

ProGait: A Multi-Purpose Video Dataset and Benchmark for Transfemoral Prosthesis Users

Xiangyu Yin Boyuan Yang Weichen Liu Qiyao Xue
Abrar Alamri Goeran Fiedler Wei Gao
University of Pittsburgh

{eric.yin, by.yang, weichenliu, qiyao.xue, abal14, gfiedler, weigao}@pitt.edu

Abstract

Prosthetic legs play a pivotal role in clinical rehabilitation, allowing individuals with lower-limb amputations the ability to regain mobility and improve their quality of life. Gait analysis is fundamental for optimizing prosthesis design and alignment, directly impacting the mobility and life quality of individuals with lower-limb amputations. Vision-based machine learning (ML) methods offer a scalable and non-invasive solution to gait analysis, but face challenges in correctly detecting and analyzing prosthesis, due to their unique appearances and new movement patterns. In this paper, we aim to bridge this gap by introducing a multi-purpose dataset, namely ProGait, to support multiple vision tasks including Video Object Segmentation, 2D Human Pose Estimation, and Gait Analysis (GA). ProGait provides 412 video clips from four above-knee amputees when testing multiple newly-fitted prosthetic legs through walking trials, and depicts the presence, contours, poses, and gait patterns of human subjects with transfemoral prosthetic legs. Alongside the dataset itself, we also present benchmark tasks and fine-tuned baseline models to illustrate the practical application and performance of the ProGait dataset. We compared our baseline models against pre-trained vision models, demonstrating improved generalizability when applying the ProGait dataset for prosthesis-specific tasks. The ProGait dataset is available at <https://huggingface.co/datasets/ericxy98/ProGait>, and the source codes of our benchmark tasks are available at <https://github.com/pittisl/ProGait>.

1. Introduction

Prosthetic legs have been a transformative solution in clinical rehabilitation, allowing individuals with lower-limb amputations the ability to regain mobility and improve their quality of life. The success of a prosthetic leg relies heavily on its alignment on individual users that is crucial to comfort, stability, and efficient movement [34], and gait analysis plays a pivotal role in achieving optimality in such alignment, by providing insights into the user’s individualized

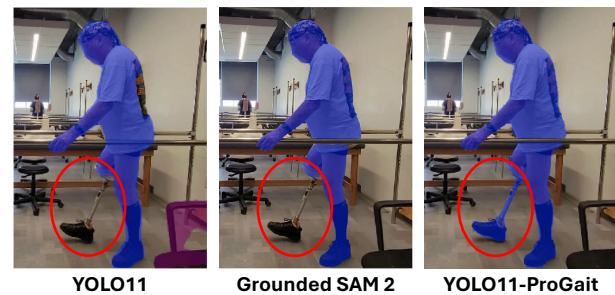


Figure 1. Limitation of current vision-based ML models on detecting the whole body of users with prosthetic legs. Left: Pre-trained YOLO11 model [15]; Middle: Grounded SAM2 model [26] with text prompt “a person with prosthetic leg”; Right: YOLO11 model fine-tuned on our ProGait dataset

walking patterns [2, 27, 35] and further indicating fitness of the prosthetic leg on each user. Moreover, gait analysis helps identify misalignments, assess biomechanical efficiency, and make precise adjustments to the prosthesis, ultimately enhancing mobility and per-user functionality.

Traditionally, gait analysis relied on specialized motion capture systems or embedded on-body motion sensors to measure biomechanical parameters [3, 5, 21, 25]. While these approaches provide high-fidelity data, they are often expensive, intrusive, and limited to controlled environments, thereby constraining their accessibility and practicality in real-world applications at scale.

Vision-based machine learning (ML) approaches have recently emerged as a promising alternative [17, 28], and can extract spatio-temporal information about human motion directly from video data, enabling non-invasive, scalable, and cost-effective gait analysis. However, current vision models face significant limitations in accurately detecting and analyzing individuals with transfemoral prosthetic legs [4], as shown in Figure 1. This is because current models are mostly trained on large datasets that only contain able-bodied individuals and hence fail to account for the unique appearance and movement patterns of prosthetic limbs. This gap in model performance not only hampers

