International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 9 (2025) 41–54

Paper Type: Original Article

Visual Analysis of Human Pose Estimation under Frame Degradation Using MediaPipe and ViTPose

Nada E. Elshami [1,2,*], Ahmad Salah [3], Amr Abdellatif [1] and Heba Mohsen [2]

[1] Faculty of Computers and Informatics, Zagazig University, Zagazig, 44519, Egypt;
Emails: nada.emad020@fci.zu.edu.eg; amaemam@fci.zu.edu.eg

[2] Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt, New Cairo, Egypt; Emails: nada.hussien@fue.edu.eg; hmohsen@fue.edu.eg.

[3] College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Sultanate of Oman; ahmad.salah@utas.edu.om

Received: 18 Jul 2025 **Revised:** 31 Aug 2025 **Accepted:** 28 Nov 2025 **Published:** 11 Dec 2025

Abstract

Human Pose Estimation (HPE) is a significant task in most computer vision applications. However, in the presence of visually degraded inputs, such as low-resolution or rotated video frames, its accuracy tends to reduce. This paper compared two frequently applied pose estimation models including MediaPipe (MP) and ViTPose in terms of their performance on carefully chosen frames extracted from three of our daily videos. In order to emulate non-optimal conditions, we used three kinds of visual filters on the videos, that is, loosy video compression (approximately 70% of the original size), clockwise 90-degree rotation, and 180-degree rotation. Then we used the original frames and compared them with their filtered counterparts using visual overlays of the predicted landmarks. Our results assist in shedding some light on the model reaction to such changes, as they provide a visual representation that could be used to explain anomalies in performance regarding different circumstances. These observations have been pivotal in determining the weakness of HPE systems in unpredictable environments and future opportunities to enhance pose estimation models with a view of their wider and real-life applications.

Keywords: Human Pose Estimation; MediaPipe; ViTPose; Rotation; Low Resolution; Model Performance.

1 | Introduction

Human Pose Estimation (HPE) contributes as one of the core elements of video analytics to effectively describe and analyze human body motion by providing insight through applications that include healthcare, sports analysis, and surveillance, among others. While the majority of HPE models are created and benchmarked against datasets that contain high-resolution and well-aligned imagery, real-world video recordings often include visual artifacts, such as rotations, lower resolutions, blur, or noise—factors that can degrade model performance.

This research provides a study of the reaction of two popular HPE models to these visual distortions. In contrast to using only quantitative measures of accuracy, we focus on a visual method of analysis by overlaying the predicted poses on both original and altered video frames, allowing for a comparative examination of model robustness under degraded visual conditions.

✉ Corresponding Author: nada.emad020@fci.zu.edu.eg

2 | Related Work

The majority of studies on HPE aim at evaluating accuracy, speed, and generalization on specially designed datasets. MediaPipe (MP), for example, is more optimized for real-time inference, whereas ViTPose leverages transformer-based architectures to achieve higher prediction accuracy. Though these are notable strengths, there has been little emphasis on evaluating the visual performance of these models under challenging conditions—such as rotated or compressed images or videos as inputs as discussed in recent robustness studies [1-6]. This paper seeks to help fill that gap by offering a visual comparison of pose predictions using overlaid landmarks, enabling a side-by-side analysis of model behavior in such non-ideal scenarios.

3 | Methodology

3.1 | Video Dataset and Preprocessing

In order to analyze the results of HPE models in various real-life environments, we compiled a small series of publicly downloadable videos. The environmental differentiation used in selecting the material explored human crowding, movement and low-light situations. Also, these complex scenarios were analyzed both before and after implementing particular visual filters, considered video compression, 90-degree rotation and 180-degree rotation, to determine the model robustness once again. Three videos have been selected based only on the visual evaluation of how it performs when it comes to pose prediction with these different conditions.

- Video 1: A person standing in a natural outdoor scene [7].
- Video 2: A man walking along a roadside at night [8].
- Video 3: A blurred crowd of people walking in an urban setting [9].

The duration of each chosen video is short ranging from 5 to 11 seconds. The evaluation is unaffected by the overall length of the video because the analysis is done frame by frame. To maintain uniformity and remove any potential selection bias during comparison, frames were extracted uniformly every 0.5 seconds.

These videos serve as the baseline (ground truth) for comparison. We applied two types of degradations: compression and rotation, to each video in order to assess the robustness of two HPE models. MP and ViTPose.

3.1.1 | Video Transformation Procedures

- Rotation: We applied two rotational transformations: 90° clockwise and 180° to simulate changes in camera orientation.

  Each video was rotated by 90° and 180° using the Clideo online video editor [10] before undertaking the extraction of frames. The rotated frames were directly fed to the HPE models to test its performance in those angles. This was done by comparing its output in terms of predictions against the annotated keypoints based on MediaPipe and ViTPose definitions. This was done without the need for modification in coordinate points or angles because the comparison is based on visual inspection.

- Video Compression: Each video was compressed to approximately 70% of its original file size using the Clideo online video editor [10]. The compression percentage is calculated using Eq. (1).

$$\begin{aligned} &\mathit{Compression\ Percentage} \\ &\qquad = \frac{\mathit{Original\ File\ Size} - \mathit{Compressed\ File\ Size}}{\mathit{Original\ File\ Size}}\ X\ 100 \\ &\qquad = \left(1 - \frac{\mathit{Compressed\ File\ Size}}{\mathit{Original\ File\ Size}}\right) X\ 100 \end{aligned} \tag{1}$$

- Frame Extraction: To enable both visual and coordinate-based comparison between models, each video was split into frames at a fixed interval of 0.5 seconds. The pseudocode for the frame extraction process is described in Algorithm **Table 1**.

Table 1. Algorithm 1- Extract Frames from Video.

Algorithm 1: Extract Frames from Video Every 0.5 Seconds

```
Load video from predefined path
function getFrame(time in seconds)
    Set video pointer to time in seconds × 1000 milliseconds
    Attempt to read a frame at the specified timestamp
    if frame is successfully read then
        Save frame as an image using a sequential filename
    end if
    return success status
end function
Initialize sec ← 0, frameRate ← 0.5, count ← 1
success ← getFrame(sec)
while success is true do
    Increment count
    sec ← sec + frameRate
    Round sec to two decimal places
    success ← getFrame(sec)
end while
```

For clarity, this procedure for frame extraction was intended to yield a fair representation of video frames to be compared in terms of the performance of an HPE model, rather than achieving an exact timing match. It is thus adequate to have a 0.5-second window for a uniform selection of video frames, considering that in this study, because of a limited length of all analyzed videos and a constant frame rate, differences resulting from FPS fluctuations or dropped and rounded video frames remain negligible. Following this preprocessing stage, we conducted visual assessments of the landmark predictions output by each HPE model for each filter.

