GOLF SWING CLASSIFICATION AND ROOT CAUSE FEEDBACK

BOWEN JIANG [JBWENJOY@SEAS], ZACHARY OSMAN [OSMANZ@SEAS], QINGQUAN BAO [QQBAO@SEAS]

ABSTRACT. We explore the explainability of deep learning models in golf swing analysis to identify areas for performance improvement. Using Grad-CAM, we examine SwingNet, a model for swing phase sequencing, and find that while it captures general biomechanical features, it often focuses on irrelevant background regions and lacks fine-grained detail. To address this, we develop a posture-only ResNet-based binary classifier that distinguishes professional from amateur swings. Grad-CAM visualizations show that professional swings maintain consistent focus across all phases, whereas amateur swings reveal specific weaknesses in certain swing stages. Our findings highlight the potential of a posture-centered approach for pinpointing problematic swing stages in amateurs, offering a foundation for actionable, data-driven feedback.

1. INTRODUCTION

The golf swing is a complex interplay of biomechanics, skill, and precision. Understanding the factors behind good or bad shots is essential for improving performance and advancing golf training methods. While machine learning has shown promise in golf swing analysis, most approaches focus on prediction or motion comparison without providing actionable feedback. This project aims to bridge that gap by exploring whether deep learning can support root cause analysis and personalized feedback for golfers.

Previous studies [8, 5, 2, 4, 3] have used neural networks to classify swing phases, differentiate professionals from amateurs, and detect motion discrepancies through latent space analysis. While insightful, these methods fail to deliver detailed error explanations or tailored improvement suggestions. Our goal is to develop a system that identifies root causes of performance issues and offers personalized, data-driven insights to enhance training efficiency.

To this end, we take two key steps in this project: First, we explore the explainability of SwingNet [6], a model trained to sequence golf swing phases, using Grad-CAM [7] to visualize its predictions. SwingNet captures broad biomechanical features, but struggles with irrelevant background focus and lacks fine-grained recognition of swing details. Second, we develop a custom ResNet-based binary classifier that uses posture-only images extracted via MMpose [1] to classify swings as professional or amateur. Grad-CAM analysis reveals that professional swings exhibit consistent attention across all phases, while amateur swings highlight specific deficiencies, such as unstable finishes or inconsistent downswing mechanics.

Our contributions are threefold:

(1) We are the first to investigate the explainability of golf-related neural networks to our best knowledge.
(2) We propose a background-agnostic, posture-centered pipeline for classifying professional/amateur golf swings.
(3) Our approach provides insights into which swing stages contribute most to classification, enabling targeted diagnosis of amateur swing deficiencies.

2. RELATED WORK

2.1. **Pose Extraction.** Pose extraction simplifies human motion into keypoints (e.g., joints like elbows and knees), enabling analysis of complex body dynamics. MMPose [1], developed by OpenMMLab, provides a robust pipeline for 2D and 3D pose estimation, supporting state-of-the-art models for applications in sports, healthcare, and robotics.

2.2. **Golf Swing Analysis.** AI-driven golf analysis uses pose estimation, temporal modeling, and neural networks to evaluate and improve golf swings. Techniques like Liu et al. (2020) [5] classify swing quality by comparing a golfer's motion against an ideal swing using 17 keypoints, processed through a ResNet model. Jiang et al. (2022) [2] enhance keypoint accuracy for fast, occluded motions with a line segment detection (LSD) algorithm. Liao et al. (2022) [4] propose a motion synchronization system using neural networks and ResNet50 for fine-grained comparisons. Kim et al. (2020) [3] reduce noise in pose data and predict swing outcomes through smoothing filters and critical phase detection. These methods enable precise swing analysis, providing valuable insights for athletes and coaches.

The GolfDB dataset [6] offers labeled video clips of professional golfers' swings with annotations for key swing phases. It utilizes the SwingNet model, combining MobileNetV2 and bidirectional LSTM to segment swing phases. The

[1]GitHub repository: `https://github.com/ESE546-Team18/Golf-Swing-Classification-and-Root-Cause-Feedback`

dataset powers systems like AI Golf [8], which uses MMPose to extract skeleton poses from keyframes and classifies swings using a ResNet50 model to distinguish between professional and amateur swings.

