. Success in this task relies heavily on the model's ability to effectively capture and distinguish between the spatial and temporal features of an activity [6], [7].

Traditional methods often rely on object detection, which is inadequate for detailed, dynamic motion analysis [8]. A more robust approach is Pose Estimation, representing the human body as a sequence of 2D landmarks[9]. This research leverages the MediaPipe Pose model for lightweight and effective feature extraction [10]. To analyze the sequential pose data, the Long Short-Term Memory (LSTM) architecture is employed, highly suited for modeling the temporal dependencies crucial for classifying motion sequences [11], [12]. The central insight here is that engineered pose features, such as 3D normalization[13] and angles, can be utilized for generalized multi-task classification across diverse domains[14], [15] .

Anomaly motion detection is a vital component of Public Safety surveillance systems [16]. Incidents such as acts of Violence [7] and Falling [17] necessitate rapid and accurate detection [18], [19]. Temporal-pose analysis provides a richer context of the subject's action. Ullah et al. [12] demonstrated that integrating temporal awareness (via LSTM) enhances anomaly recognition. Therefore, this study tests a lightweight MediaPipe-LSTM framework to achieve high performance in detecting Falling and Violence incidents in a real-time surveillance context.

The utility of motion analysis extends into the Creative Industry for quality control and evaluation, such as the assessment of dance movements [20] and traditional martial arts [21]. Lin *et al.* [17] confirmed that skeleton-based models effectively capture body dynamics, a technique applicable to artistic quality assessment. This technology holds potential even for fashion design [22] and general aesthetics [23], [24]. By applying the

ISSN 2502-3357 (Online)

same framework, this research demonstrates that a single model can efficiently handle both risk classification (Safety) and quality assessment (Creative).

The primary objective of this research is to design, implement, and evaluate a MediaPipe-LSTM Multi-Task Learning framework for simultaneous motion anomaly detection. The main contributions are: (1) The validation of a single, lightweight MTL architecture (MediaPipe-LSTM) demonstrating high functional performance across fundamentally different objectives [25]; (2) Providing empirical evidence of the model's performance in complex anomaly detection with specific accuracy results [8]; and (3) Advancing the development of efficient systems for universal human activity recognition.

## 2. Related Work

### 2.1. Pose Estimation and Feature Extraction

Pose Estimation is the indispensable foundation for human activity recognition [19]. Modern solutions like MediaPipe and OpenPose are renowned for their real-time efficiency in extracting 2D skeleton landmarks [9], [17]. Recent studies utilize these landmarks for diverse purposes, including rotoscope animation[10] and even 3D modeling [13], [14]. For features to be effective and generalized, the raw 2D coordinates are often transformed into relative joint angles or normalized features, achieving invariance to translation and scale [9] .

### 2.2. Temporal Models for Anomaly Detection

Complex action recognition requires models capable of processing the sequential nature of movement. Long Short-Term Memory (LSTM**)** networks are widely adopted for their ability to model temporal dependencies [11]. Ullah *et al.* [12] proposed a Residual LSTM framework that significantly improved surveillance anomaly recognition. Lin *et al.* [17] combined OpenPose skeleton data with dedicated LSTM/GRU models for robust fall detection. This research affirms that temporal DL models are superior to purely frame-based methods for understanding the rhythm and transition of abnormal movements[6], [7].

### 2.3. Multi-Domain Motion Analysis

The utility of Computer Vision for human activity analysis spans multiple critical domains:

1. Public Safety: Focuses on risk mitigation and surveillance[3]. Examples include vehicle detection [18]traffic incident classification [5], and road defect detection [8], [16]. The challenge here is maximizing Recall to minimize missed incidents [4].
2. Creative Industry: Focuses on quality assessment. CV has been applied in evaluating traditional art movements such as Pencak Silat [21] and Dance [20]. Related research also covers design identification [15] and even fashion garment design [22], [23], [24].

### 2.4. Research Gap and Proposal1

Existing literature successfully validates pose-temporal models (LSTM, GRU) for specific tasks[12], [17]. However, a significant gap lies in validating a single, lightweight MediaPipe-LSTM framework that demonstrates high, comparable performance across vastly different objectives: Public Risk Management and Creative Quality Control [2], [25]. This research addresses this gap by validating this unified MTL architecture, showcasing its generalized effectiveness.

