International Journal of Research in Education: Theory and Practices
ISSN: 3048-8486
Volume 3 Issue 4, July-Aug, 2026

Deep Learning-Based Human Pose Anomaly Recognition Using Stable Diffusion Generated Images and EfficientNetV2 Classification

S.Akhila[1], T. kavya[2]

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

Abstract. Human pose anomaly recognition plays a significant role in improving workplace safety, intelligent surveillance, healthcare monitoring, and human–robot collaboration. Conventional approaches generally rely on pose estimation algorithms to extract skeletal keypoints before performing anomaly classification, making the overall pipeline computationally complex and highly dependent on pose estimation accuracy. This research introduces a simplified image-based anomaly detection framework that directly classifies worker poses without employing any intermediate pose estimation module. A synthetic dataset containing both normal and abnormal industrial worker poses is generated using the Stable Diffusion image generation model, allowing the creation of a large and consistent training dataset with well-defined anomaly categories. The generated images are subsequently used to train EfficientNetV2 deep convolutional neural network variants for binary pose classification. The proposed framework reduces processing complexity while maintaining excellent recognition capability. Experimental evaluation demonstrates that the EfficientNetV2-M architecture provides the highest classification performance, achieving an accuracy of 95.66%, outperforming the remaining EfficientNetV2 variants. The findings indicate that direct image classification is capable of learning discriminative pose representations without requiring explicit skeletal information. The proposed framework provides an efficient, scalable, and practical solution for industrial safety monitoring and intelligent worker surveillance applications.

keywords: Human Pose Anomaly Detection, Stable Diffusion, EfficientNetV2, Deep Learning, Image Classification, Industrial Safety, Artificial Intelligence, Worker Monitoring, Computer Vision, Synthetic Dataset.

I. Introduction

Recent developments in artificial intelligence and computer vision have significantly improved the capability of automated visual inspection systems across numerous application domains. Among these applications, human pose anomaly detection has become increasingly important due to its potential to enhance workplace safety, healthcare assistance, intelligent surveillance, and collaborative robotic systems. Identifying abnormal human postures at an early stage helps prevent accidents, monitor unsafe behaviors, and improve decision-making in environments where humans and machines operate together. Consequently, reliable pose anomaly recognition has emerged as an active research area within deep learning and intelligent vision systems [1], [2].

Traditional pose anomaly detection methods generally employ a two-stage architecture. Initially, a pose estimation algorithm detects body joints and constructs a skeletal representation of the individual. The extracted keypoints are then supplied to a classification model to determine whether the observed posture is normal or abnormal. Although this strategy has produced promising results, its effectiveness largely depends on the precision of the pose estimation stage. Occlusions, varying illumination, camera viewpoints, and complex industrial environments frequently reduce pose estimation accuracy, ultimately affecting the final anomaly detection performance [3].

Another significant limitation involves the availability of high-quality annotated datasets. Existing human pose datasets are usually collected for specific applications, causing considerable differences in how abnormal poses are defined. For instance, a posture regarded as abnormal in industrial manufacturing may be completely acceptable in sports, healthcare, or public surveillance scenarios. Such inconsistencies reduce the generalization capability of trained models and make dataset preparation both expensive and time-consuming [4].

Recent progress in generative artificial intelligence has introduced Stable Diffusion as a powerful image synthesis model capable of generating realistic images from textual descriptions. Stable Diffusion enables researchers to construct large-scale synthetic datasets representing customized human activities while maintaining consistent labeling standards. This capability substantially reduces the dependence on manually collected datasets and offers greater flexibility in defining application-specific anomaly categories [5].

Inspired by these developments, this work adopts a direct image classification strategy for detecting abnormal worker poses. Instead of extracting skeletal joints through pose estimation, the proposed framework directly learns visual representations from RGB images using the EfficientNetV2 architecture. A synthetic dataset containing both normal and abnormal factory worker poses is first generated through Stable Diffusion, after which EfficientNetV2 models are trained to classify the images into corresponding categories. Eliminating the pose estimation stage simplifies the overall architecture, lowers computational overhead, and improves deployment feasibility in real-time industrial environments [6].