2.3. **AI Explainability.** AI explainability aims to make the decisions of complex models, such as deep neural networks, transparent and interpretable for human users. Grad-CAM (Gradient-weighted Class Activation Mapping) [7] is a popular method for visualizing the regions in an input image that most strongly affect a model's predictions. By leveraging the gradients of the target class with respect to the final convolutional layer of a neural network, Grad-CAM produces a heatmap overlay that highlights influential areas, offering insight into the model's focus.

3. Approach

3.1. **Overview.** Following previous research, our project is approached using a hierarchical structure that combines the SwingNet from GolfDB, MMPose, ResNet18, and the Grad-CAM. The SwingNet extracts key event frames from videos and feeds them into MMPose for 2D human skeleton extraction. Fine-tuned ResNet18 is then used to classify the swing pose as a professional one (good swing) or an amateur one (bad swing). Grad-CAM is applied both to the SwingNet and the ResNet to visualize which parts of the input images contribute most significantly to the network's decisions, providing interpretability to our model's classification results. A sketch of the pipeline is shown in Fig. 1.

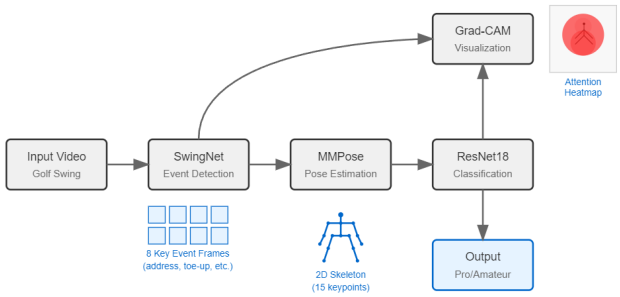


FIGURE 1. Overview of the Pipeline

Why using pose without background: While overlaying pose annotations on the original images could provide additional contextual information [8], the amateur videos in our dataset were captured in a limited set of locations with the same individuals. This lack of diversity increases the risk of overfitting to irrelevant cues, such as background scenery or clothing color. By using only the extracted pose skeletons, we eliminate these potentially misleading factors and ensure the model focuses solely on motion and posture.

3.2. **Dataset Creation Using SwingNet.** We take 200 videos from GolfDB [6] as professional data and 70 from ourselves as amateur data. The videos are 160x160 and trimmed to contain only the swing motion. Then the videos are fed into the pre-trained SwingNet to extract the eight key-frames during a swing as an example shown in Figure 2.

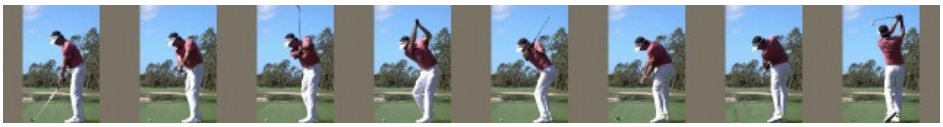


FIGURE 2. Frames of the Eight Key Events (from GolfDB). From left to right: 1. Address, 2. Toe-up, 3. Mid-backswing, 4. Top, 5. Mid-downswing, 6. Impact, 7. Mid-follow-through, 8. Finish.[6]

3.3. **2D Pose Extraction Using MMPose.** We employed MMPose's RTMPose framework, which offers an excellent balance between accuracy and computational efficiency. Specifically, our pipeline uses two pre-trained models:

(1) **RTMDet-m for person detection**, trained on the COCO and Object365 datasets. This model ensures robust person detection even in varying lighting conditions and camera angles common in golf videos.
(2) **RTMPose-m for keypoint detection**, trained on multiple human pose datasets for 420 epochs. The model outputs 17 keypoints following the COCO format, providing comprehensive coverage of major body joints essential for golf swing analysis. It's ability to handle partial occlusions and varied poses makes it particularly suitable for golf swing sequences where rapid movements and self-occlusions are common.

Why 2D pose instead of 3D: While 3D pose could potentially provide more comprehensive coaching feedback, accurate 3D reconstruction from single-perspective 160x160 images presents significant challenges. Our experiments with MMPose's 3D pose estimation models revealed unreliable results with anatomically incorrect body configurations. Since professional golf instructors can effectively assess swing quality using 2D video playback alone, we believe 2D pose representation is sufficient for meaningful analysis.