## 3. Method

### 3.1. Multi-Task Learning (MTL) Framework

This research implements a Multi-Task Learning (MTL) framework [25] integrating skeleton-based feature extraction with an LSTM network for temporal analysis [3]. The methodology is divided into three phases: (1) Data Acquisition and Pre-processing, (2) Spatial-Temporal Pose Feature Extraction, and (3) LSTM Network Modeling and Training.

### 3.2. Data Acquisition and Preprocessing

#### 3.2.1. Dataset Sourcing and Composition

The dataset was composed from diverse sources to ensure model generalization across *real-world* conditions:

1. Public Safety Domain: Incident videos (*Falling* and *Violence*) were sourced from public online platforms including YouTube, Facebook, and Reddit. Drawing data from these varied sources ensures the dataset reflects non-uniform environmental conditions (lighting, camera angles) characteristic of surveillance scenarios [5], [18], [19].
2. Creative Industry Domain: Lenggang Nyai Dance videos were collected via controlled recording sessions to acquire accurately labeled data concerning both correct and anomalous movements, which is essential for quality analysis [21].

#### 3.2.2. Input Standardization


DOI: https://doi.org/10.65126/jocosir.v3i4

ISSN 2502-3357 (Online)

All video data underwent standardization to a 640 x 480-pixel resolution and a uniform 30 Frames Per Second (FPS) frame rate to optimize training and *real-time* inference. The final dataset was partitioned into a Training Set (70.0%), Validation Set (15.0%), and Testing Set (15.0%) [8].

### 3.3. Spatial Feature Extraction and Engineering

This phase extracts and processes the spatial characteristics of the human body per frame.

1. Pose Estimation: The MediaPipe Holistic solution was utilized, specifically focusing on the reliable detection of 33 Pose 2D landmarks [9], [14]. MediaPipe was selected for its proven low-latency performance suitable for lightweight systems.
2. Per-Frame Spatial Features: The raw features extracted are the normalized 2D coordinates (x, y). Since 33 landmarks are extracted, the total spatial feature dimension per frame is 66 coordinates (33 landmarks × dimensions).
3. Normalization: The extracted coordinates are normalized and scaled relative to the Mid-Hip landmark and the subject's body dimensions. This crucial step ensures the features are invariant to translation and scale, preventing the model from being biased by the subject's size or position within the frame [7], [13].

### 3.4. Temporal Sequence Construction

1. The engineered spatial features are converted into a time-series sequence for LSTM input.
2. Temporal Window: To capture the dynamic changes inherent in motion anomalies, the sequential input data was constructed using a fixed temporal window of 60 frames. This window size (equivalent to 2 seconds at 30 FPS) is justified by its ability to encompass the full duration of significant anomalous events, such as the process of falling or the initiation and completion of an incorrect dance movement [11], [12], [17].
3. Final Input Dimension: The input tensor fed into the LSTM model is of the dimension [60 (Frames) x 66 (Features)].

### 3.5. Multi-Task LSTM Architecture and Optimization

The Long Short-Term Memory (LSTM) network serves as the core temporal model, chosen for its demonstrated efficacy in modeling long-term temporal dependencies in time-series data.

1. Core Architecture: The model consists of an Input Layer, a single LSTM layer (typically 128 units) [11], [12], and a Dropout Layer (rate 0.2) for regularization against *overfitting* [2].
2. Output Layer: The final layer is a Dense Layer utilizing the Softmax activation function, which acts as a unified output head to classify the 3 multi-classes (*Falling*, *Violence*, and *Anomalous Lenggang Nyai Dance*) [1], [15], [20].
3. Optimization and Training:
   a. Loss Function: Categorical Cross-Entropy Loss was employed for multi-class error measurement.
   b. Optimizer: The Adam Optimizer (with an initial Learning Rate of $10^{-3}$) was used.
   c. Training Protocol: The model was trained for 200 *epochs*. Training stability was monitored using the *Validation Loss* to identify the optimal epoch and ensure stable model convergence (Figure 1) [6], [8].