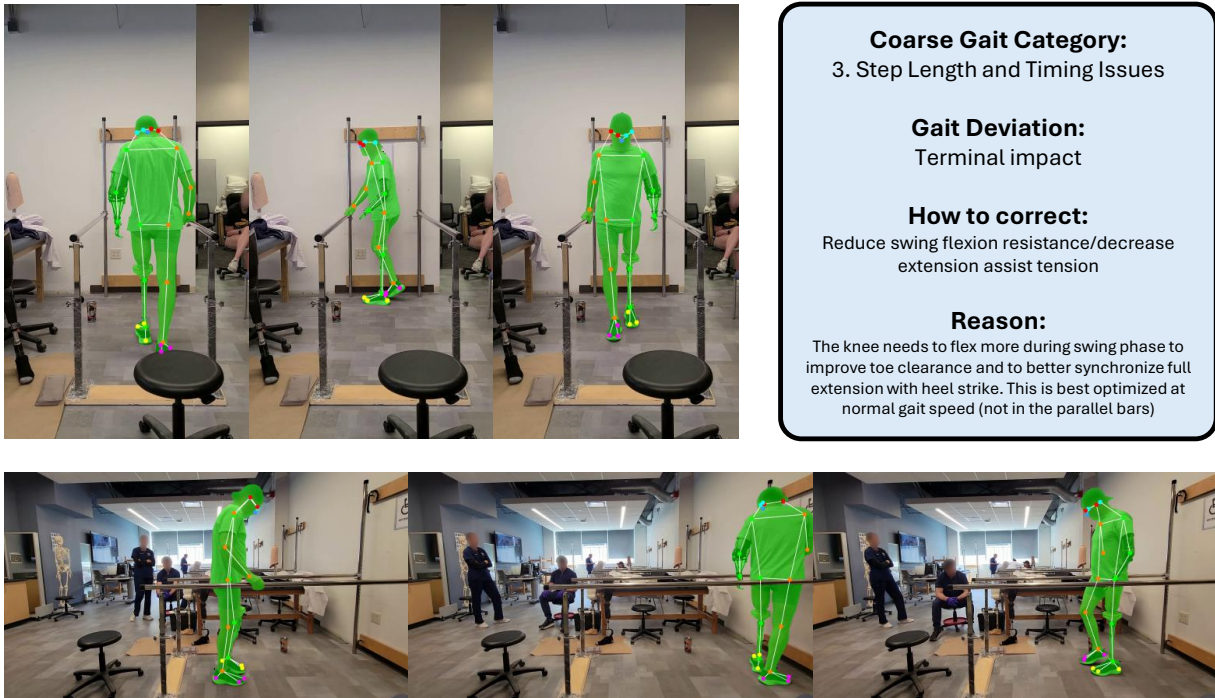


Figure 2. Examples of video frames and annotations from the ProGait dataset. **Top-left:** Frontal view with segmentation masks and pose keypoints. **Bottom:** Sagittal view with segmentation masks and pose keypoints. **Top-right:** Textual descriptions for gait analysis. This sample was collected in the inside parallel bar scenario. Additional samples are available in the supplementary material.

accurate gait analysis, but also impedes many downstream tasks such as rehabilitation assessment [7] and prosthesis optimization [22] that are important to human well-beings.

To address this challenge, in this paper, we present a new multi-purpose video dataset, namely *ProGait*, that depict the presence, contours, poses, and gait patterns of human subjects with transfemoral prosthetic legs. As shown in Figure 2, our dataset is designed to mainly support three important tasks in vision-based analysis: 1) Video Object Segmentation (VOS), 2) 2D Human Pose Estimation (HPE), and 3) Gait Analysis (GA). By providing high-quality annotations and diverse scenarios, this dataset enables evaluation and fine-tuning of vision models, to improve their performance on prosthesis-specific tasks, as well as in-depth studies of prosthesis dynamic alignment. More video samples are illustrated in Appendix C.

We collected 412 video clips from four above-knee amputees, when testing multiple newly-fitted prosthetic legs through walking trials. The videos encompasses the following to primary scenarios as shown in Figure 3:

1. **Walking inside the parallel bars without assistance:** Subjects navigate the parallel bars independently, focusing on balance and stability.
2. **Walking outside the parallel bars (in hallways) with assistance:** Subjects ambulate in indoor spaces with external support, simulating real walking conditions.¹

¹The assistants do not have direct contact or provide any physical as-

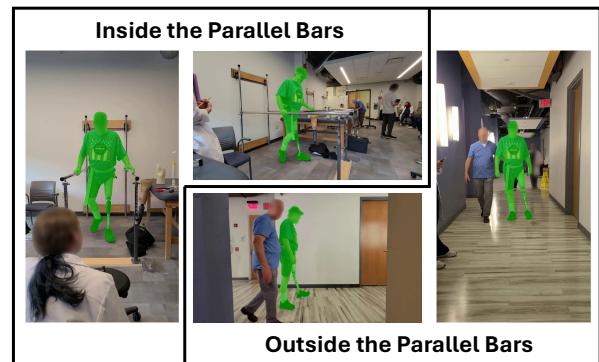


Figure 3. The two scenarios

Each walking trial includes both frontal and sagittal views, providing comprehensive perspectives for analysis. To ensure diversity and generalizability, the trials on each subject involve various types and configurations of prosthetic legs, different background contexts and lighting conditions, and heterogeneous presence of other human individuals. The dataset covers a diverse range of normal and abnormal gait patterns, each of which is accompanied by detailed textual descriptions from researchers in rehabilitation sciences and human engineering. These descriptions outline the correlations between abnormal gait deviations and the necessary corrective adjustments in order to regain

assistance during the data collection. Their presence was solely for safety assurance, and they intervene only in the event of a potential fall.

normal gaits, as well as detailed reasons about why such adjustments are needed. All the video data collection, annotations and text descriptions have been approved by the institutional IRB at University of Pittsburgh.

We also provided benchmark tasks and fine-tuned YOLO11 [15] and RTMPose [12] models as baselines, to demonstrate how this dataset can be used to practically advance research in enhancing the accessibility and effectiveness of prosthetic solutions through vision models. We evaluate our baseline models against SOTA vision models in a zero-shot setup, with results showing that our baseline models outperform the SOTA models by 9% in VOS task and 10-30% in 2D HPE task. We also trained a top-down classification model for identifying 9 different gait patterns, and the overall accuracy is up to 81.2% when taking sagittal views alone. These findings validate the effectiveness of our dataset in improving the generalizability of vision models for individuals with prosthetic legs.