3.4. **Dataset Augmentation.**

Mirror-flipping: Mirror-flipping is first applied to randomly selected data for argumentation. Since the number of amateur videos is much less than professional ones, we used a higher argumentation probability on amateur data and a lower one on professional data for a better balance.

Handling Unclean Data Using Event Frame Dropout: SwingNet outputs the probability of belonging to each of the events (including the blank event) for every frame. Due to different background environment conditions, our videos sometimes get imperfect results from SwingNet, with some key events missing. We observed that the last two events are more likely to be missing, which could introduce bias into our amateur dataset.

Inspired by neural network dropout techniques, we used a data augmentation strategy as shown in Fig. 3. The procedure is:

(1) Randomly select clean samples from both professional and amateur datasets.
(2) Randomly disable two of the eight key event frames and create two new data points without these events.
(3) For amateur samples, only perform dropout among the first six events for a more balanced event distribution.

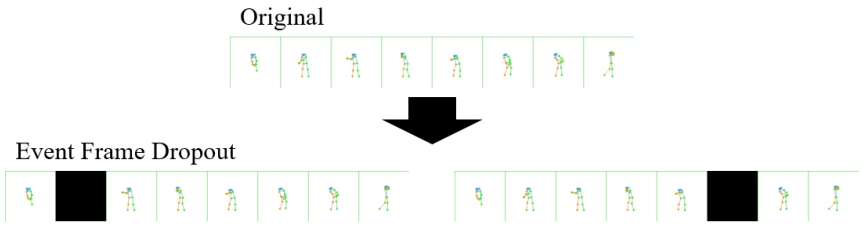


FIGURE 3. An Example of Event Frame Dropout: the original image is replaced by two new images with the 2nd and the 6th event removed

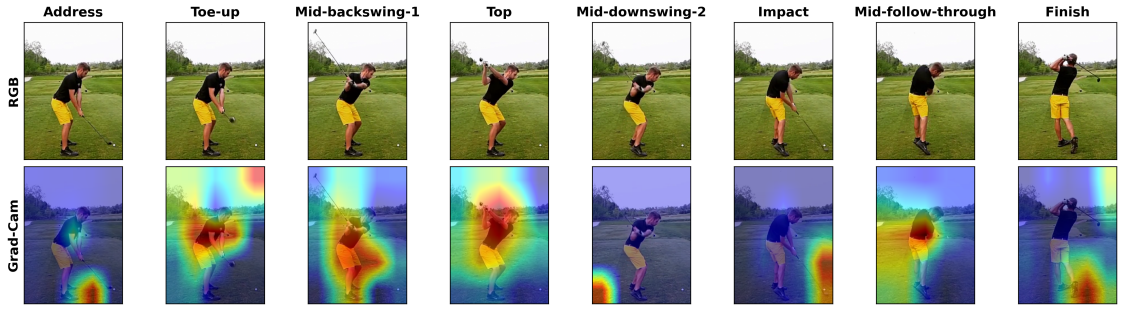
Ultimately, we end up having 369 professional and 200 amateur swing sequences. This represents a more balanced distribution compared to the original 20:7 ratio, while carefully avoiding excessive bias that could be introduced through aggressive data augmentation.

3.5. **Fine-tuning ResNet.** To classify golf swings as professional or amateur, we fine-tuned a pre-trained ResNet-18 model, chosen for its balance between efficiency and accuracy. The original fully connected layer was replaced with a sequence of a 2048-unit hidden layer with ReLU activation and 30% dropout, followed by the output layer. The model was trained using cross-entropy loss and the Adam optimizer with different learning rates for the original ResNet weights and the newly added layers. Please refer to Appendix. A for details.