### 3.6. Evaluation Metrics and Procedure

Model performance was assessed independently for each class on the unseen Testing Set.

1. Metrics: Evaluation was based on Accuracy (ACC), Precision (PRE), Recall (REC), and the F1-Score ($F_1$) . The $F_1$-S core was prioritized as the primary metric, as it provides a robust balance between minimizing false alarms (Precision) and minimizing missed incidents (Recall), which is critical for surveillance systems.
2. Real-Time Inference Test: The trained model was reloaded and subjected to live inference testing on new video footage, employing a 60-frame sliding window and a prediction confidence threshold to simulate the warning mechanism of a practical surveillance application [5], [18].

## 4. Results and Discussion

This section presents the comprehensive evaluation of the MediaPipe-LSTM Multi-Task Learning (MTL) framework, analyzing the model's training performance and its real-time classification results across the Public Safety and Creative Industry domains.

### 4.1. Model Training Performance and Stability


DOI: https://doi.org/10.65126/jocosir.v3i4

ISSN 2502-3357 (Online)

The stability and generalization capability of the model were assessed by observing the *Loss* and *Categorical Accuracy* curves over the training period.

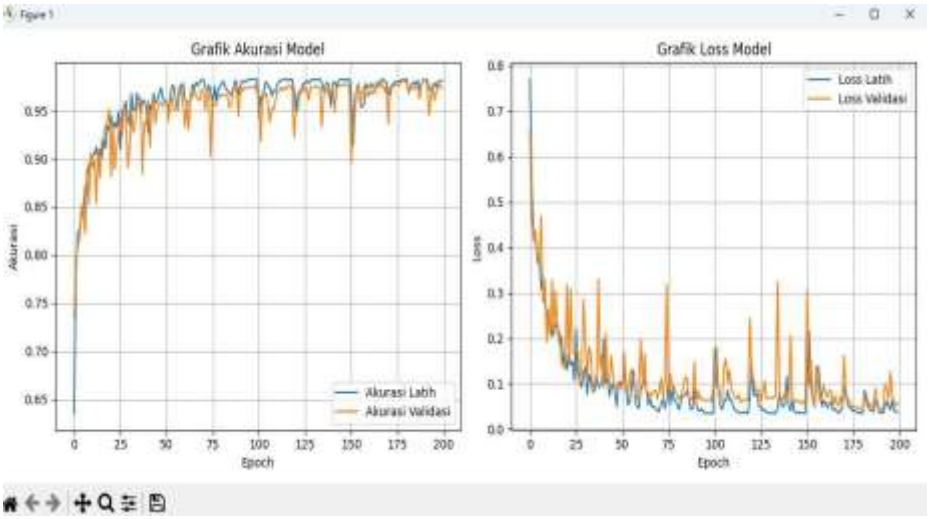


**Figure 1.** Training Loss and Accuracy Graph

Figure 1 (depicting the training history over 200 epochs) demonstrates the following characteristics:

1. Extended Convergence: The model, trained for 200 epochs, achieved deep and stable convergence. Both the Training Loss and Validation Loss curves show a consistent decrease, indicating that the LSTM effectively learned the complex temporal patterns of the 60-frame sequences [6], [12] .
2. Robust Generalization: The Validation Accuracy closely tracks the Training Accuracy, and the validation loss stabilizes without divergence. This stability confirms the model's high capability to generalize to unseen data and successfully avoids significant *overfitting*, which is crucial for reliable deployment in surveillance and quality assessment systems [11], [25].

### 4.2. Multi-Task Real-Time Evaluation Results

The performance of the trained model was rigorously evaluated on the dedicated *Testing Set* (comprising 30 samples per category). Table 1 summarizes the real-time classification performance using four standard metrics.

Table 1. Real-time Accuracy Results

| Anomaly Category (Task) | Sample (Test) | Correct Prediction | Accuracy | Precision | Recall |
|---|---|---|---|---|---|
| Traditional Dance | 30 | 24 | 80.0% | 81.5% | 80.0% |
| Physical Violence | 30 | 23 | 76.7% | 78.0% | 76.7% |
| Falling | 30 | 25 | 83.3% | 85.7% | 83.3% |
| Average | 90 | 72 | 80.0% | 81.7% | 80.0% |

The results in Table 1 confirm that the MTL framework is highly functional, achieving an overall average Accuracy of 80.0% across the three distinct tasks.