2. Related Work

2.1. Gait Analysis and Vision AI Models

Gait analysis has been the cornerstone of understanding human locomotion, and used to rely on motion capture systems [3, 21], force plates [1, 20], and embedded motion sensors [5, 25] to capture biomechanical parameters such as joint angles, ground reaction forces, and stride lengths. These methods, however, are limited by their reliance on specialized equipment, controlled laboratory settings and high operational costs. Vision-based AI models, such as YOLO [25] and the Segment Anything Model (SAM) [18], allow more scalable, flexible and non-invasive gait assessment [17, 28], by extracting the spatio-temporal features from video data and transforming gait analysis into typical vision tasks such as pose estimation, object tracking, and semantic segmentation.

Despite these advancements, current vision models often struggle to accurately detect and analyze individuals with prosthetic legs, due to the unique appearance and movement patterns of prosthetic limbs [4]. This limitation calls for specialized datasets and tailored models to improve the generalizability of vision techniques for prosthetic users.

2.2. The Existing Datasets

Datasets have been available for vision-based gait analysis. The Gait Abnormality in Video Dataset (GAVD) [23] and the Health&Gait [33] dataset offer thousands of video sequences with detailed annotations, such as semantic segmentation and human pose. However, they lack representation of individuals with prosthetic limbs. A more specialized dataset focuses on kinematics and kinetics of 18 above-knee amputees walking at various speeds [11]. Although providing detailed motion capture data and offering insights into the biomechanics of prosthetic users, it does

not provide raw videos and is instead limited by the fact that participants only use their personal prosthetic devices.

In contrast, the ProGait dataset presented in this paper addresses these gaps by including individuals with prosthetic legs, when testing multiple prosthetic designs. This diversity allows for more comprehensive evaluations and aligns with the need for scalable vision-based approaches that do not rely on expensive motion capture systems.

3. The ProGait Dataset

The ProGait dataset primarily focuses on enabling the following three tasks: 1) Video Object Segmentation (VOS), 2) 2D Human Pose Estimation (HPE), and 3) Gait Analysis (GA). ProGait contains paired video clips that record the frontal and sagittal views of walking trials conducted by human subjects with prosthetic legs. Each video pair includes the corresponding annotations, such as bounding boxes, segmentation masks and pose keypoints of the subject, accompanied by descriptive gait assessments in text. These comprehensive annotations aim to support multi-task learning and fine-tuning of vision-based models. A sample of such a video pair and associated annotations and text descriptions can be found in Figure 2.

3.1. Video Data Collection

The video data was collected from four human subjects (see Appendix B for detailed information). Each subject was tested with multiple newly-fitted prosthetic legs and performed multiple walking trials both inside and outside the parallel bars as depicted in Figure 3. During the trials conducted outside the parallel bars in the open space, a health-care professional accompanied the subjects to provide assistance and ensure their safety, particularly to prevent falls.

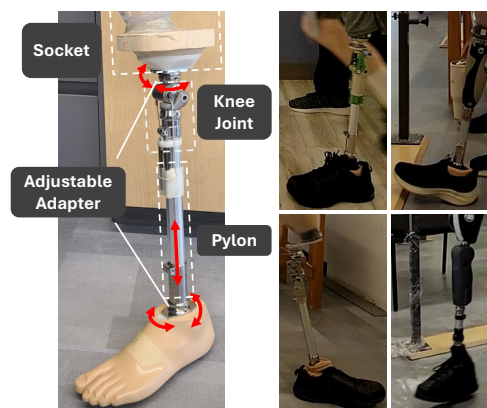


Figure 4. Components of a transfemoral prosthetic leg (left) and different types of prosthetic legs (right)

As shown in Figure 4, each prosthetic leg differs from each other, such that they have different types of knees (mechanical, hydraulic or computerized), different angles of knee and ankle joints, and different lengths of the pylon.

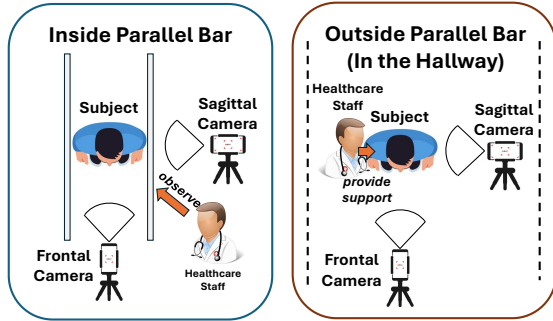


Figure 5. Video Recording Setup

Such differences affect the subject’s gait patterns in different ways, and also result in very distinct visual appearances that make it hard for the vision models to recognize.

To capture the walking trials, we used two video cameras positioned at fixed locations to record the frontal and sagittal views of the subjects, as illustrated in Figure 5. The videos were recorded at a resolution of 1920×1080 and a frame rate of 30fps. Each walking trial consists of multiple round trips. In some cases, the healthcare staff may partially or fully obstruct the subject’s view, particularly in the sagittal view during trials outside the parallel bars, as they walk alongside the subject to provide support. When this occurs, only a one-way trip is selected. The duration of selected walking trips lasts for 8-20 seconds in trials inside the parallel bars, and for 2-40 seconds for trials outside the parallel bars. During each trial, the subject walked approximately 5 meters inside the parallel bars and 10–15 meters outside the bars. The walking speed varied depending on the subject’s comfort with the prosthesis.

In total, we collected 144 walking trials, resulting in 412 video clips contained in the dataset. Table 1 presents the basic distribution of video data, including the number of video samples per subject and scenario.

