The Fingerprint of a Shot: Pose-Based Embeddings for Personalized Basketball Form Analysis

Landon Dolvin Michael Zhang
Massachusetts Institute of Technology (MIT)
landond@mit.edu, mmz@mit.edu

Abstract

We present a two-phase pipeline for analyzing basketball shooting form using pose-sequence embeddings learned through contrastive learning. In Phase 1, we employ self-supervised contrastive pretraining to learn general motion representations from unlabeled 2D keypoint sequences. Phase 2 then refines these embeddings using supervised triplet loss for two downstream tasks: shooter identification and shot-type classification. Despite subtle visual differences between shot types, our method achieves strong classification accuracy across 100 trials, including 91.6% on shooter identity and 90.2% on shot-type recognition for one participant. Beyond classification, we introduce interpretable tools, such as directional deviation heatmaps, pose overlay scrubbing, and t-SNE visualizations, that allow athletes to compare their form to baselines, diagnose inconsistencies, and track improvement. Our results suggest embedding-based pose analysis may be a promising framework for training feedback in sports applications, as well as for personalized motion understanding more broadly.

1. Introduction

Shooting is obviously one of the most critical and individualized skills in basketball; however, most players and trainers rely heavily on qualitative assessments of form – such as video review or subjective feedback. These methods are inherently limited by human perception and bias, especially when it comes to detecting fine-grained deviations or understanding how form varies across shot types and contexts.

Recent advances in computer vision and pose estimation offer a compelling alternative: extracting structured, interpretable keypoints from video data. This enables quantitative analysis of shooting mechanics without the need for manual annotation or intrusive sensors. However, raw pose data is high-dimensional and can be difficult to analyze directly. To address this, we propose a novel two-phase pipeline that learns low-dimensional embeddings of pose sequences via contrastive learning, enabling both classification and visual interpretation of shooting form.

Our pipeline first applies self-supervised contrastive pretraining to learn general representations of mechanics from pose sequences without requiring labels. We then fine-tune the model in a supervised setting using triplet loss for two downstream classification tasks: (1) shooter identification and (2) shot-type classification. In addition to strong quantitative results on both tasks, we introduce a collection of visual tools - including t-SNE embedding maps, deviation heatmaps, and a pose-scrubbing tool, which allow users to diagnose discrepancies and gain actionable insights.

Our method is designed to support a range of use cases, from helping players understand how their form breaks down over time (e.g., due to fatigue or context) to comparing themselves to others and modeling their form after elite shooters. In this paper, we demonstrate that such a system can produce meaningful embeddings and serve as a foundation for future training tools in basketball and beyond.

2. Related Work

Our approach builds on two key lines of research: contrastive learning for self-supervised representation learning and pose-based modeling of human motion.

Contrastive learning has recently emerged as a powerful technique for learning meaningful embeddings without labeled data. Frameworks like SimCLR [1] and MoCo [3] optimize neural encoders to pull together representations of augmented views of the same input while pushing apart different inputs. These methods have demonstrated strong performance across vision, language, and other domains, and they form the basis of our self-supervised Phase 1.

Pose-based action recognition methods represent motion using sequences of skeletal keypoints extracted from video. This paradigm has proven effective for a variety of tasks, including action classification [2], sign language recognition [6] and sports performance analysis. Temporal modeling techniques such as LSTMs and 1D temporal convo-

lutions are frequently employed to capture dynamic dependencies across time, as we do in our LSTM-based embedding network.

Triplet loss and supervised metric learning are commonly used to refine embedding spaces for classifications or retrieval tasks. Triplet loss was introduced in the context of face recognition [5], where it enabled robust identification by enforcing distance constraints in the learned space. More recent work has shown its utility for fine-grained action recognition [4], where class distinctions may be subtle and hard to separate using traditional classification objectives. In our Phase 2, we apply triplet loss to encourage embedding separation across shooter identities or shot types, depending on the task.

While contrastive learning and triplet-based metric learning have seen wide application in domains such as person identification or biometric verification, our work is, to the best of our knowledge, the first to apply this two-stage embedding learning framework to basketball shooting form analysis. This opens up new opportunities for athlete-specific performance feedback grounded in both quantitative and visual interpretation.