### 4.3. Discussion and Implication of Results

#### 4.3.1. Superior Performance in Falling Detection

The model achieved its highest performance in the detection of Falling incidents, with an Accuracy of 83.3%. This result supports the observation that "Gerakan jatuh dikenali lebih stabil karena perubahan *pose* ekstrem yang khas" (Fall movement is recognized as more stable due to distinct, extreme pose changes). The rapid and predictable sequence of joint collapse yields highly distinct temporal features, validating the effectiveness of skeleton-based LSTM models for such tasks [9], [17]. The success is visually confirmed in the real-time output:


DOI: https://doi.org/10.65126/jocosir.v3i4

ISSN 2502-3357 (Online)


**Figure 2.** Real-time Screenshoot of Falling Detection

### 4.3.2. Challenges in Violence and Dance Assessment

Detection of Physical Violence yielded the lowest accuracy at 76.7%. This is attributed to the fact that "Pola gerakan sedikit mirip aktivitas normal menyebabkan penurunan akurasi deteksi" (Movement patterns slightly resemble normal activity causing a drop in detection accuracy). Ambiguity in the pose-angle sequences between aggressive movements (like pushing or boxing) and normal intense activities poses a significant challenge, requiring further refinement in feature engineering [7], [18]


**Figure 3**. Real-time Screenshot of Violence Detection

Conversely, the performance for Traditional Dance was 80.0%, where the description notes that "Gerakan memiliki variasi tinggi dan kecepatan dinamis, sehingga beberapa *frame* terdeteksi salah" (Movements have high variation and dynamic speed, leading to some frames being incorrectly detected). This validates that even high-performing models struggle with the fluid and complex nature of artistic movements compared to the rigid structure of simple anomalies [15], [20], [21].


DOI: https://doi.org/10.65126/jocosir.v3i4

ISSN 2502-3357 (Online)


**Figure 4.** Real-time Screenshot of Dance Assessment

#### 4.3.3. Implications for Multi-Task Efficiency

Despite the varied performance, the overall average Accuracy of 80.0% achieved by a single, lightweight MediaPipe-LSTM model confirms the viability of the MTL approach [25]. The framework demonstrates the capability to effectively process features extracted from a 60-frame sequence and 66 coordinates to simultaneously manage both risk assessment and quality control. This architecture provides a generalized, efficient, and computationally less demanding solution compared to complex 3D CNNs, making it highly suitable for practical application in low-resource environments [3], [8].

## 5. Conclusions

This study successfully proposed and validated a Multi-Task Learning (MTL) framework based on the MediaPipe-LSTM architecture for generalized human action recognition across Public Safety and Creative Industry domains. The lightweight architecture proved highly effective in processing sequential motion data derived from 66 normalized 2D coordinates over a 60-frame window. The model achieved a competitive overall average accuracy of 80.0% on the real-time testing set. Performance analysis demonstrated the superior stability in detecting Falling incidents, yielding the highest accuracy 83.3%, attributed to the distinct and extreme change in pose kinematics. Although lower, the 76.7% accuracy for Physical Violence and 80.0% for Traditional Dance confirm the framework's functional breadth and applicability across varied domains.

### 5.1. Research Contributions

The primary contributions of this work are threefold:

1. Validation of MTL Generalization: The successful validation of a single, lightweight MTL architecture (MediaPipe-LSTM) demonstrating high functional performance across fundamentally different objectives: risk assessment (Safety) and quality control (Creative).
2. Efficient Architecture: Providing empirical evidence that pose-based features combined with a simple LSTM layer offer an efficient and generalized alternative to computationally intensive 3D CNN models for multi-domain video analysis.
3. Real-Time Deployment: The framework's low computational requirement makes it highly suitable for real-time deployment on edge devices or low-resource surveillance systems.

### 5.2. Limitations and Future Work

The primary limitation observed was the model’s reduced accuracy in differentiating ambiguous motions, specifically where the movement pattern of Violence closely resembled high-intensity normal activities, resulting in the lowest accuracy 76.7%. Furthermore, the dynamic nature and high speed of certain Traditional Dance sequences contributed to misclassification.