Subject ID	Scenario	Trials	Samples
P1	Inside Bars	20	67
	Outside Bars	32	95
P2	Inside Bars	7	21
	Outside Bars	34	100
P3	Inside Bars	1	2
	Outside Bars	9	25
P4	Inside Bars	22	54
	Outside Bars	19	48

Table 1. Basic Distribution of ProGait Dataset

3.2. Annotations

To provide high-quality annotations for the three tasks of using the ProGait dataset, we employed SOTA vision models to provide initial annotations in an automated manner, followed by manual refinements to ensure accuracy.

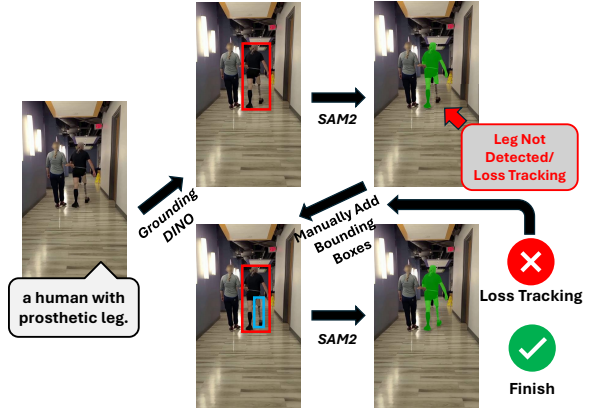


Figure 6. Pipeline of annotation

Video Object Segmentation (VOS): There are three major challenges when utilizing the pre-trained vision models for tracking the subject and generating segmentation mask:

1. Precisely detecting the prosthesis as part of the human body.
2. Correctly distinguishing the subject from the healthcare staff.
3. Ensuring consistent tracking despite occlusions.

The second and third challenges are particularly pronounced in scenarios outside the parallel bars, where multiple people, including the healthcare professional, are present alongside the subject.

As shown in Figure 6, we address these challenges by developing a Human-in-the-Loop annotation pipeline with the help of Grounded SAM [26]. The process begins with extracting the initial frame of the video and using the GroundingDINO model [19] with the prompt “a human with a prosthetic leg” to annotate the subject’s bounding box. Manual refinement is then performed to add additional bounding boxes, allowing the prosthesis to be tracked as a separate object. These bounding boxes are used as input for Segment Anything Model 2 (SAM2) [24], which generates segmentation masks and tracks objects across frames. To ensure accuracy, we visually inspect for any loss of tracking. If the tracking fails, supplementary bounding boxes are added at the problematic frames, and segmentation masks are regenerated to maintain consistency. This semi-automated approach reduces the need for frame-by-frame manual annotation while maintaining high annotation quality.

2D Human Pose Estimation (HPE). In this dataset, we focus on 23 pose keypoints, 17 for body and 6 for foot, defined in the COCO-WholeBody dataset [14]. After generating the segmentation mask, we apply it to the original videos and estimate the poses by applying the pre-trained RTMW model [13] within the MMPose framework [6] on the masked video. However, the first challenge mentioned above remains unresolved. As illustrated in Figure 7, the pre-trained model frequently fails to accurately detect key-

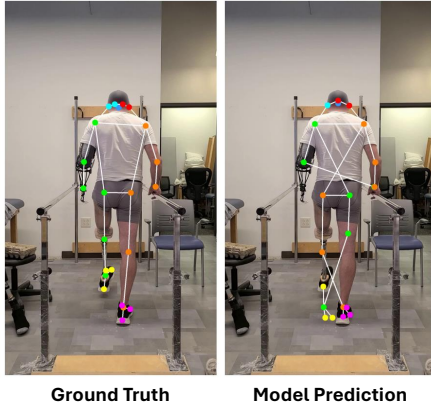


Figure 7. Zero-shot pose estimation with RTMW

points on the prosthetic knees and feet. Unlike the segmentation task, where the prosthesis can be tracked as an independent object, incorrect skeleton keypoints require labor-intensive, frame-by-frame manual correction.

To tackle this issue, we first randomly extract a small set of frames (~ 100) from the videos and manually correct the keypoints. We then fine-tune the RTMW model on these annotated frames and use the fine-tuned model to infer keypoints for a larger set of frames ($> 1,000$). Next, we visually inspect the annotations from the fine-tuned model and select high-quality samples for a second round of fine-tuning. Finally, we apply the fine-tuned RTMW model to automatically annotate pose keypoints across all video sequences, and manually correct any remaining errors. In fact, only $< 25\%$ of videos require manual correction, significantly reducing the amount of workload for dataset preparation.

Example 1

- 4
- insufficient swing phase flexion
- reduce flexion resistance (loosen spring compression in extension assist mechanism) of the knee
- not enough clearance during swing through causes tripping hazard and excessive terminal impact at extension. Part of the problem could be the unnatural slow walking speed in the parallel bars, yet it is safer to make the adjustment here and undo it later if needed at normal gait speed

Example 2

- 6
- not enough toe clearance
- reduce swing flexion resistance
- hard to see from given perspective, but likely this socket is too extended as well, making for uneven step lengths and increased (relative) hip flexion in early swing. This causes the low clearance (tripping hazard, asymmetric gait)

Example 3

- 9
- normal gait
- knee fine tuning tbd outside parallel bars
- good outcome for initial optimization session

Figure 8. Gait annotations

Gait Analysis (GA). To facilitate gait analysis, we engaged researchers in rehabilitation sciences and human engineering to provide detailed textual descriptions for each video sample. As shown in Figure 8, these descriptions consist of four key components: (1) the general gait category, (2) the specific gait deviation, (3) recommendations on how to

adjust the prosthesis to correct the gait, and (4) the reasons of these recommendations.