3. Methodology

3.1. Data Collection and Preprocessing

Our dataset was constructed from a controlled basketball shooting session involving two players, Landon Dolvin and Luke Wagner, hereafter referred to as LD and LW, respectively. Each participant recorded 90 total normal shots, split into three categories: Catch and Shoots (30 shots), Side Steps (30 shots), and Step Backs (30 shots). To evaluate model robustness to poor mechanics, we also included 10 shot sequences for each shooter in which they intentionally executed distorted or ”bad” form versions of their shots.

Recordings were captured from a stationary and therefore consistent position in a well-lit indoor gym environment using an iPhone 13 camera. All videos were then manually trimmed to include only the frames from the point at which the shooter gathered the ball to release, and then later downsampled to 64 frames per sequence for consistency across samples. By trimming the videos to only include the gather-to-release segment, we eliminate the model’s ability to ”cheat” when later classifying shot types—such as by simply recognizing the pre-shot setup (e.g., whether the ball was passed for a Catch and Shoot, or whether the shooter approached from the left or right, as in a Step Back or Side Step, respectively).

3.2. Pose Extraction and Normalization

To facilitate self-supervised pretraining and downstream analysis, each video was processed to extract 2D skeletal keypoints using the YOLOv8-Pose model. This model outputs 17 joint coordinates per frame, capturing key anatomical points such as shoulders, elbows, hips, knees, and ankles. As illustrated in **Figure 1** , the extracted poses provide an expressive representation of a shooter’s form at a specific moment in time. These pose vectors were then stacked across all 64 frames.

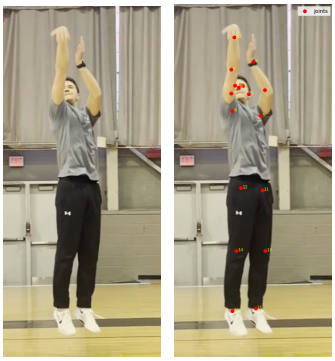

Figure 1. Original video frame (left) alongside its corresponding pose estimation output (right), generated using the YOLOv8-Pose model to extract skeletal keypoints for downstream motion analysis.

For alignment and comparability across samples, each pose sequence was normalized by translating keypoints relative to the left shoulder (joint 5). This ensured that motion dynamics—rather than absolute joint locations—were the primary features used for learning representations.

3.3. Phase 1: Contrastive Pretraining

To pretrain the model with meaningful representations of shooting motion, we adopt a self-supervised contrastive learning framework in Phase 1. The goal is to train the network to produce embeddings that cluster similar shooting sequences close together in the learned space, without relying on explicit labels.

For each video, we generate positive pairs by creating 100 different augmentations of the pose sequence. These augmentations apply Gaussian jitter and random joint masking, encouraging the model to be robust to minor variations in motion while preserving the underlying shooting form. Negative pairs are sampled from different videos. A contrastive loss is used to minimize the distance between embeddings of positive pairs while maximizing the distance from negatives.

The architecture consists of an LSTM network that processes each 64-frame pose sequence and outputs a fixed-dimensional embedding. This sequential encoder captures temporal dynamics in joint motion.

3.4. Phase 2: Supervised Fine-Tuning

After learning general motion representations in Phase 1, we fine-tune the model in a supervised manner to perform two downstream classification tasks: shooter identity (predicting whether LD or LW shot the ball), and shot type

(classifying a shot as a Catch and Shoot, Step Back, or Side Step). The latter task is considerably more challenging, as form differences are often subtle and difficult to distinguish—even for the human eye.

For this phase, we utilize Triplet Loss, which encourages the model to refine its embeddings such that they are both invariant to intra-class variation and discriminative across classes. Each training example consists of an anchor, a positive sample from the same class, and a negative sample from a different class. The loss enforces that the distance between the anchor and positive is smaller than the distance to the negative by at least a specified margin.