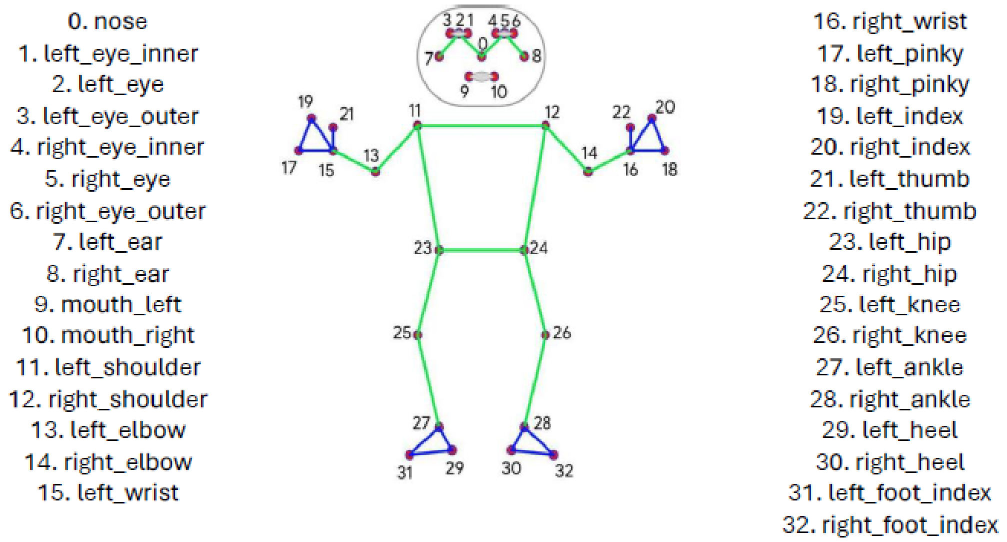


Figure 1. MediaPipe Keypoints.

3.2| Pose Estimation Models

We used two commonly used models for landmark prediction:

3.2.1| MediaPipe

MediaPipe is an abstraction framework, created by Google, that was used to make perception pipelines cross-platform and real-time. It combines lightweight and fast machine learning models and techniques based on computer vision to provide fast and lightweight solution to problems like pose estimation, facial feature landmark detection and hand tracking. MediaPipe uses BlazePose model, a two-stage detector and tracker, to detect 33 landmark points on the human body as shown in Figure 1. MediaPipe Keypoints, and the omission is rapid and precise. MediaPipe is designed to run on mobile and embedded systems, thus it works especially well for real-time applications because it has low-latency processing, resource-light computation, and its modularity. It is simple and durable regarding controlled environments and has won it a lot of fans in spheres of fitness monitoring, healthcare and augmented reality, and surveillance [11-16].

3.2.2| ViTPose

Vision Transformer-based model with dense landmark predictions. ViTPose is a state-of-the-art human pose estimation model using plain vision transformers (ViT) to make a highly accurate guess at 17 anatomical keypoints human by human as shown in Figure 2, e.g. joints and limbs. ViTPose is simpler, non-hierarchical, has less structural complexity than previous models and depends neither on CNNs nor on a complex transformer architecture: it uses a simple transformer backbone with a lightweight decoder, where errors in the simple backbone are corrected in the lightweight decoder. It achieves extraordinary performance over numerous benchmarks, such as 80.9 AP on the MS COCO test-dev set and performs well in a variety of situations, such as occlusion, disparate resolutions and training across dataset. It has the advantages of scale (100M through 1B parameters), training and deployment flexibility, and effective knowledge transfer between model sizes, being a very generalizable solution to a wide variety of applications in healthcare, surveillance, sports, AR/VR, and others [17-20].