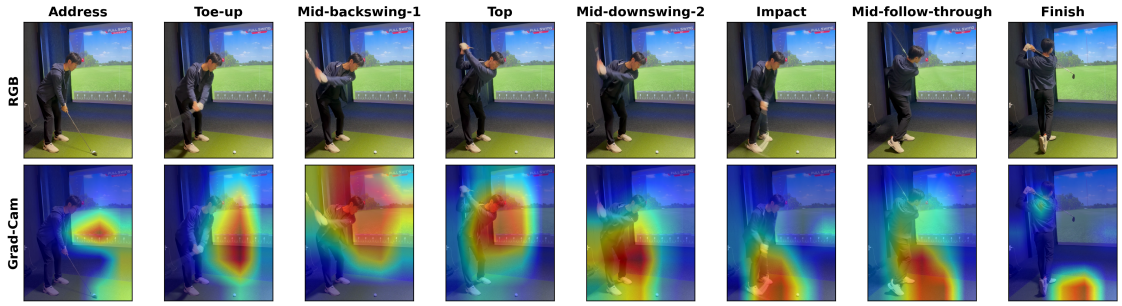
3.6. **Incorporating Grad-CAM.** We use Grad-CAM to visualize the heatmap of gradients on the original images. Following the Grad-CAM repository's recommendation, we selected the final convolutional layer as the target layer for analysis of the ResNet. For each image in our dataset, we obtained the predicted class from our trained model and set it as the target for Grad-CAM. We created a CAM object to handle the model and layer selection, then passed the input image and target class into the object to generate the heatmap visualization. This process allowed us to overlay the heatmap on the input image, providing a clear view of the specific parts of the image that influenced the classification.

4. EXPERIMENTAL RESULTS

4.1. **Explanability of Swingnet.** We evaluated the pretrained SwingNet model, trained with GolfDB, on two test videos and analyzed its predictions using Grad-CAM to visualize the contributing pixels for each stage of the golf swing. The visualizations in Figure 4a reveal that SwingNet has successfully learned several key features to differentiate swing stages. For example: During the Toe-up stage, the model focuses on the golfer's twisting and leaning body posture. In the Mid-backswing, the focus shifts toward the legs, indicating their role in stabilizing the motion. At the Top stage, the heatmap highlights the head and upper body, showing the model's recognition of balance and rotational alignment. During the Impact stage, SwingNet emphasizes the club-ball contact area, showing the critical strike moment.



(A) Professional player video result.



(B) Amateur player video result.

FIGURE 4. Two case studies of Grad-Cam on Swingnet

These observations suggest that the model has captured certain biomechanical patterns essential for identifying swing stages. However, several notable issues arise:

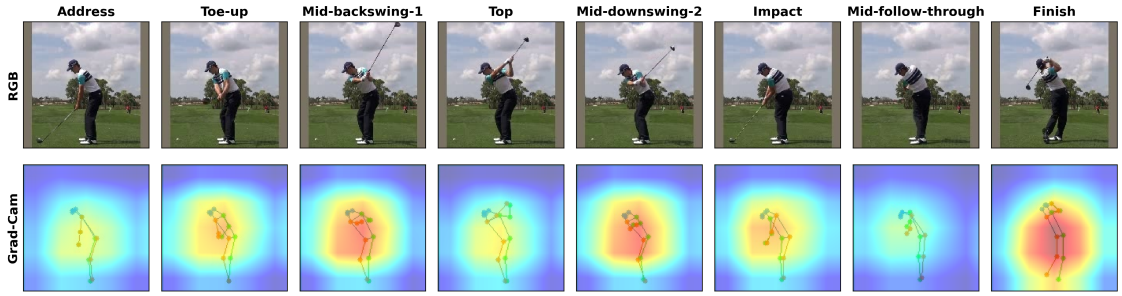
(1) Irrelevant Focus on Background Features: In some stages, the model's attention includes irrelevant regions in the background. For instance, during the Mid-downswing stage (confidence score: 0.889), the heatmap highlights empty grass in the left corner rather than focusing on the golfer's posture or club position.
(2) Limited Recognition of Fine-Grained Features: SwingNet's attention lacks the granularity expected from a human coach's perspective. For example: In the Mid-backswing stage, a coach would emphasize the golfer's arm alignment parallel to the ground, which is overlooked in the heatmap. At the Top stage, SwingNet does not identify the alignment of the club parallel to the ground, a hallmark of proper form.

These limitations highlight the model's difficulty in identifying the subtle, biomechanical nuances that define professional golf techniques. In contrast, when applied to the amateur player's video in Figure 4b, the limitations are even more pronounced. Our dataset, derived from non-professional golfers, falls outside the distribution of SwingNet's training data, exacerbating the issues observed. The heatmaps for amateur golfers show substantial attention to irrelevant areas, such as the indoor simulator's screen, rather than the golfer's posture or movement. Besides, it shows a weak ability to generalize as SwingNet struggles to detect meaningful swing features in amateur performances.