Triplets are generated from pose sequences that are manually labeled by shooter or shot type. The result is an embedding space where sequences are well-clustered either by shooter or, more granularly, by shot type for a specific shooter.

3.5. Evaluation Metrics

3.5.1 Classification Accuracy and Confusion Matrices

To quantitatively assess the effectiveness of our learned embeddings in downstream classification tasks, we report classification accuracy and include confusion matrices for both shooter and shot-type prediction.

Accuracy is computed as the proportion of correctly predicted labels on a randomly held-out validation set. We repeat training and evaluation over multiple randomized trials to account for variation due to noise in validation set selection and optimization, reporting both the mean accuracy and standard deviation.

In addition to accuracy statistics, we visualize confusion matrices for these randomized trials, which provide a more detailed view of performance by breaking down true versus predicted label distributions. This helps us identify specific limitations within the model's learned embeddings. These confusion matrices are especially important in the shot-type classification task, where the three classes are visually similar and difficult to distinguish.

All evaluations are performed using k-Nearest Neighbors (k-NN) with $k = 3$ on the learned embedding vectors.

3.5.2 t-SNE Visualization

To qualitatively evaluate the structure of the learned embedding space, we apply t-distributed Stochastic Neighbor Embedding (t-SNE) to project the high dimensional embeddings into two dimensions. This allows us to visualize how well the model separates classes in both Phase 1 and Phase 2.

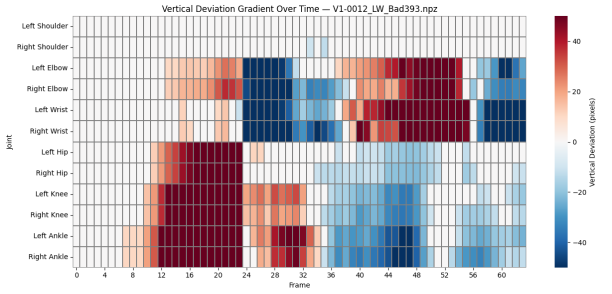

Figure 2. Vertical deviation heatmap comparing LW's normal shooting form to his purposely distorted form, highlighting frame-by-frame VERTICAL joint displacement (Red represents deviation upwards).

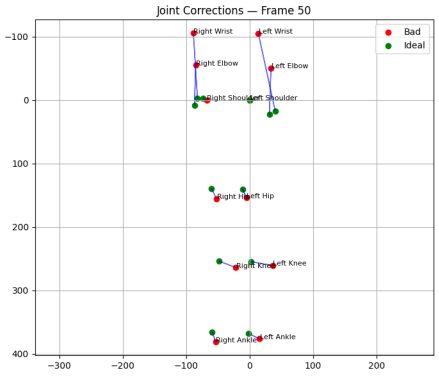

Figure 3. Pose overlay at frame 50 comparing LW's normal shooting form to his purposely distorted form. Joint corrections are visualized as arrows between corresponding keypoints, illustrating deviations frame-by-frame (interactive scrubber omitted here for static presentation).

4. Additional Analysis Tools

4.1. Directional Deviation Heatmaps

To provide frame-level interpretability of form discrepancies, we introduce directional deviation heatmaps that compare a given pose sequence (i.e., shot form) to a reference or base shooting form of choice. For each frame and joint, we compute horizontal and vertical deviations in pixel space relative to this base form—typically an "ideal" shot depending on the use case. These heatmaps make discrepancies in form easy to visualize and serve as high-level diagnostic tools for generating interpretable insights.

As an example, **Figure 2** shows the vertical deviation heatmap for LW's purposely distorted shot compared to his average normal form. Here, positive (red) values represent upward deviation, and negative (blue) values represent downward deviation. These heatmaps enable interpretable, joint-level feedback for analyzing breakdowns in shooting form.