While our benchmark task for GA primarily focuses on the classification problem over the general gait category, we provide full access to the other three components in the dataset, enabling advanced applications such as prosthesis optimization and clinical decision-making. In total, the annotations include 9 different gait categories, with each category comprising several fine-grained gait deviations. These deviations are classified according to their relevance to critical prosthetic alignment factors, such as knee and ankle positioning. The distribution and details of these categories are listed in Appendix A.

3.3. De-identification

To protect the privacy of individuals, we implement a comprehensive de-identification process for all video clips. Using a similar approach to segmentation mask annotation, we employ GroundingDINO and SAM2 models to detect sensitive elements, such as human faces and identifiable signage, in the first few frames of video clips. We then manually add supplementary bounding boxes to cover these elements in later frames. These identified areas are then subjected to Gaussian blurring to obscure sensitive information. This semi-automated process ensures consistent and reliable de-identification while preserving the dataset’s integrity. Additionally, we conduct manual inspections to verify that all sensitive details are properly anonymized and that no identifiable features remain.

4. Benchmarks and Baseline Models

To facilitate the effective use of the ProGait dataset, we establish several benchmark tasks and provide fine-tuned baseline models as reference implementations. These benchmarks serve as an initial guide for researchers and practitioners, helping them evaluate the model performance across key tasks relevant to prosthetic gait analysis.

4.1. Benchmark Tasks and Metrics

Video Object Segmentation. For the segmentation task, we use the mean Intersection over Union (mIoU) as the evaluation metric. The mIoU measures the overlap between the predicted mask and the ground-truth mask across video frames through a pixel-wise calculation, such that

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \frac{|P_i \cap G_i|}{|P_i \cup G_i|} \quad (1)$$

where N is the total number of frames, P_i is the predicted mask for frame i and G_i is the ground-truth mask for frame i . Intersection (\cap) measures the number of pixels that are correctly predicted as belonging to the class, and Union (\cup) measures the total number of pixels that are either in the predicted mask or the ground truth. This metric, hence, evaluates if the segmentation models can accurately delineate both the natural body parts and prosthetic components,

which is crucial for downstream applications such as pose estimation and biomechanical analysis.

2D Human Pose Estimation. We evaluate the accuracy of pose estimation following the COCO-WholeBody standard [14], and use the Average Precision (AP) across different Object Keypoint Similarity (OKS) thresholds ranging from 0.5 to 0.95, namely the AP@[.5,.95] as the primary metric. This metric can be presented as:

$$AP@[0.5, 0.95] = \frac{1}{K} \sum_{t \in \{0.5, 0.55, \dots, 0.95\}} AP_t \quad (2)$$

where K is the number of OKS thresholds (in this case, $K = 10$ because the range is from 0.5 to 0.95 with a step of 0.05), and AP_t is the Average Precision calculated by computing the area under precision-recall curves at a specific OKS threshold t . OKS measures the normalized distance between the predicted and ground-truth keypoints:

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (3)$$

where d_i is the Euclidean distance between the i -th corresponding ground truth and the detected keypoint, v_i is the visibility flag of the ground truth, s is the object scale, k_i is a per-keypoint constant that controls falloff.

Gait Classification. This task categorizes video clips into predefined gait classes, by capturing the variations in movement patterns due to different prosthetic configurations, walking conditions, or rehabilitation progress. We provide an end-to-end pipeline for training and evaluating different classification models. Performance is measured using Top-1 accuracy, which indicates the proportion of correctly classified gait patterns. Balanced accuracy, defined as the average of sensitivity (recall) for each class, is also used as a reference to account for the unbalanced class distribution.

Given that gait patterns can be subtle and complex, multiple gait deviations can, and often do, co-occur in clinical reality. However, in a clinical setting, physical therapists and prosthetists typically address one primary alignment issue at a time to avoid the problem of “analysis paralysis” [10, 16]. This iterative process allows for precise adjustment and clear assessment of the impact of each modification, preventing the introduction of new problems while attempting to fix multiple simultaneously. Therefore, in our benchmark studies, we only consider the primary gait deviation for our classification task.

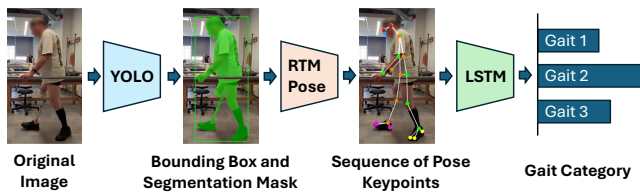


Figure 9. The pipeline of baseline models

4.2. Baseline Models

To benchmark our dataset in the aforementioned tasks, we fine-tuned several vision models listed below as baselines. These models operate in a pipeline, as shown in Figure 9, where the output of one model serves as the next’s input.

- For Video Object Segmentation task, we use **YOLO11** [15], the latest iteration in the Ultralytics YOLO series, capable of tracking multiple objects across frames.
- For 2D Human Pose Estimation task, we use **RTM-Pose** [12] from the MMPose framework. RTMPose takes the detected bounding boxes from the tracking model (YOLO11) as input and predicts pose keypoint positions.
- For Gait Classification task, we use **A custom LSTM classifier**, which processes the sequence of poses over time and classifies the gait pattern for each video sample.