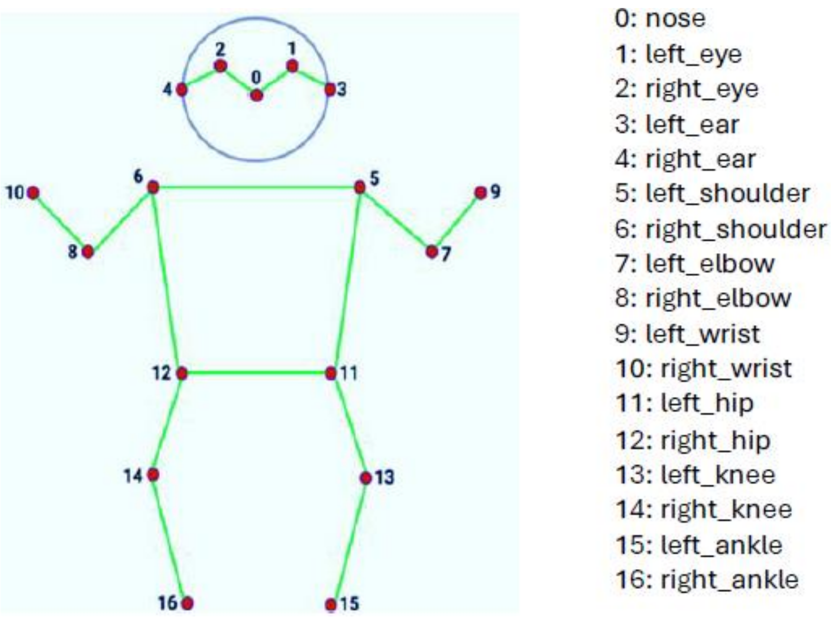
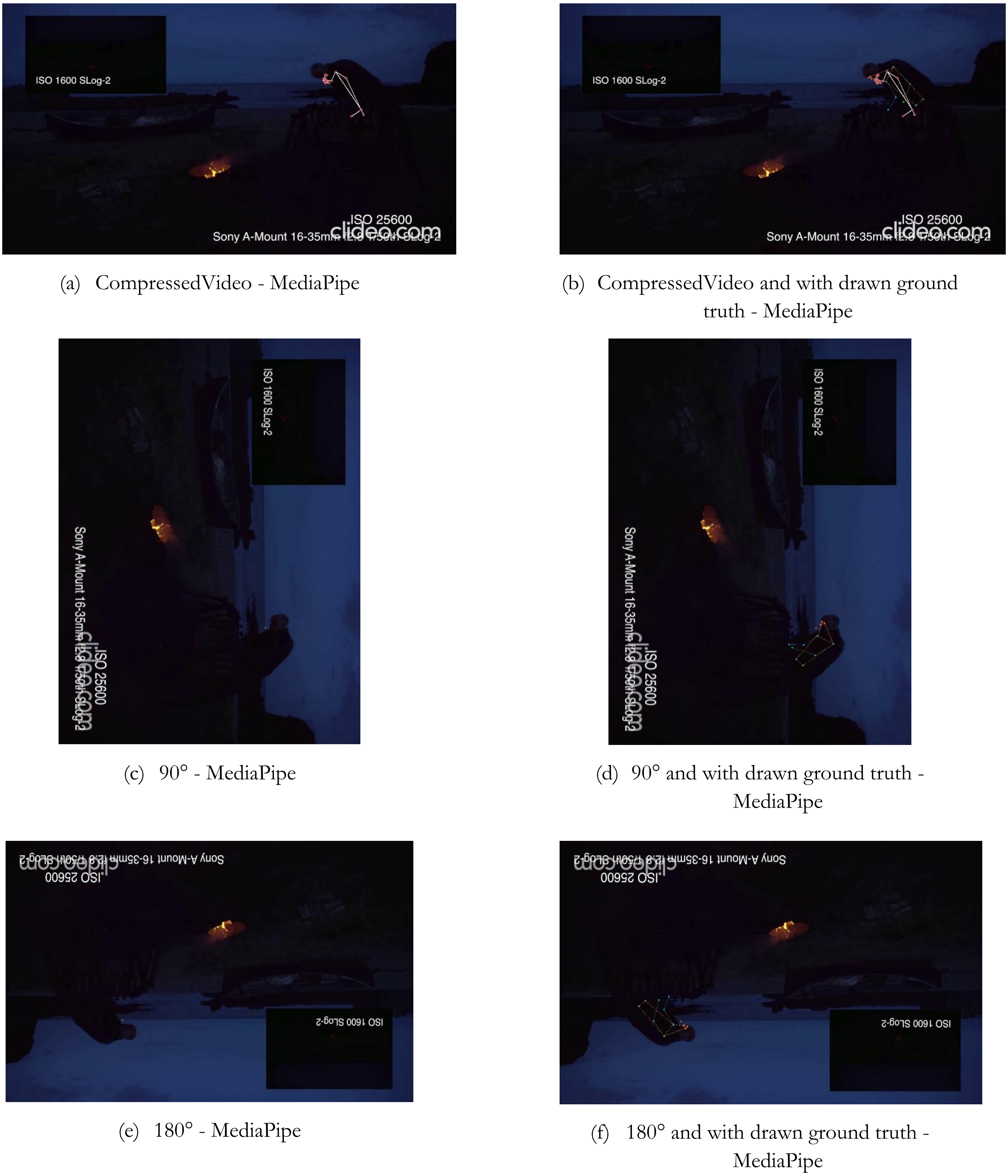


Figure 2. ViTPose Keypoints.

4 | Results and Discussion

4.1 | Videos Visual Results

4.1.1 | Video 1: A Person Standing in a Natural Outdoor Scene



(a) CompressedVideo - MediaPipe

(b) CompressedVideo and with drawn ground truth - MediaPipe

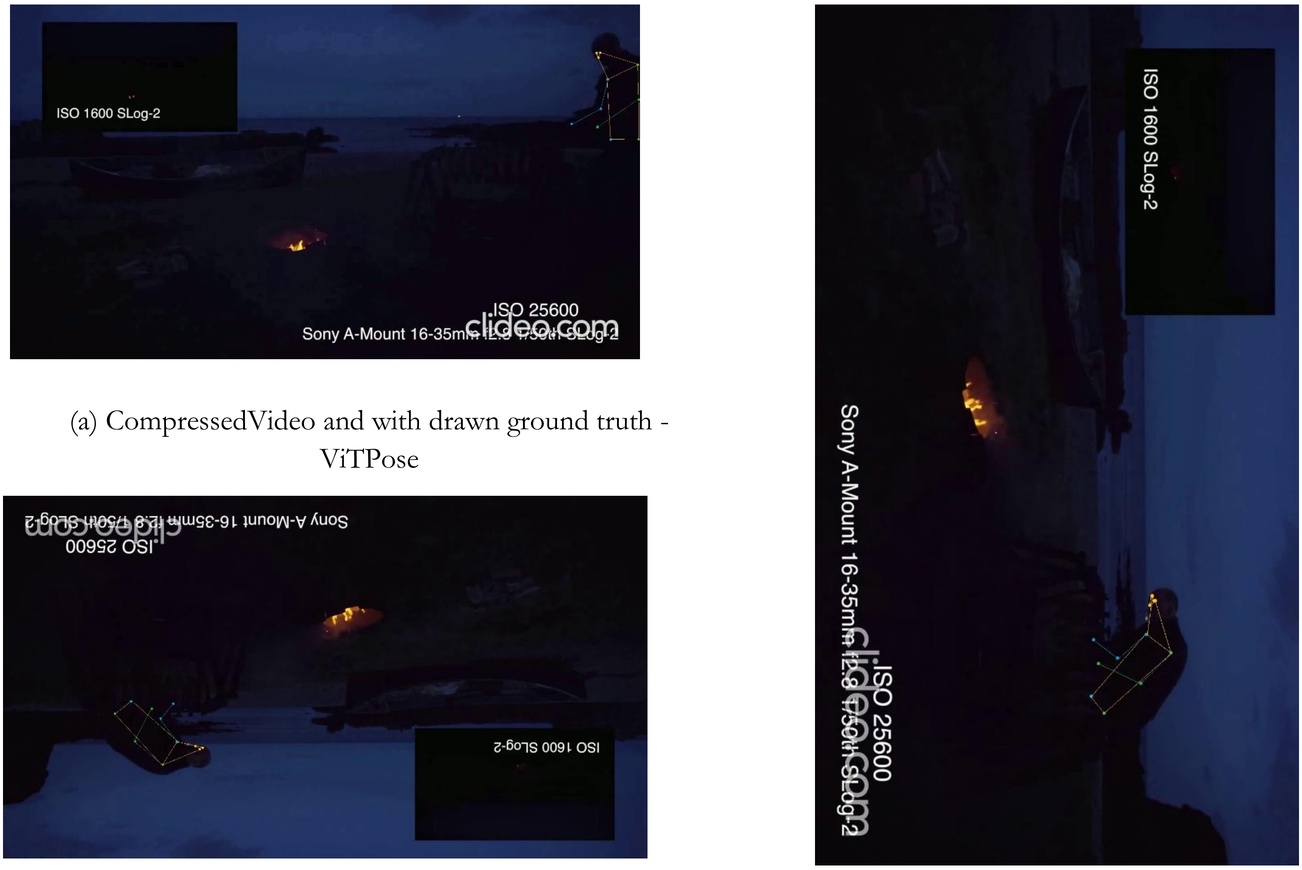
(c) 90° - MediaPipe

(d) 90° and with drawn ground truth - MediaPipe

(e) 180° - MediaPipe

(f) 180° and with drawn ground truth - MediaPipe

Figure 3. MediaPipe landmark outputs on Video 1.