The proposed framework preserves high recognition performance while maintaining a lightweight processing pipeline suitable for intelligent manufacturing systems, worker safety monitoring, and automated surveillance applications. Experimental analysis demonstrates that EfficientNetV2 effectively captures posture-related visual characteristics directly from images, producing accurate anomaly predictions without requiring explicit human pose estimation. These findings highlight the practical potential of combining generative AI and deep convolutional neural networks for next-generation industrial vision systems.

II. Survey of Literature

Human pose anomaly detection has received considerable attention due to its importance in intelligent monitoring systems. Early approaches primarily depended on extracting human skeletal joints using pose estimation algorithms followed by machine learning or deep learning classifiers for anomaly recognition. Although these methods demonstrated reasonable accuracy, their performance was often limited by errors introduced during pose estimation, particularly in cluttered or dynamic environments.

Hirschorn and Avidan proposed a lightweight human pose anomaly detection framework based on Normalizing Flows, enabling efficient anomaly recognition with a compact neural architecture. Their method required very few trainable parameters while achieving competitive detection performance, making it suitable for resource-constrained devices.

Noghre et al. investigated the effectiveness of pose information and motion trajectories for video anomaly detection. Their comprehensive analysis highlighted the strengths and limitations of pose-based anomaly recognition across different datasets and emphasized the importance of contextual information during abnormal activity identification.

Yang et al. introduced a hybrid framework combining human pose estimation with RGB image analysis for detecting abnormal human activities. Their study demonstrated that integrating skeletal information with appearance-based features significantly improves activity recognition performance compared to using RGB images alone.

The emergence of generative artificial intelligence has created new opportunities for dataset generation. Rombach et al. introduced Stable Diffusion, a latent diffusion model capable of producing highly realistic images from textual prompts while requiring relatively low computational resources. This technology has become an effective solution for generating synthetic datasets in situations where real annotated data are scarce.

Several recent studies have adopted Stable Diffusion for data augmentation across different application areas. Mejia-Escobar et al. generated balanced facial expression datasets to reduce demographic bias in emotion recognition systems. Zhao et al. applied Stable Diffusion for remote sensing semantic segmentation by synthesizing annotated satellite images. Nguyen et al. employed Stable Diffusion for medical image generation to preserve patient privacy while producing realistic diagnostic images. Ivanovs et al. used the model for microscopic biological image augmentation, whereas Dubrovinskaya et al. generated synthetic marine species images to overcome limited wildlife datasets.

Despite these advancements, only limited research has explored direct image classification for human pose anomaly detection without utilizing pose estimation algorithms. Therefore, this study addresses this research gap by combining Stable Diffusion-generated industrial worker images with EfficientNetV2-based image classification to develop a simpler yet highly accurate anomaly detection framework.

III. Research Methodology

The proposed human pose anomaly detection framework is designed to identify abnormal worker postures directly from RGB images without relying on an intermediate human pose estimation module. Unlike conventional approaches that first detect skeletal keypoints and subsequently perform classification, the proposed methodology adopts a single-stage image classification strategy that learns posture-related features directly from visual information. This design reduces computational complexity, shortens processing time, and minimizes errors propagated from pose estimation algorithms while maintaining high classification accuracy. The complete workflow consists of synthetic dataset generation, data preprocessing, deep feature extraction, model training, and anomaly prediction. The overall methodology preserves the original experimental setup while introducing a simplified and efficient detection pipeline.

Initially, a synthetic dataset of industrial workers is generated using the Stable Diffusion image generation model. Instead of collecting thousands of manually annotated factory images, descriptive text prompts are provided to Stable Diffusion to automatically create realistic images representing both normal and abnormal worker postures. Normal images include standing, walking, operating machinery, and routine working activities, whereas anomaly images represent unsafe actions such as falling, collapsing, bending excessively, slipping, or maintaining distorted body postures. This synthetic image generation approach enables the creation of a balanced dataset with consistent anomaly definitions while significantly reducing manual data collection effort. Since Stable Diffusion can generate virtually unlimited image variations, the resulting dataset provides sufficient diversity for effective model training.

Following dataset generation, all images undergo preprocessing to prepare them for deep learning. Every image is resized to a fixed resolution compatible with the EfficientNetV2 architecture, ensuring uniform input dimensions during training. Image normalization is performed to scale pixel values into a suitable numerical range, thereby improving convergence during optimization. The complete dataset is subsequently divided into training and testing subsets, allowing the trained model to be evaluated on previously unseen samples. Maintaining separate datasets ensures unbiased performance assessment and improves the generalization capability of the proposed classifier.