4.2. Frame-by-Frame Pose Overlay Scrubbing Tool

To further support closer analysis of shooting form, we also implement an interactive frame-by-frame overlay tool that visualizes joint-level discrepancies between pose se-

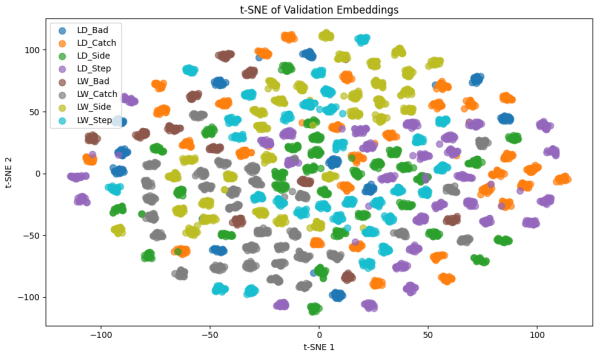

Figure 4. t-SNE visualization of validation embeddings after Phase 1 (Self-Supervised Contrastive Pretraining). Positive pairs were generated from augmented segments of the same shot, encouraging the model to learn pose-consistent embeddings. Points are color-coded by shooter and form type.

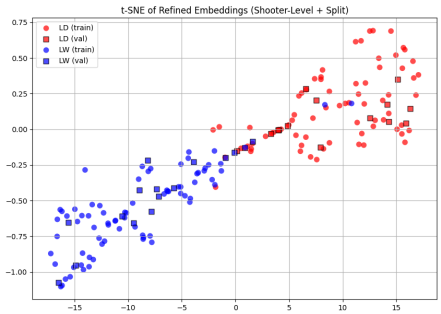

Figure 5. t-SNE visualization of refined embeddings after Phase 2 (Shooter Classification). Supervised contrastive training encourages separation between shooters. Points represent shot sequences, color-coded by shooter and split into training and validation sets.

quences. For each frame, the tool displays the corresponding joint positions from two shots, with blue lines connecting the direction and magnitude of necessary corrections to go from one form to the other. This visualization allows users to scrub through the sequence and observe how form deviates over time. **Figure 3** shows an example from this tool, with the interactive functionality omitted for static presentation.

5. Experimental Results

5.1. Embedding Quality and Clustering Behavior

As a first step in evaluating the learned embeddings, we utilize t-SNE visualizations at multiple stages of the pipeline.

Figure 4 displays the validation set embeddings immediately after Phase 1 (Self-Supervised Contrastive Pretraining). As described above, the model was trained to pull together augmented segments of the same video while pushing apart embeddings from different shot sequences. The color-coded groupings demonstrate strong performance, with distinct clusters already forming for each

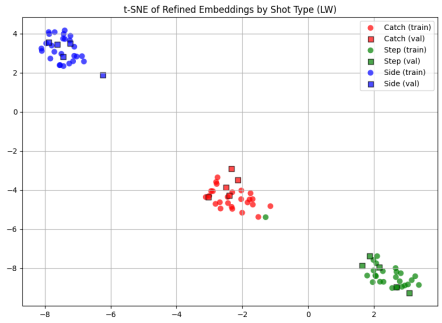

Figure 6. t-SNE visualization of refined embeddings for LW's shot types. Supervised contrastive training separates Catch, Step, and Side shots into distinct clusters. Points are color-coded by shot type and split by training vs. validation sets.

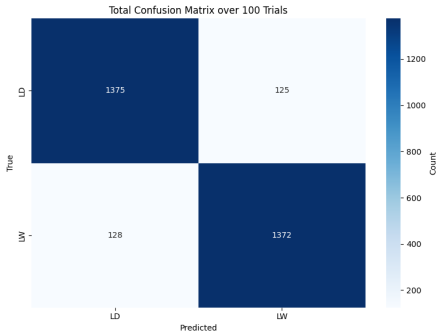

Figure 7. Confusion matrix for shooter classification over 100 trials. The model accurately distinguishes between LD and LW with minimal cross-shooter confusion. Rows represent true labels; columns represent predicted labels.

video. While not the primary goal of this stage, the model has already begun to capture differences in shooter and shot type, showing early signs of separation.

Figure 5 presents the embedding space after Phase 2 fine-tuning for shooter classification (LD vs. LW). The addition of supervised triplet loss significantly improves class separability, with clear clusters forming for each shooter. Both training and validation points are shown, indicating that the model generalizes beyond its training set.