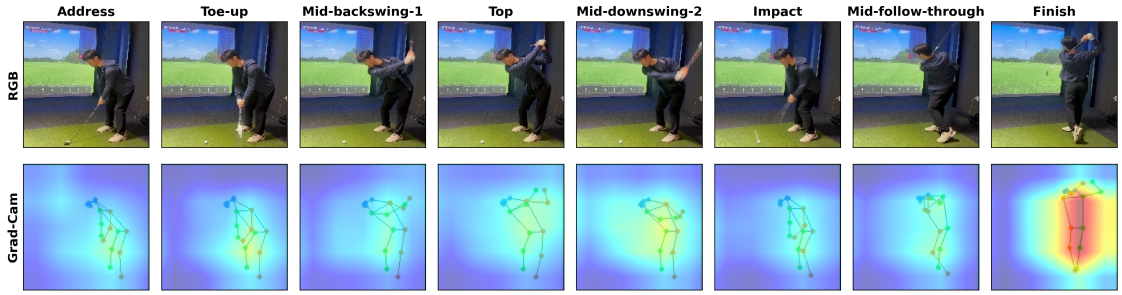
4.2. **ResNet Grad-Cam Analysis.**

The ResNet in our pipeline is used to classify keyframe poses as either professional or amateur swings. By passing a set of frames through the ResNet, we obtain a swing classification prediction. We then use Grad-CAM to visualize which parts of the input image contributed to this classification. Specifically, for a swing classified as amateur, Grad-CAM generates a heatmap on the original keyframes, highlighting the areas that contributed to the swing's poor classification. We present the results of a Grad-CAM variant, ScoreCam, on the ResNet for an amateur and a professional swing in Figure 5.

Analyzing the results of the amateur swing with ScoreCam, we observe that the heatmap is relatively weak across most of the swing, suggesting that it is not significantly flawed. However, the mid-downswing and finish frames have the highest heatmap values, indicating that these phases are most closely associated with the swing being classified as amateur. In contrast, the professional swing shows consistently high heatmap values throughout, indicating that the entire swing strongly correlates with the professional classification.



(A) Professional player video result.



(B) Amateur player video result.

FIGURE 5. Two case studies of ScoreCam on ResNet

We also tested several Grad-CAM variants to evaluate which provided the most distinct visualizations. The results of these tests, which can be found in Fig. 6 and 7 in Appendix. B, show that ScoreCam produced the most concentrated heatmap among the methods tested.

5. DISCUSSION

Our findings highlight the potential and limitations of using deep learning models for golf swing analysis. SwingNet visualizations demonstrated the model's ability to capture broad biomechanical features but revealed challenges with generalization and fine-grained feature detection, particularly for amateur swings. To address this, we developed a posture-centered binary classifier using ResNet18 and Grad-CAM, enabling a focus on critical swing stages.

Analyzing the ResNet component of our pipeline with GradCam provided valuable insights into which frames were most influential in the classification of the swing. For amateur swings, the results were inconsistent, with no clear pattern emerging across different swings. However, the finish of the swing was generally a significant contributor to the amateur classification. Interestingly, some intermediate frames were highlighted as more critical than the finish for certain swings, which is expected given the variability in amateur swing form. Different parts of the swing may vary in quality, leading to fluctuations in what is classified as the most amateur-like.

In contrast, professional swings exhibited more consistent heatmaps, indicating that most, if not all, keyframes contributed to the professional classification. This is understandable, as professionals have more refined technique, and their swings tend to be more uniform across all phases.

Further explorations for this project may include:

(1) **Expanding dataset diversity.** Our current dataset, based on a few videos of friends and ourselves, lacks diversity and is not representative of the broader amateur golf community. Including players of various skill levels and adopting a more detailed classification system would enhance real-world applicability.

(2) **Adopting advanced pose analysis architectures.** While ResNet18 performs well on our dataset, more diverse and complex data would benefit from models like Video Transformers for temporal modeling, Temporal Convolutional Networks (TCN) for sequence analysis, or Graph Neural Networks (GNN) for joint relationship modeling.