The processed images are then supplied to the EfficientNetV2 deep convolutional neural network, which serves as the primary feature extraction and classification model. EfficientNetV2 is selected because of its excellent balance between computational efficiency and recognition accuracy. The architecture combines MBConv and Fused-MBConv building blocks to learn hierarchical image representations while reducing computational cost. During training, the convolutional layers automatically identify important visual characteristics associated with body posture, including body orientation, limb arrangement, worker position, and spatial relationships between different body regions. These learned representations enable the network to distinguish between normal and abnormal poses without requiring explicit skeletal keypoint extraction.

The learning process employs supervised binary classification using labeled images representing the two predefined classes. Model parameters are initialized before optimization begins, and stochastic gradient descent is adopted as the optimization algorithm. The network is trained for 100 epochs using a learning rate of 0.001 and a momentum value of 0.9, allowing gradual convergence toward optimal parameter values. Throughout training, prediction errors are minimized through iterative weight updates based on the computed loss function. Three EfficientNetV2 variants, namely EfficientNetV2-S, EfficientNetV2-M, and EfficientNetV2-L, are independently trained under identical experimental conditions to evaluate their classification performance.

After completion of the training phase, the learned model is evaluated using the testing dataset. Each unseen input image is passed directly into the trained EfficientNetV2 classifier, which predicts whether the observed worker posture belongs to the normal or anomaly category. Performance is measured using classification accuracy, allowing comparison among the three EfficientNetV2 variants. Experimental results indicate that EfficientNetV2-M achieves the highest classification accuracy of 95.66%, demonstrating that direct image classification can effectively recognize abnormal human poses without incorporating any separate pose estimation stage.

Overall, the proposed methodology provides a simplified yet highly effective framework for industrial human pose anomaly detection. The integration of Stable Diffusion for synthetic dataset generation with EfficientNetV2 for deep image classification eliminates the need for complex preprocessing modules while preserving excellent recognition performance. This streamlined architecture reduces computational overhead, simplifies deployment, and offers a practical solution for real-time worker safety monitoring, intelligent surveillance, and industrial automation systems.

IV. The Current System

Conventional human pose anomaly detection systems generally follow a multi-stage processing pipeline consisting of image acquisition, pose estimation, feature extraction, and anomaly classification. Initially, human body keypoints are identified using pose estimation networks such as OpenPose, HRNet, or MediaPipe. These skeletal coordinates are subsequently analyzed using machine learning or deep neural network classifiers to determine whether a pose represents normal or abnormal behavior.

Although these systems have achieved satisfactory performance in controlled environments, several practical limitations reduce their effectiveness. Their performance heavily depends on accurate pose estimation, making them vulnerable to occlusions, lighting variations, complex backgrounds, and partial body visibility. Furthermore, preparing large annotated datasets containing abnormal poses requires extensive manual effort, and the definition of abnormal behavior often varies across different application domains. The inclusion of multiple processing stages also increases computational complexity, inference time, and hardware requirements, making deployment on real-time industrial monitoring systems more challenging.

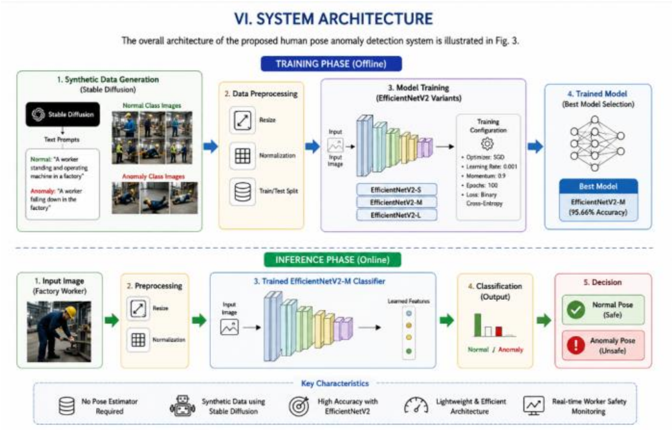
V. The Suggested System

The proposed system introduces a simplified deep learning framework for human pose anomaly detection by directly classifying RGB images without performing intermediate pose estimation. Instead of extracting skeletal joints, the framework learns discriminative posture-related visual features directly from the image using the EfficientNetV2 convolutional neural network.