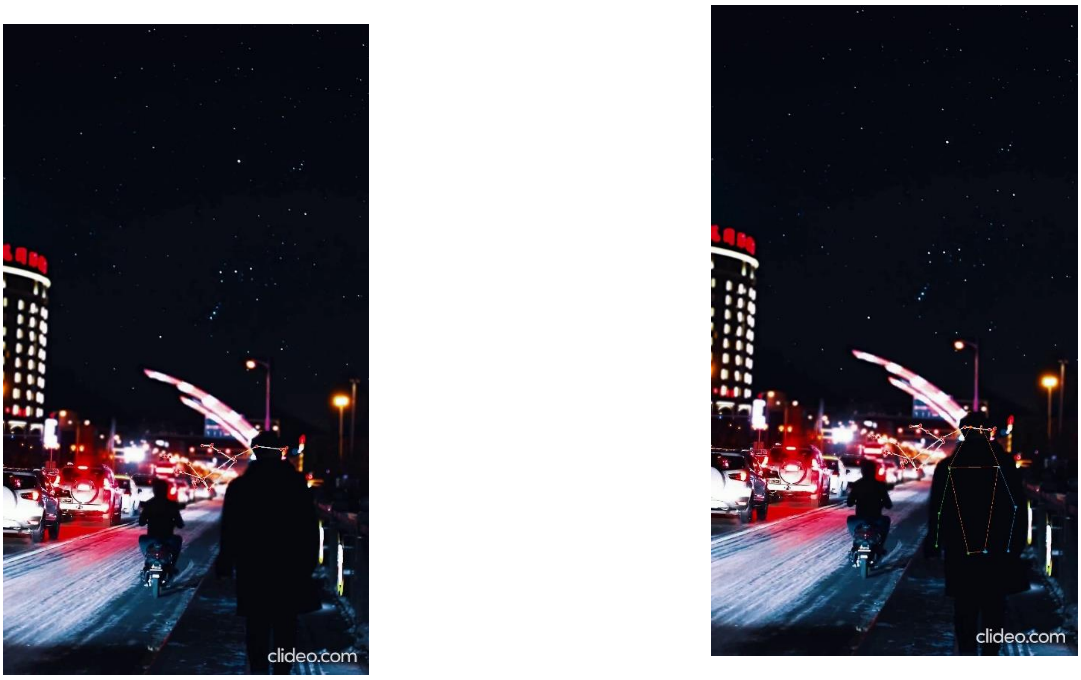
(a) CompressedVideo and with drawn ground truth - ViTPose

(c) 180° and with drawn ground truth - ViTPose

(b) 90° and with drawn ground truth - ViTPose

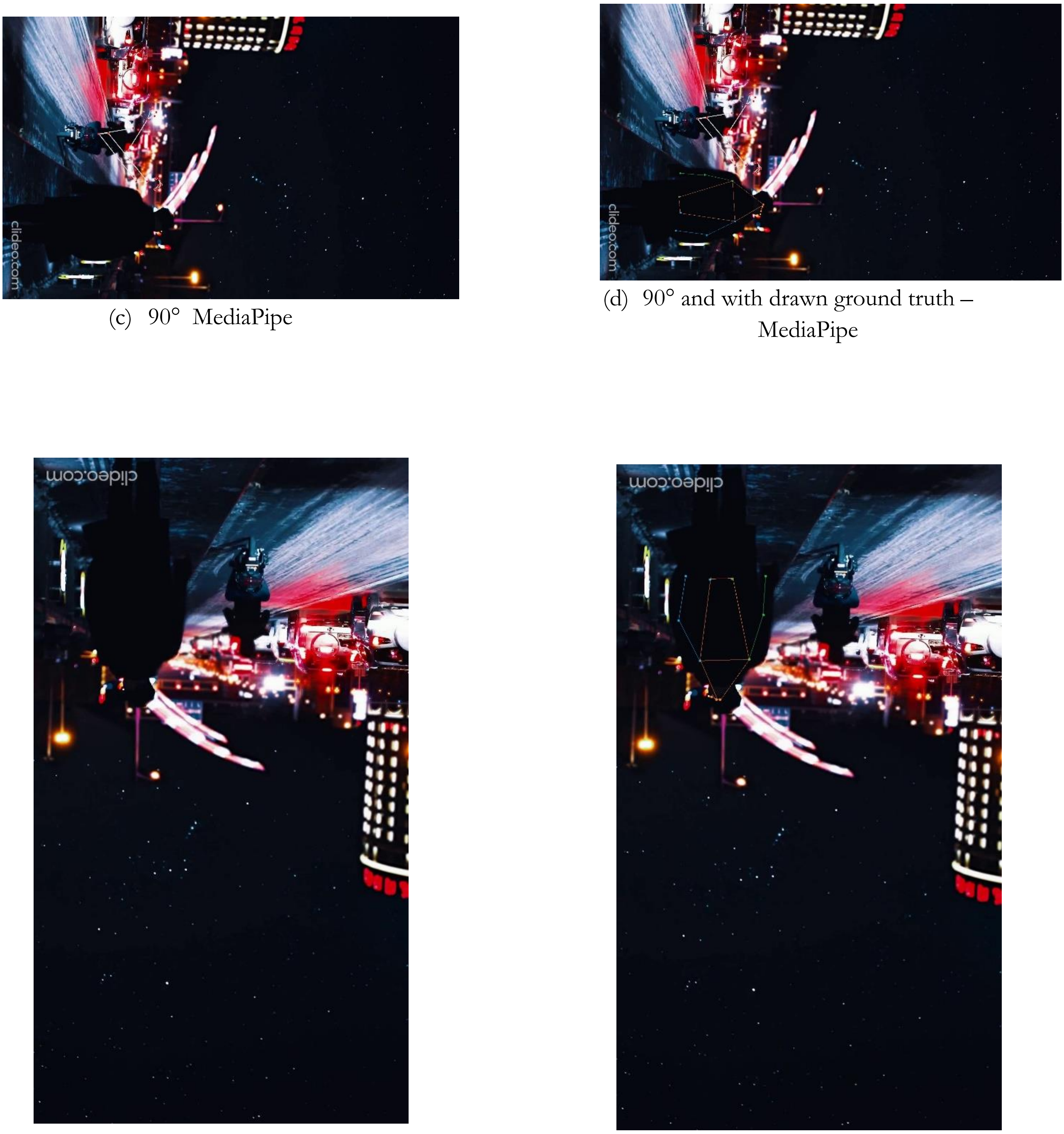
Figure 4. ViTPose landmark outputs on Video 1.