(3) **Incorporating multi-modal data.** Combining additional data sources, such as depth information, IMU sensors, or language models, could improve system performance. A multi-modal approach would provide more holistic feedback, enabling the development of a comprehensive AI golf coaching system.

(4) **Fine-Grained Feature Localization:** Currently, our pipeline identifies which swing stages are problematic but does not pinpoint the specific body parts contributing to errors. Incorporating fine-grained feature localization, such as joint-specific heatmap analysis or pose alignment metrics, could provide actionable, body-part-level feedback to amateur golfers.

REFERENCES

[1] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[2] Zhongyu Jiang, Haorui Ji, Samuel Menaker, and Jenq-Neng Hwang. Golfpose: Golf swing analyses with a monocular camera based human pose estimation. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2022.

[3] Theodore Kim, Mohamed Zohdy, and Michael Barker. Applying pose estimation to predict amateur golf swing performance using edge processing (february 2020). *IEEE Access*, PP:1–1, 08 2020.

[4] Chen-Chieh Liao, Dong-Hyun Hwang, and Hideki Koike. Ai golf: Golf swing analysis tool for self-training. *IEEE Access*, 10:106286–106295, 2022.

[5] Jen Jui Liu, Jacob Newman, and Dah-Jye Lee. Body motion analysis for golf swing evaluation. In George Bebis, Zhaozheng Yin, Edward Kim, Jan Bender, Kartic Subr, Bum Chul Kwon, Jian Zhao, Denis Kalkofen, and George Baciu, editors, *Advances in Visual Computing*, pages 566–577, Cham, 2020. Springer International Publishing.

[6] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[8] Zihan Yi. Ai golf: Golf swing discrepancy detector. https://github.com/harryyizihan/ai_golf_swing, 2023.

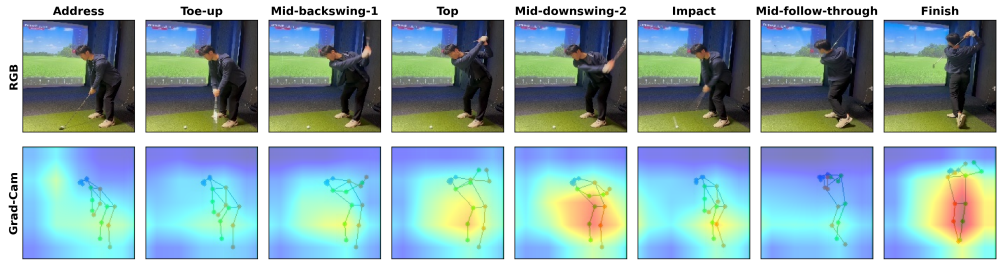
APPENDIX A. FINE-TUNING RESULT OF RESNET

ResNet is fine-tuned using a learning rate of 1×10^{-5} for the original ResNet-18 layers and 1×10^{-3} for newly added layers. Learning rates decay with $\gamma = 0.7$. Fine-tuned ResNet showed rapid convergence, improving from 63.71% to 99.35% training accuracy over 20 epochs and achieved 96.55% accuracy on the test set. Table 1 summarizes the model's performance metrics on the test set. The confusion matrix revealed only 4 misclassifications out of 116 test samples, demonstrating the model's strong capability in distinguishing professional and amateur swings based on key event frames.

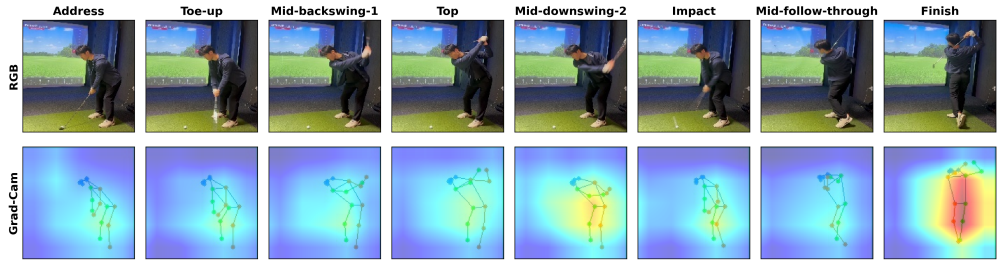
TABLE 1. Performance Metrics of the Fine-tuned Model