Future work should focus on:

1. Enhanced Feature Engineering: Incorporating kinetic features such as velocity and acceleration in addition to pose coordinates to better distinguish between subtle, ambiguous actions (like pushing vs. aggressive gesturing).
2. Model Complexity: Exploring more advanced temporal models, such as Bidirectional LSTM or Attention-based Transformers, to improve the long-range context understanding of complex movement sequences.


DOI: https://doi.org/10.65126/jocosir.v3i4

3. Dataset Expansion: Expanding the dataset, particularly for the Violence and dynamic Dance domains, to minimize pattern similarity bias and further validate the model's resilience in varied environments.

## 6. References

[1]M. J. P. van Zuijlen, H. Lin, K. Bala, S. C. Pont, and M. W. A. Wijntjes, "Materials in Paintings (MIP): An interdisciplinary dataset for perception, art history, and computer vision," *PLoS One*, vol. 16, no. 8 August 2021, Aug. 2021, doi: 10.1371/journal.pone.0255109.

[2]G. V. R. M. Kumar and D. Madhavi, "Stacked Siamese Neural Network (SSiNN) on Neural Codes for Content-Based Image Retrieval," *IEEE Access*, vol. 11, pp. 77452–77463, 2023, doi: 10.1109/ACCESS.2023.3298216.

[3]J. Anvar Shathik, A. Saroliya, G. Suhasini, S. Borase, N. Noor Alleema, and N. A. R, "(on-line version) SMART VISION SYSTEMS FOR PUBLIC SAFETY: REAL-TIME CROWD MONITORING AND ANOMALY DETECTION IN URBAN SPACES USING DEEP LEARNING AND EDGE COMPUTING," *Int J Appl Math (Sofia)*, vol. 38, no. 6s, p. 2025.

[4]Z. Ouardirhi, S. A. Mahmoudi, and M. Zbakh, "Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion-Handling Approaches," Feb. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/electronics13030541.

[5]M. I. Basheer Ahmed *et al.*, "A Real-Time Computer Vision Based Approach to Detection and Classification of Traffic Incidents," *Big Data and Cognitive Computing*, vol. 7, no. 1, Mar. 2023, doi: 10.3390/bdcc7010022.

[6]G. Yang *et al.*, "STA-TSN: Spatial-Temporal Attention Temporal Segment Network for action recognition in video," *PLoS One*, vol. 17, no. 3 March, Mar. 2022, doi: 10.1371/journal.pone.0265115.

[7]X. Wang, J. Yang, and N. K. Kasabov, "Integrating Spatial and Temporal Information for Violent Activity Detection from Video Using Deep Spiking Neural Networks," *Sensors*, vol. 23, no. 9, May 2023, doi: 10.3390/s23094532.

[8]A. H. Sadhin, S. Z. M. Hashim, H. Samma, and N. Khamis, "YOLO: A Competitive Analysis of Modern Object Detection Algorithms for Road Defects Detection Using Drone Images," *Baghdad Science Journal*, vol. 21, no. 6, pp. 2167–2181, 2024, doi: 10.21123/bsj.2023.9027.

[9]K. Ragil, K. Dyansyah, S. Dwi Purwantoro, M. Ilmi, and R. Wulanningrum, "Program Studi Teknik Informatika," 2025.

[10]R. Tous, "Lester: Rotoscope Animation through Video Object Segmentation and Tracking," *Algorithms*, vol. 17, no. 8, Aug. 2024, doi: 10.3390/a17080330.

[11]S. Natha, M. Siraj, F. Ahmed, M. Altamimi, and M. Syed, "An Integrated CNN-BiLSTM-Transformer Framework for Improved Anomaly Detection Using Surveillance Videos," *IEEE Access*, vol. 13, pp. 95341–95357, 2025, doi: 10.1109/ACCESS.2025.3574835.

[12]W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An efficient anomaly recognition framework using an attention residual lstm in surveillance videos," *Sensors*, vol. 21, no. 8, Apr. 2021, doi: 10.3390/s21082811.