4.1.2 | Video 2: A man walking along a roadside at night.



(a) CompressedVideo - MediaPipe

(b) CompressedVideo and with drawn ground truth - MediaPipe



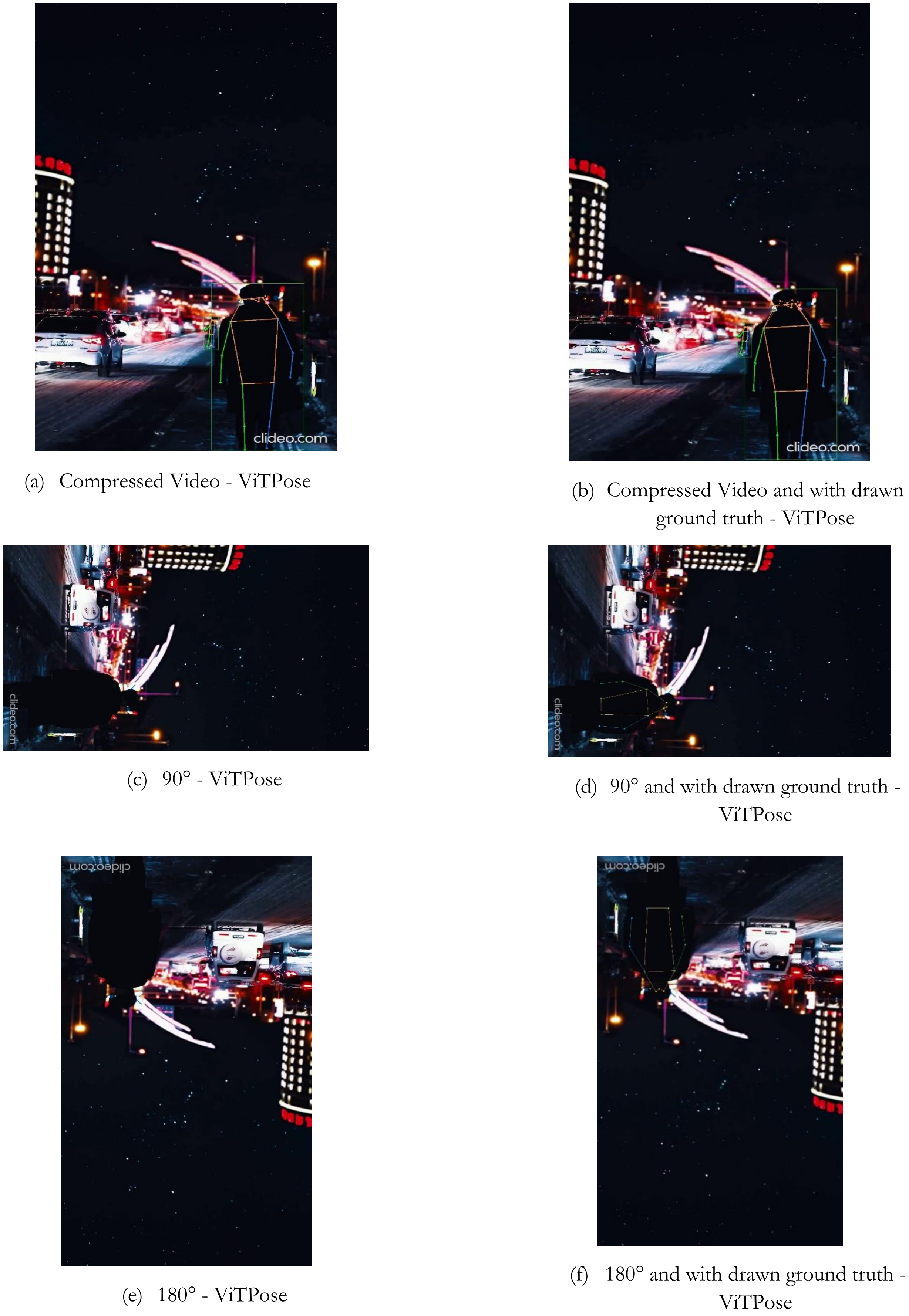
(c) 90° MediaPipe

(d) 90° and with drawn ground truth – MediaPipe

(e) 180° - MediaPipe

(f) 180° and with drawn ground truth - MediaPipe

Figure 5. MediaPipe landmark outputs on Video 2.