When fine-tuning the pre-trained models using the ProGait dataset, the dataset is divided into $\sim 70\%$ for training, $\sim 20\%$ for validation and $\sim 10\%$ for testing. To prevent data leakage, subjects appearing in the test set do not appear in the training or validation sets.

Since the majority of our pose annotations, including those in the test set, are derived from the fine-tuned RTMW model, we refrain from using it as the baseline for 2D Human Pose Estimation task. Instead, we use an RTMPose variant and fine-tune it exclusively on the training and validation sets. During the training of RTMPose and LSTM classifier, we use ground truth data as input, rather than the outputs from preceding models. Note that, RTMPose generates 133-point whole-body poses; however, we evaluate only the 23 keypoints of the body and feet, aligning with our annotations. Since gait patterns primarily involve the lower body, we exclude keypoints of face and arms, and only 12 keypoints are used for the LSTM classifier.

5. Experiments

In this section, we present benchmark results by comparing our baseline models with other off-the-shelf models, for the aforementioned three tasks on the ProGait dataset.

5.1. Experimental Setup

For fair comparisons, all evaluations are done only on the test set, and all the subjects in the test set are not present in training or validation sets. This guarantees that the test samples are entirely unseen by the baseline models. Unlike the process of fine-tuning the baseline models described in Section 4.2, the experiment results are computed across all frames of the input videos rather than sparsely sampled frames. This approach provides a more comprehensive assessment of the model’s consistency across frames.

5.2. Video Object Segmentation (VOS)

For the VOS task, we compare our fine-tuned YOLO11 model with the original checkpoint and the Grounded SAM2 with different text prompts.

Method	mIoU	mIoU (inside)	mIoU (outside)
YOLO11	0.784	0.831	0.774
Grounded SAM2 “a person.”	0.358	0.643	0.559
“an amputee.”	0.905	0.90	0.907
“a human with prosthetic leg.”	0.964	0.921	0.929
YOLO11-ProGait	0.847	0.815	0.866

Table 2. Results for Video Object Segmentation

Since the YOLO11 model is originally trained for multi-object detection and tracking, it can produce multiple mask outputs. To evaluate its performance in tracking a single subject, we compute the mIoU using only the predicted mask with the largest intersection with the ground truth. Besides, the fine-tuned YOLO11 may occasionally detect the human body as separate instances, as shown in Figure 10. This occurs because YOLO11 can only be trained on closed shapes and cannot inherently handle discrete parts of a single subject. To address this problem, we train the YOLO11 model to learn discrete parts separately and merge all detected mask instances into a single mask for evaluation.

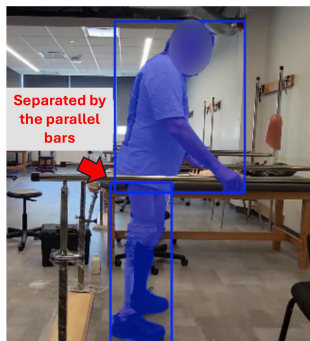


Figure 10. The fine-tuned YOLO model may detect the subject body as separate instances.

The results in Table 2 show that, Grounded SAM2 with an appropriate text prompt can achieve very good performance, but highly depends on the specific text prompt being used. Further, even with mIoU up to 96%, it can occasionally lose tracking of the prosthetic leg, as shown in Figure 1. On the other hand, although the issue of detecting body parts separately causes a slight accuracy drop in scenario of inside parallel bars, YOLO11-ProGait generally outperforms its original checkpoint. Given that prosthetic legs are typically thin and occupy a small area in the frame, the performance gap between the pre-trained and fine-tuned models remains modest, which is reasonable.

5.3. 2D Human Pose Estimation (HPE)

In the 2D HPE task, for comparison with our fine-tuned RTMPose model, we selected three pre-trained models for

2D whole-body pose estimation: HRNet [29], ViPNAS [31], and ViTPose [32] pre-trained on COCO-WholeBody dataset[14], apart from the original RTMPose checkpoint.

Method	AP	AP (inside)	AP (outside)
HRNet [29]	0.750	0.825	0.733
ViPNAS [31]	0.761	0.735	0.767
RTMPose [12]	0.855	0.876	0.850
ViTPose [32]	0.830	0.845	0.822
RTMPose-ProGait	0.947	0.968	0.942

Table 3. Results for 2D Human Pose Estimation

We first evaluated the $AP@[0.5, 0.95]$ metric for all the 23 keypoints (17 keypoints for ViTPose²), and the results are shown in Table 3. Our baseline model RTMPose-ProGait outperforms all the other pre-trained models, demonstrating the effectiveness of using the ProGait dataset to improve the prosthesis detection.

Additionally, we evaluated the AP for lower body keypoints. More specifically, we calculate the score for 10 out of the 23 keypoints, covering knees, ankles, and feet. ViTPose was excluded due to its lack of 6 feet points. The results in Table 4 shows that our RTMPose-ProGait significantly performs better than other methods on this metric. On the one hand, these results also show that the existing models struggle in detecting prostheses as part of the human body, and the ProGait dataset can well address this problem.