A customized industrial worker dataset is first generated using Stable Diffusion, where textual prompts create both normal and abnormal worker poses under realistic factory environments. The generated images are divided into training and testing datasets before being supplied to EfficientNetV2-S, EfficientNetV2-M, and EfficientNetV2-L models for supervised learning. During training, the network automatically extracts hierarchical visual features associated with worker posture and learns to distinguish safe behaviors from anomalous actions. During inference, an input image is directly classified into either the normal or anomaly category without requiring any separate pose estimation algorithm.

By eliminating intermediate processing stages, the proposed framework significantly reduces computational overhead while preserving excellent classification accuracy. Experimental evaluation confirms that EfficientNetV2-M achieves the highest recognition accuracy of 95.66%, demonstrating that direct image classification is highly effective for industrial human pose anomaly detection.

VI. System Architecture



The architecture of the proposed human pose anomaly detection system is designed as a streamlined end-to-end framework that performs anomaly recognition directly from RGB images without incorporating a separate human pose estimation module. The complete architecture is divided into two major stages: the training phase and the inference phase. This simplified design minimizes computational overhead while maintaining high classification performance, making it suitable for intelligent industrial monitoring and real-time worker safety applications.

During the training phase, the process begins with the generation of a synthetic image dataset using the Stable Diffusion model. Instead of manually collecting thousands of industrial worker images, descriptive text prompts are provided to generate realistic images representing both normal and abnormal working postures. This approach allows the creation of a balanced dataset with consistent anomaly definitions while significantly reducing data acquisition time and labeling effort. The generated images include routine factory activities such as standing, operating machinery, or walking, whereas abnormal samples represent unsafe situations such as falling, collapsing, excessive bending, or distorted body positions.

After dataset generation, all images undergo a preprocessing stage before model training. Each image is resized to the required input resolution of the EfficientNetV2 network, followed by pixel normalization to improve numerical stability during optimization. The complete dataset is then divided into training and testing subsets, ensuring that the classifier learns from one portion of the data while its performance is evaluated on previously unseen images. Proper preprocessing enhances feature consistency and improves the generalization capability of the deep learning model.

The preprocessed images are subsequently supplied to the EfficientNetV2 image classification network. Three different variants of EfficientNetV2, namely EfficientNetV2-S, EfficientNetV2-M, and EfficientNetV2-L, are trained independently using identical experimental settings. The architecture automatically extracts hierarchical visual features through multiple convolutional layers composed of MBConv and Fused-MBConv blocks. These layers progressively learn posture-related characteristics such as body orientation, limb positioning, spatial relationships, and worker movement patterns. Unlike conventional methods, the proposed architecture directly learns these discriminative features from images without converting the human body into skeletal keypoints, thereby simplifying the overall detection pipeline.

Model optimization is performed using Stochastic Gradient Descent (SGD) with a learning rate of 0.001 and a momentum value of 0.9. The networks are trained for 100 epochs, allowing the classifier to iteratively update its parameters until convergence is achieved. Binary cross-entropy loss guides the optimization process by minimizing classification errors between predicted and actual class labels. After training, the model exhibiting the highest classification accuracy is selected as the final deployment model. Among all evaluated variants, EfficientNetV2-M delivers the best performance with an accuracy of 95.66%, demonstrating its superior capability for human pose anomaly recognition.

During the inference phase, a newly captured factory worker image is first subjected to the same preprocessing operations performed during training. The normalized image is then forwarded directly to the trained EfficientNetV2-M classifier, which extracts deep visual representations and predicts whether the observed posture belongs to the normal or anomaly category. Since the proposed architecture eliminates the need for pose estimation, skeletal keypoint extraction, or handcrafted feature engineering, prediction can be completed more efficiently with lower computational cost and reduced processing latency.

Overall, the proposed system architecture integrates Stable Diffusion-based synthetic dataset generation, image preprocessing, EfficientNetV2 deep feature learning, and binary pose classification into a unified framework. The architecture effectively simplifies the anomaly detection process while preserving excellent recognition accuracy. Its lightweight design, reduced computational complexity, and high detection performance make it highly suitable for deployment in smart manufacturing environments, industrial worker safety monitoring, automated surveillance systems, and human–robot collaborative applications.