(a) Compressed Video - ViTPose

(b) Compressed Video and with drawn ground truth - ViTPose

(c) 90° - ViTPose

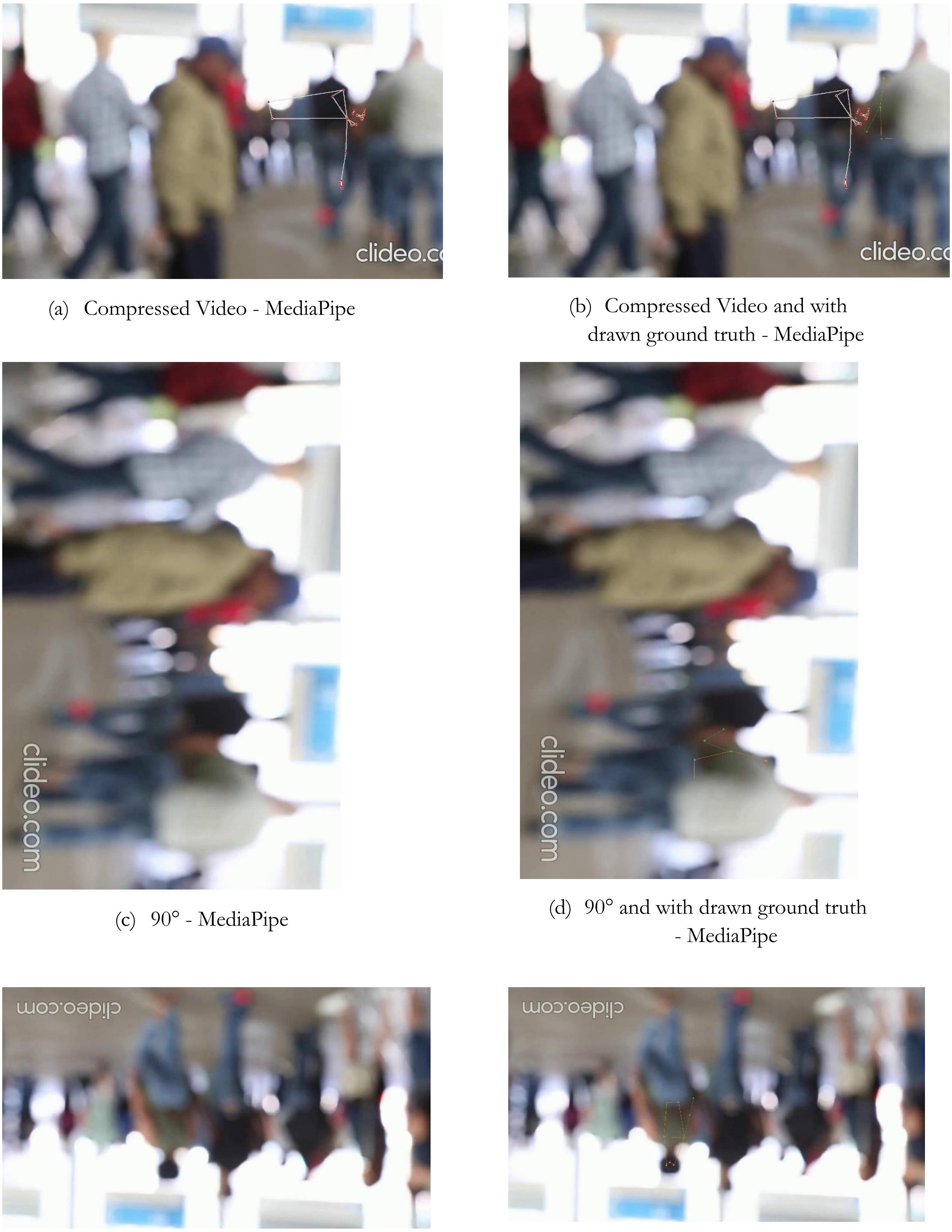
(d) 90° and with drawn ground truth - ViTPose

(e) 180° - ViTPose

(f) 180° and with drawn ground truth - ViTPose

Figure 6. ViTPose landmark outputs on Video 2.

4.1.3 | Video 3: A blurred crowd of people walking in an urban setting.



(a) Compressed Video - MediaPipe

(b) Compressed Video and with drawn ground truth - MediaPipe

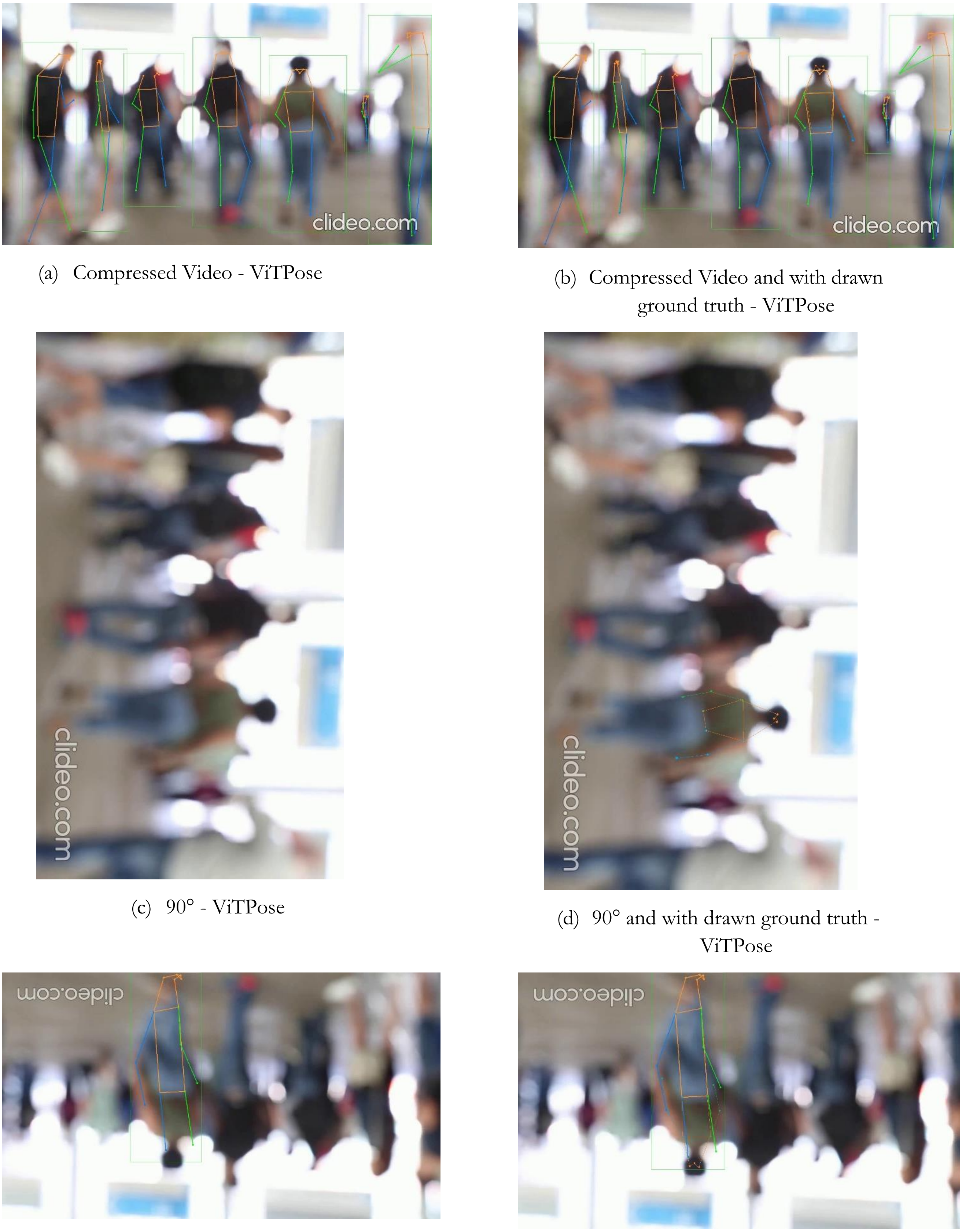
(c) 90° - MediaPipe

(d) 90° and with drawn ground truth - MediaPipe

(e) 180° - MediaPipe

(f) 180° and with drawn ground truth - MediaPipe

Figure 7. MediaPipe landmark outputs on Video 3.