Method	AP-leg	AP-leg (inside)	AP-leg (outside)
HRNet	0.625	0.691	0.539
ViPNAS	0.631	0.527	0.655
RTMPose	0.804	0.761	0.814.
RTMPose-ProGait	0.918	0.918	0.918

Table 4. Results for 2D Human Pose Estimation with AP calculated on knee, ankle, and foot points only

5.4. Gait Classification

To bridge the gap between pose sequences and gait analysis, we trained a custom LSTM network with 128 hidden dimensions to process pose sequences. As an example shown in Figure 11, the (x, y) coordinates of pose keypoints show periodic motion over time. We utilize such time-series data to classify the poses into 9 general gait categories. Since no off-the-shelf model exists for this specific task, we conducted a 5-fold cross-validation and evaluated the model across five different setups, as shown in Table 5.

Notably, the LSTM network performs exceptionally well when provided with pose sequences from the sagittal view

²Due to its adherence to the standard COCO 17-keypoints format, ViTPose cannot detect the 6 feet points.

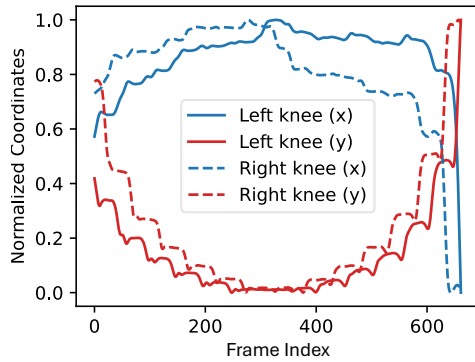


Figure 11. The sequence of pose keypoint coordinates

alone. However, when both frontal and sagittal views are fed into the model, accuracy drops significantly. This suggests that treating pose sequences from both views simultaneously may introduce confusion. Additionally, it indicates that the sagittal view may be a more suitable angle for observing gait patterns.

	Top-1 Acc.	Balanced Acc.
Frontal view	0.510	0.545
Sagittal view	0.826	0.790
Inside parallel bars	0.364	0.437
Outside parallel bars	0.486	0.320
Whole dataset	0.372	0.403

Table 5. Results of Gait Classification, with all 23 keypoints taken as input

We further compared the gait classification performance using all 23 keypoints vs. using only the 12 keypoints corresponding to the lower body. As shown in Table 6, excluding upper-body keypoints has minimal impact on classification accuracy. While upper-body movements may introduce variations in gait patterns, gait is fundamentally driven by lower-body dynamics. This suggests that lower-body keypoints alone are sufficient for gait-related downstream tasks.

	Top-1 Acc.	Balanced Acc.
Frontal view	0.474	0.521
Sagittal view	0.773	0.812
Inside parallel bars	0.364	0.457
Outside parallel bars	0.566	0.423
Whole dataset	0.384	0.413

Table 6. Results for Gait Classification, with lower-body keypoints taken as input only

For additional demonstration of ProGait’s generality, we performed more experiments over multiple gait recognition methods, including GaitGraph2[30], GaitBase[8], and

GPGait[9], by applying our dataset onto the pre-trained model and using our dataset to fine-tune the pre-trained models. For the pre-trained model, we used the generated gait embeddings for 1-nearest-neighbor classification. In fine-tuning, we appended a classification head of two FC layers to the pre-trained model, and only trained these two layers. We also conducted experiments on ScoNet [36], a silhouette-based clinical gait classification model for scoliosis detection. We trained the model from scratch using the same original settings, and adapted its architecture by changing the output dimension from 3 to 9 to align with our 9-class classification task.

Results in Table 7 show that these gait recognition/classification methods achieve reasonable performance on the gait classification task, and with simple fine-tuning/retraining, they can be well-adapted to our ProGait dataset with significantly higher accuracy. These results also showed that our simple LSTM classifier provides competitive performances.

	Pre-trained	Fine-tuned/ Re-trained
GaitGraph2 [30]	0.200	0.440
GPGait [9]	0.388	0.457
GaitBase [8]	0.313	0.340
ScoNet	N/A	0.333
LSTM-ProGait	N/A	0.372

Table 7. Results for Gait Classification on the whole dataset, measured in Top-1 accuracies.

6. Discussion and Conclusion

In this paper, we introduce ProGait, a versatile dataset designed to advance vision-based models for detecting lower-limb prostheses, facilitate comprehensive gait analysis to aid in prosthesis design and alignment, and support the development of assistive technologies that enhance mobility and overall quality of life for prosthesis users. By providing high-quality data tailored for multiple applications, ProGait aims to bridge the gap between computer vision research and real-world clinical needs, fostering innovations in prosthetic engineering and rehabilitation.

Beyond the tasks introduced in this paper, the textual descriptions of gait patterns and their corresponding reasoning unlock vast possibilities for the ProGait dataset. In future work, we plan to leverage this textual information alongside large language models (LLMs), to provide prosthesis users with quick and convenient assessments of stability, comfort, and mobility, while also reducing the workload of healthcare professionals. Ultimately, our goal is to enable the development of adaptive, interactive prosthetic legs powered by microcomputers, paving the way for intelligent, user-responsive prosthetic solutions.

Acknowledgments

We thank the reviewers and the area chair for their insightful comments and feedback. This work was supported in part by National Science Foundation (NSF) under grant number IIS-2205360, CCF-2217003, CCF-2215042, and National Institutes of Health (NIH) under grant number R01HL170368.