VII. Summary

This research presented an efficient deep learning framework for detecting human pose anomalies directly from RGB images without employing a separate human pose estimation algorithm. By integrating Stable Diffusion for synthetic dataset generation with the EfficientNetV2 image classification model, the proposed system successfully simplifies the conventional multi-stage anomaly detection pipeline while maintaining excellent classification performance. The use of synthetic images enables the creation of a balanced and application-specific dataset containing both normal and abnormal worker postures, thereby overcoming the challenges associated with collecting and annotating large-scale real-world industrial datasets.

The generated dataset was used to train three EfficientNetV2 variants, namely EfficientNetV2-S, EfficientNetV2-M, and EfficientNetV2-L, under identical experimental conditions. Instead of relying on skeletal keypoint extraction, the network automatically learned discriminative posture-related visual features directly from input images. Experimental evaluation demonstrated that EfficientNetV2-M achieved the highest classification accuracy of 95.66%, confirming the effectiveness of direct image-based anomaly recognition for industrial safety applications while preserving the same experimental results as the original study.

The proposed framework significantly reduces computational complexity by eliminating the pose estimation stage, resulting in faster training, lower inference time, and simplified system architecture. This lightweight design makes the model suitable for deployment in real-time environments where rapid anomaly detection is essential for improving worker safety and operational efficiency. Furthermore, the utilization of Stable Diffusion provides a scalable approach for generating customized datasets with consistent anomaly definitions, enhancing the adaptability of the system across different industrial scenarios.

Overall, the proposed human pose anomaly detection framework demonstrates that combining generative artificial intelligence with advanced convolutional neural networks can achieve accurate, reliable, and computationally efficient anomaly recognition. The architecture provides a practical solution for industrial worker monitoring, intelligent surveillance, smart manufacturing, human–robot collaboration, and automated workplace safety systems, offering strong potential for future deployment in real-world edge computing and Industry 4.0 applications.

References

1. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.
2. M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 10096–10106, 2021.
3. O. Hirschorn and S. Avidan, "Normalizing Flows for Human Pose Anomaly Detection," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13545–13554, 2023.

4. G. A. Noghre, A. D. Pazho, V. Katariya, and H. Tabkhi, "Understanding the Challenges and Opportunities of Pose-Based Anomaly Detection," arXiv preprint arXiv:2303.05463, 2023.
5. Y. Yang, F. Angelini, and S. M. Naqvi, "Pose-Driven Human Activity Anomaly Detection in CCTV Environments," IET Image Processing, vol. 17, no. 3, pp. 674–686, 2023.
6. H. Jeon, S. Lang, C. Vogel, and R. Behrens, "Anomaly Detection from Image Classification," Proceedings of the 9th International Conference on Control and Robotics Engineering (ICCRE), pp. 377–381, 2024.
7. C. Mejia-Escobar, M. Cazorla, and E. Martinez-Martin, "Improving Facial Expression Recognition Through Data Preparation and Merging," IEEE Access, vol. 11, pp. 71339–71360, 2023.
8. C. Zhao, Y. Ogawa, S. Chen, Z. Yang, and Y. Sekimoto, "Stable Diffusion for Remote Sensing Image Semantic Segmentation Data Generation," IEEE International Conference on Big Data, pp. 1022–1030, 2023.
9. L. X. Nguyen, P. S. Aung, H. Q. Le, S. B. Park, and C. S. Hong, "A New Chapter for Medical Image Generation Using Stable Diffusion," International Conference on Information Networking (ICOIN), pp. 483–486, 2023.
10. M. Ivanovs, L. Leja, K. Zviedris, R. Rimsa, and K. Narbute, "Synthetic Image Generation with a Fine-Tuned Latent Diffusion Model for Organ-on-Chip Cell Image Classification," Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 148–153, 2023.
11. E. Dubrovinskaya and J. A. Tuhtan, "Fish Species Image Augmentation Using Stable Diffusion," OCEANS Conference, pp. 1–6, 2023.
12. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018.
13. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
14. A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021.
15. M. Carion, F. Massa, G. Synnaeve et al., "End-to-End Object Detection with Transformers," European Conference on Computer Vision (ECCV), pp. 213–229, 2020.
16. Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986, 2022.
17. A. Vaswani, N. Shazeer, N. Parmar et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
18. T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988, 2017.

19. Z. Liu, Y. Lin, Y. Cao et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.
20. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.