(a) Compressed Video - ViTPose

(b) Compressed Video and with drawn ground truth - ViTPose

(c) 90° - ViTPose

(d) 90° and with drawn ground truth - ViTPose

(e) 180° - ViTPose

(f) 180° and with drawn ground truth - ViTPose

Figure 8. ViTPose landmark outputs on Video 3.

4.2 | Discussion

As can be seen in Video 1 Figure 3, the MediaPipe (MP) model shows profound deviations when compared to the manually created ground truth; the ground truth was manually created in Microsoft Paint by labeling every joint, pixel by pixel, to as closely resemble the corresponding MediaPipe and ViTPose keypoints as possible as outlined in Figure 1 and Figure 2. It can be seen that in video compression environments Figure 3 (a) and Figure 3 (b), clear deviations in the predicted poses from the ground truth can be noticed. In both rotation cases (90 and 180 degrees), the model was able to detect a limited number of landmarks but with highly inaccurate placement, which highlights its inability to detect human poses under circumstances defined by low visibility and darkness. On the other hand, the ViTPose model fails to detect any poses in all filtered cases displayed in this video. As a result, Figure 4 only shows the manually created ground-truth poses, as no output was generated using the model.

In Video 2, the MP system creates pose estimations that are imprecise and do not match any real individual in the frame. The imprecision is most notable under video compression and 90-degree rotation, as shown in Figure 5 (a) and Figure 5 (c), compared to the reference data shown in Figure 5 (b) and Figure 5 (d). Additionally, the MP system is challenged when processing multiple individuals, often trying to predict a single pose or misattributing landmarks. In contrast, the ViTPose model performs almost excellently under video compression, producing highly accurate estimations with minimal deviations in head landmarks, as shown in Figure 6 (a) and Figure 6 (b). However, it fails drastically when dealing with frames rotated by 90 or 180 degrees, as presented in Figure 6 (c), (d), (e), and (f), thus pointing out the vulnerability of the model to changes in orientation. However, under rotated conditions (90° and 180°), ViTPose fails to detect any meaningful poses, whereas MP occasionally produces partial or distorted poses. Thus, MP shows a limited detection capability even when poses are inaccurate, while ViTPose fails to respond at all under such transformations.

Video 3 is a unique case because its indistinct background adds a level of noise and ambiguity to the scenario. The MP system can detect some landmarks in frames of the compressed video as shown in Figure 7 (a) and Figure 7(b); however, these landmarks are not matched to any real individuals in the frame. In the 90-degree and 180-degree rotation frames, the MP system is unable to identify any landmarks whatsoever, as shown in Figure 7 (c), (d), (e), and (f). Additionally, it still fails in multi-person detection, generating no meaningful outputs in the crowded scenes. Ground truth in these cases has been manually labeled to give an idea of what ought to be detected in the original frame. For comparison, ViTPose does well in the compressed video scenario, producing close to accurate pose estimations of most people in the frame, ignoring the crowd and blur challenges, as shown in Figure 8 (a) and Figure 8 (b). It detects no landmarks in the 90-degree rotation frame but one full-body pose in the 180-degree rotation frame. This pose is, however, shown in the inverse, upside-down pose and is not as accurate as the true scene. It also fails to detect other people in the frame, as shown in Figure 8 (c), (d), (e), and (f).

Briefly, ViTPose is more robust in cases involving blurriness, making it a preferable choice in such cases compared to MP, even with its poorer performance in cases involving rotational distortions. Improvement in robustness regarding orientation would further increase its effectiveness. Under illumination conditions, as evidenced by examples given in Video 1, MP shows relatively better performance compared to ViTPose, despite inaccuracies in their approximations. With further developments, MP can potentially be a better choice in cases where ViTPose fails in landmark identification altogether. In cases involving rotation, MP demonstrates limited yet comparatively better detection capability than ViTPose, as it occasionally identifies human poses even if they are misaligned or inaccurately drawn. ViTPose, on the other hand, generally fails to produce any detections under rotated conditions. Both models have some limitations that highlight the need for pose estimation improvements under deteriorated real-world conditions.

It is noteworthy that compression degradation is a natural fusion of several degradation types, including resolution degradation, block degradations, and blur artifacts. This makes it quite tricky to identify a single precise reason for its performance degradation in a particular experiment or study. In this work, all degradation types are considered as one category to assess how well the models of HPE generally perform in such conditions. In our future research, we will aim to study how multiple degradation types work together simultaneously in order to have a clear understanding of their individual impacts.

In the current research, the performance of pose estimation was mostly evaluated through visual inspection of differences between the ground truth labeled by human expertise and estimates of keypoints provided by individual HPE models. This qualitative assessment enables one to visualize differences in structure and pose estimates in degradation conditions. Nonetheless, a more precise quantitative assessment in terms of differences in x and y-coordinates ($\Delta x, \Delta y$) of ground truth and estimates of keypoints, divided by model ground truth distances, would offer a fair assessment of individual model performance. This quantitative assessment framework was also described at a conceptual level within the current study but will be developed in forthcoming research.

4.3| Models Settings

4.3.1 | MediaPipe Pose

MediaPipe Pose, a human pose estimation framework developed by Google Research, was used for analyzing human poses in a frame-by-frame manner. In fact, a full version of Google's BlazePose was utilized through MediaPipe's Python API (v0.10.9) in a static image analysis mode that prevents tracking of human poses from one frame to another.

Every parameter was set to its defaults (mode=False, smooth=True, detectionCon=0.5, trackCon=0.5) to ensure that all experiments were reproduced in an identical manner. This was done in a Python implementation that utilized OpenCV for grabbing and displaying video frames as well as drawing functionality provided by MediaPipe's drawing utilities (mp.solutions.drawing_utils) for depicting keypoints and skeleton lines [21].

This was run in a Google Colab workspace with its in-built hardware infrastructure; using Python 3.10 with a CPU runtime to provide a controlled environment. It was also specifically done in the state with the image mode being static, and it guarantees that each frame is handled individually without the assistance of either tracking or image filtering feature. This permits none of the processing of one frame influencing the rest in regards to motion tracking and is thus simply done based on image degradation capabilities like rotation or compression.