References

- [1] Steven C Budsberg, Mary C Verstraete, and Robert W Soutas-Little. Force plate analysis of the walking gait in healthy dogs. *American journal of veterinary research*, 48(6):915–918, 1987. 3
- [2] Andres M Cárdenas, Juliana Uribe, Josep M Font-Llagunes, Alher M Hernández, and Jesús A Plata. The effect of prosthetic alignment on the stump temperature and ground reaction forces during gait in transfemoral amputees. *Gait & Posture*, 95:76–83, 2022. 1
- [3] Elena Ceseracciu, Zimi Sawacha, and Claudio Cobelli. Comparison of markerless and marker-based motion capture technologies through simultaneous data collection during gait: proof of concept. *PloS one*, 9(3):e87640, 2014. 1, 3
- [4] Anthony Cimorelli, Ankit Patel, Tasos Karakostas, and R James Cotton. Validation of portable in-clinic video-based gait analysis for prosthesis users. *Scientific Reports*, 14(1):3840, 2024. 1, 3
- [5] Teunis Cloete and Cornie Scheffer. Benchmarking of a full-body inertial motion capture system for clinical gait analysis. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4579–4582. IEEE, 2008. 1, 3
- [6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020. 4
- [7] Alberto Esquenazi. Gait analysis in lower-limb amputation and prosthetic rehabilitation. *Physical Medicine and Rehabilitation Clinics*, 25(1):153–167, 2014. 2
- [8] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9707–9716, 2023. 8
- [9] Yang Fu, Shibe Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19595–19604, 2023. 8
- [10] Hiroshi Hashimoto, Toshiki Kobayashi, Fan Gao, and Masataka Kataoka. A proper sequence of dynamic alignment in transtibial prosthesis: insight through socket reaction moments. *Scientific reports*, 13(1):458, 2023. 6
- [11] Sarah Hood, Marshall K Ishmael, Andrew Gunnell, KB Foreman, and Tommaso Lenzi. A kinematic and kinetic dataset of 18 above-knee amputees walking at various speeds. *Scientific data*, 7(1):150, 2020. 3
- [12] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RtmPose: Real-time multi-person pose estimation based on mmPose. *arXiv preprint arXiv:2303.07399*, 2023. 3, 6, 7
- [13] Tao Jiang, Xinchun Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024. 4
- [14] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020. 4, 6, 7
- [15] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 1, 3, 6
- [16] Niels Jonkergouw, Maarten R Prins, Arjan WP Buis, and Peter van der Wurff. The effect of alignment changes on unilateral transtibial amputee’s gait: a systematic review. *PloS one*, 11(12):e0167466, 2016. 6
- [17] Taha Khan, Peter Grenholm, and Dag Nyholm. Computer vision methods for parkinsonian gait analysis: A review on patents. *Recent Patents on Biomedical Engineering (Discontinued)*, 6(2):97–108, 2013. 1, 3
- [18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3
- [19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4
- [20] Tao Liu, Yoshio Inoue, Kyoko Shibata, and K Shiojima. A mobile force plate and three-dimensional motion analysis system for three-dimensional gait assessment. *IEEE Sensors Journal*, 12(5):1461–1467, 2011. 3
- [21] Alexandra Pfister, Alexandre M West, Shaw Bronner, and Jack Adam Noah. Comparative abilities of microsoft kinect and vicon 3d motion capture for gait analysis. *Journal of medical engineering & technology*, 38(5):274–280, 2014. 1, 3
- [22] Mark A Price, Philipp Beckerle, and Frank C Sup. Design optimization in lower limb prostheses: A review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(8):1574–1588, 2019. 2
- [23] Rahm Ranjan, David Ahmedt-Aristizabal, Mohammad Ali Armin, and Juno Kim. Computer vision for clinical gait analysis: A gait abnormality video dataset. *arXiv preprint arXiv:2407.04190*, 2024. 3
- [24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4
- [25] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 1, 3
- [26] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng

- Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1, 4
- [27] Thomas Schmalz, Siegmund Blumentritt, and Rolf Jarasch. Energy expenditure and biomechanical characteristics of lower limb amputee gait: The influence of prosthetic alignment and different prosthetic components. *Gait & posture*, 16(3):255–263, 2002. 1
- [28] Jasvinder Pal Singh, Sanjeev Jain, Sakshi Arora, and Uday Pratap Singh. Vision-based gait recognition: A survey. *Ieee Access*, 6:70497–70527, 2018. 1, 3
- [29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7
- [30] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1569–1577, 2022. 8
- [31] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16072–16081, 2021. 7
- [32] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 7
- [33] Jorge Zafra-Palma, Nuria Marín-Jiménez, José Castro-Piñero, Magdalena Cuenca-García, Rafael Muñoz-Salinas, and Manuel J Marín-Jiménez. Health & gait: a dataset for gait-based analysis. *Scientific Data*, 12(1):44, 2025. 3
- [34] M Zahedi, W Spence, S Solomonidis, and J Paul. Alignment of lower-limb prostheses. *J Rehabil Res Dev*, 23(2):2–19, 1986. 1
- [35] Tengyu Zhang, Xuefei Bai, Fei Liu, and Yubo Fan. Effect of prosthetic alignment on gait and biomechanical loading in individuals with transfemoral amputation: A preliminary study. *Gait & Posture*, 71:219–226, 2019. 1
- [36] Zirui Zhou, Junhao Liang, Zizhao Peng, Chao Fan, Fengwei An, and Shiqi Yu. Gait patterns as biomarkers: A video-based approach for classifying scoliosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 284–294. Springer, 2024. 8