Metric	**Amateur (Bad)**	**Professional (Good)**
Precision	0.9348	0.9857
Recall	0.9773	0.9583
F1-score	0.9556	0.9718
Overall Accuracy		96.55%
ROC-AUC Score		0.9678

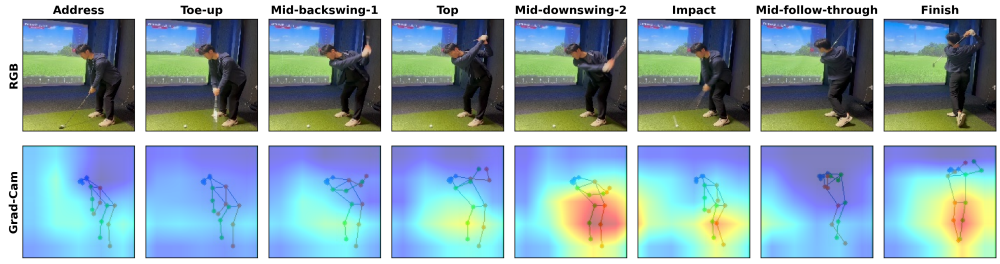
APPENDIX B. TESTING DIFFERENT GRAD-CAM VARIANTS



(A) AblationCam

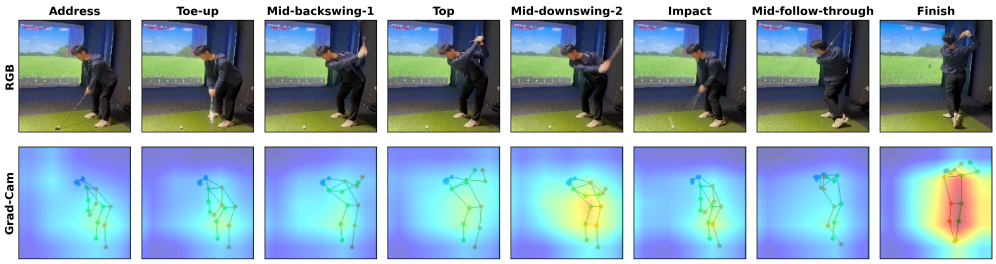


(B) ElementWiseCam

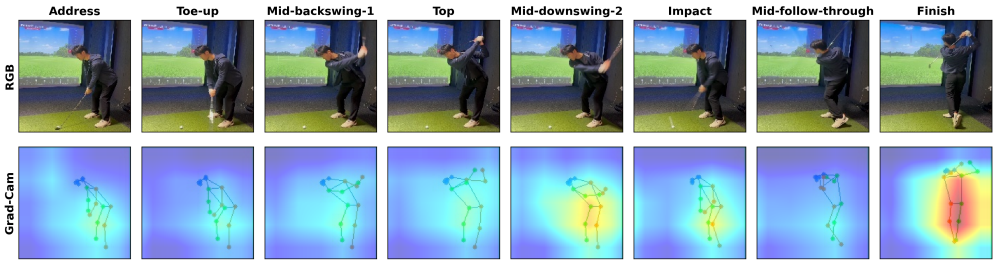


(C) HiResCam

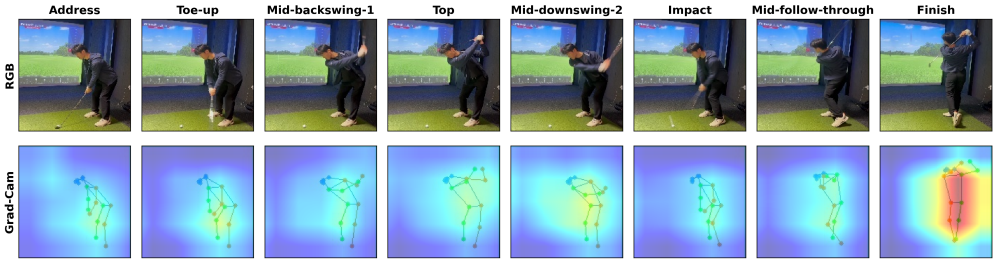
FIGURE 6. Testing Grad-CAM Variants (Part 1)



(A) LayerCam



(B) GradCamPlusPlus



(C) ScoreCam

FIGURE 7. Testing Grad-CAM Variants (Part 2)