4.3.2 | ViTPose – B

Transformer ViTPose-B (Base) was utilized as a model of Vision Transformer architectures in human pose estimation. In this research, inference was done through use of a publicly accessible interface, known as Hugging Face Space, that incorporates YOLOX-l for person localization and ViTPose-B pre-trained in coco for key-point estimation. In order to provide a suitable comparison, this model was run with its set defaults. The most important variables were as follows: box score threshold=0.5 units, visualization threshold=0.3 units, dot radius=4 units, line width=2 units, and a total of 60 processed frames. The reason for choosing to use the hosted inference environment was for ensuring robust methodology replication, as it leverages official pretrained models and pipelined detection/estimation routines in line with the open-source ViTPose library [22], [23].

In MediaPipe Pose, only the most confident detected person was considered in a single frame as per its single-person detection functionality. ViTPose, coupled with the YOLOX-l detector, allowed single as well as multiple persons to be analyzed simultaneously. Nevertheless, as it was pertinent to conduct a study to analyze the performance of pose estimation in degradation conditions in general, a comparison between the models was done with respect to keypoints according to the most confident detection in a single frame.

5 | Conclusion and Future Work

This work conducted visual analysis of the strength of two frequently used models within the area of Human Pose Estimation (HPE), i.e. MediaPipe and ViTPose, in low visibility and visual impairment situations, e.g. when the object is transformed in rotation and/or video compression. Taking a qualitative action to determine our evaluation, we examined the strength of each model regardless of quantitative evaluation. The visual analysis identified that the reasoning between the models was quite different in terms of the modified input responses, although it created a clear picture as to the inherent weaknesses of each methodology and its potential advantages when applied to difficult environments.

Although subjective reviews are valuable information, it would be more accurate to include quantitative scales in further studies of the assessments of model performance. In particular, degradation of a frame could be evaluated more accurately by calculating the variations in the landmarks coordinates using values of the (X, Y) of the original and degraded frames. Also, it would be a good improvement to extend the research to cover a broader spectrum of degradation types, e.g., blurring, noise, and occlusion, to be able to better reflect the situation in the actual circumstances.

Additional enhancements may involve utilization of image degradation recovery mechanisms prior to the implementation of pose estimation, which would in turn lessen the infliction of degradation. Lastly, investigation of fine-tuning of models with augmented training datasets with various distortions may be of worth in order to improve generalization and robustness, especially in adversarial environments, where pure input is not guaranteed.

Author Contribution

Conceptualization, N.E.E. and A.S.; Methodology, N.E.E. and A.S.; Software, N.E.E.; Validation, A.S. and H.M.; Formal Analysis, N.E.E.; Investigation, N.E.E. and A.S.; Resources, A.S.; Data Curation, N.E.E.; Writing—Original Draft Preparation, N.E.E.; Writing—Review and Editing, A.S., H.M., and A.A.; Visualization, N.E.E. and A.S.; Supervision, A.S., H.M., and A.A.; Project Administration, A.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The dataset used in this study will be made available upon request to the corresponding author for research purposes.

Conflicts of Interest

The author declares that there are no conflicts of interest related to the content or publication of this research.

Reference

[1] K. Yun, J. Park and J. Cho, "Robust human pose estimation for rotation via self-supervised learning," IEEE Access, vol. 8, p. 32502–32517, 2020.

[2] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, p. 8680–8694, 2022.

[3] J. Dong, Y. Chen, Y. Wu, J. Han and E. Ding, "Robust pose estimation under occlusion via compositional human recovery," IEEE Transactions on Image Processing, vol. 31, p. 1969–1982, 2022.

[4] F. Zhang, X. Zhu, M. Ye and S. Yang, "Distribution-aware coordinate representation for human pose estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[5] C. Zheng, L. Xu, W. Xie and M. Wang, "Degraded image pose estimation: A benchmark and study on low-quality images," IEEE Transactions on Image Processing, vol. 32, p. 4658–4671, 2023.

[6] A. Samkari, H. Kim and J. Kim, "Evaluating the robustness of human pose estimation models to compression artifacts," Sensors, vol. 23, no. 12, p. 5678, 2023.

[7] "A person on a mountain view," [Online]. Available: https://www.youtube.com/watch?v=7RyiS-mrp1c.

[8] P. Contributors, "A Man Walking by the Roadside in the City at Nighttime," 2019. [Online]. Available: https://www.pexels.com/video/a-man-walking-by-the-roadside-in-the-city-at-nighttime-3226454/.

[9] P. Contributors, "Blurred Crowd of People Walking," 2021. [Online]. Available: https://www.pexels.com/video/blurred-crowd-of-people-walking-852107/.

[10] Clideo, "Clideo Online Video Editor," [Online]. Available: https://clideo.com/.

[11] Google, "MediaPipe," 2023. [Online]. Available: https://mediapipe.dev.

[12] A. K. Singh, V. A. Kumbhare and K. Arthi, "Real-time human pose detection and recognition using MediaPipe," in International Conference on Soft Computing and Signal Processing.

[13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee and e. al., "Mediapipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[14] U. Dedhia, P. Bhoir, P. Ranka and P. Kanani, "Pose Estimation and Virtual Gym Assistant Using MediaPipe and Machine Learning," in 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), IEEE.

[15] J.-W. Kim, J.-Y. Choi, E.-J. Ha and J.-H. Choi, "Human pose estimation using MediaPipe pose and optimization method based on a humanoid model," Applied Sciences, vol. 13, no. 4, p. 2700, 2023.

[16] W. Simoes, L. Reis, C. Araujo and J. Maia Jr, "Accuracy assessment of 2D pose estimation with MediaPipe for physiotherapy exercises," Procedia Computer Science, vol. 251, p. 446–453, 2024.

[17] Y. Xu, J. Zhang, Q. Zhang and D. Tao, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation," Advances in Neural Information Processing Systems, 2022.

[18] Y. Xu, J. Zhang, Q. Zhang and D. Tao, "ViTPose+: Vision Transformer Foundation Model for Generic Body Pose Estimation," arXiv preprint arXiv:2212.04246, 2022.

[19] Y. Xu, Q. Zhang, J. Zhang and D. Tao, "Vitae: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias," Advances in Neural Information Processing Systems, vol. 34, 2021.

[20] Q. Zhang, Y. Xu, J. Zhang and D. Tao, "ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond," arXiv preprint arXiv:2202.10108, 2022.

[21] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee and e. al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint, 2019.

[22] Hysts, "ViTPose Video Demo," [Online]. Available: https://huggingface.co/spaces/hysts/ViTPose_video.

[23] G. Blocks, "ViTPose Demo," [Online]. Available: https://huggingface.co/spaces/Gradio-Blocks/ViTPose.