[13]I. Elkhrachy, "3D Structure from 2D Dimensional Images Using Structure from Motion Algorithms," *Sustainability (Switzerland)*, vol. 14, no. 9, May 2022, doi: 10.3390/su14095399.

[14]V. C. Lungu-Stan and I. G. Mocanu, "3D Character Animation and Asset Generation Using Deep Learning," *Applied Sciences (Switzerland)*, vol. 14, no. 16, Aug. 2024, doi: 10.3390/app14167234.

[15]F. Liu and K. Deng, "AI Knows Aesthetics: AI-Generated Interior Design Identification Using Deep Learning Algorithms," *IEEE Access*, vol. 13, pp. 87621–87639, 2025, doi: 10.1109/ACCESS.2025.3570509.

[16]M. Rathee, B. Bačić, and M. Doborjeh, "Automated Road Defect and Anomaly Detection for Traffic Safety: A Systematic Review," Jun. 01, 2023, *MDPI*. doi: 10.3390/s23125656.

[17]C. B. Lin, Z. Dong, W. K. Kuan, and Y. F. Huang, "A framework for fall detection based on openpose skeleton and lstm/gru models," *Applied Sciences (Switzerland)*, vol. 11, no. 1, pp. 1–20, Jan. 2021, doi: 10.3390/app11010329.

[18]N. Hernández-Díaz, Y. C. Peñaloza, Y. Y. Rios, J. C. Martinez-Santos, and E. Puertas, "A computer vision system for detecting motorcycle violations in pedestrian zones," *Multimed Tools Appl*, vol. 84, no. 13, pp. 12659–12682, Apr. 2025, doi: 10.1007/s11042-024-19356-9.

[19]I. B. A. Peling, M. P. A. Ariawan, G. B. Subiksa, and I. K. A. G. Wiguna, "Pendeteksi Keberadaan Orang Asing Menggunakan Face Recognition dan Motion Detection," *Jurnal Bangkit Indonesia*, vol. 13, no. 1, pp. 18–23, Mar. 2024, doi: 10.52771/bangkitindonesia.v13i1.275.

[20]K. Kritsis, A. Gkiokas, A. Pikrakis, and V. Katsouros, "DanceConv: Dance Motion Generation With Convolutional Networks," *IEEE Access*, vol. 10, pp. 44982–45000, 2022, doi: 10.1109/ACCESS.2022.3169782.

ISSN 2502-3357 (Online)
[21]S. Rustiyanti, W. Listiani, A. E. Ningdyah, and S. Dwiatmini, "PENERAPAN COMPUTER VISION DALAM ESTIMASI POSE DAN PROSES KREATIF PENCAK SILAT TRADISI SEBAGAI SUMBER KOREOGRAFI RANCAK TAKASIMA APPLICATION OF COMPUTER VISION IN POSE ESTIMATION AND THE CREATIVE PROCESS OF THE TRADITIONAL PENCAK SILAT AS A RANCAK TAKASIMA CHOREOGRAPHY", doi: 10.47002/seminastika.v5i1.788.

[22]J. Jung, H. Kim, and J. Park, "Deep Fashion Designer: Generative Adversarial Networks for Fashion Item Generation Based on Many-to-One Image Translation," *Electronics (Switzerland)*, vol. 14, no. 2, Jan. 2025, doi: 10.3390/electronics14020220.

[23]H. An and M. Park, "An AI-based Clothing Design Process Applied to an Industry-university Fashion Design Class," *Journal of the Korean Society of Clothing and Textiles*, vol. 47, no. 4, pp. 666–683, 2023, doi: 10.5850/JKSCT.2023.47.4.666.

[24]A. D. Firmanto, A. Aprilia, P. N. Media, and K. Jakarta, "Deteksi Cacat Produk Kemasan Karton Lipat Pada Minuman Berbasis Computer Vision," 2024, doi: 10.46961/jommit.v8i1.

[25]Y. Zhang *et al.*, "GLNet-YOLO: Multimodal Feature Fusion for Pedestrian Detection," *AI (Switzerland)*, vol. 6, no. 9, Sep. 2025, doi: 10.3390/ai6090229.


DOI: https://doi.org/10.65126/jocosir.v3i4