Alternatively, **Figure 6** presents the embedding space when trained on shot-type labels (Catch and Shoot, Step Back, and Side Step) for LW. Despite the subtle visual differences between these three types of shots—especially when performed by the same shooter—the model is still able to learn well-separated clusters. This indicates that the learned representations are expressive enough to capture fine-grained distinctions in motion.

5.2. Shooter Classification Performance

To better evaluate the robustness of the learned embeddings to randomization noise, we conducted 100 randomized trials of the shooter classification task. The aggregated confusion matrix is shown in **Figure 7**, where the model

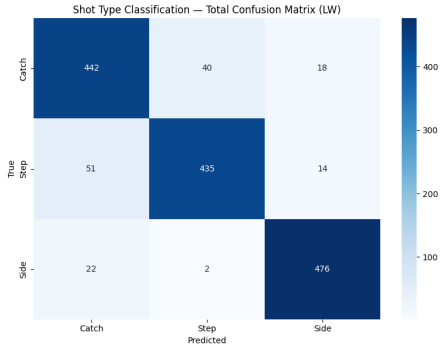

Figure 8. Shot type classification confusion matrix for LW. The model reliably distinguishes between Catch, Step, and Side shots, with minimal confusion across categories. Rows indicate true shot types; columns indicate predicted labels.

consistently distinguishes between LD and LW with minimal cross-shooter confusion. The strong diagonal dominance indicates stable prediction performance across trials for both classes.

Overall, the model achieved a mean classification accuracy of **91.57%** with a standard deviation of **4.75%**, highlighting its robustness to variation in validation splits and optimization conditions.

5.3. Shot-Type Classification Performance

To assess the model's ability to differentiate on the alternative shot-type classification task for LW, we present results in **Figure 8**. The model accurately classifies Catch and Shoot, Step Back, and Side Step shots with minimal confusion across categories. The strong diagonal in the confusion matrix once again reflects reliable and consistent predictions across all three shot types.

This task is particularly challenging because, during preprocessing, each shot was trimmed to begin after the gather phase, meaning the model **cannot rely on visual cues from the shot setup**—such as whether the ball was passed (Catch and Shoot) or whether the player moved right and laterally or left and backward (Side Step or Step Back). Instead, the model must infer shot type solely from the pose dynamics during the shooting motion itself, which may appear nearly identical to the human eye. Despite this difficulty, the model still achieves an impressive average classification accuracy of **90.20%** with a standard deviation of **7.02%** across 100 randomized trials. This result demonstrates that the learned embedding space captures subtle, yet meaningful, differences in motion that distinguish between these nuanced shot types.

6. Discussion

6.1. Use as a Shot Analysis Tool

One of the most powerful byproducts of learning a meaningful embedding space is the ability to detect and analyze

Figure 9. Side-by-side comparison of LW's normal shooting form (left) and intentionally distorted form (right).

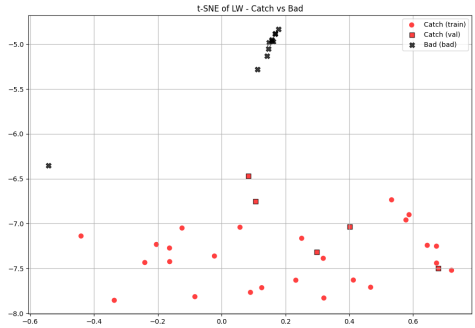

Figure 10. t-SNE embedding of LW's Catch and Shoot attempts, contrasting normal (good) form with purposely distorted (bad) form. Clustering indicates clear separation in the learned representation space, with bad shots (black) tightly grouped and distinct from the distributed good form embeddings (red).

deviations in a shooter's form—often caused by fatigue, discomfort with a particular shot type, or simply poor mechanics. For instance, some players naturally struggle more with certain types of shots; even in the NBA, athletes are often categorized as strong "Catch and Shoot" shooters or "Off the Bounce" shooters (i.e., shooting off the dribble). Because each embedding encodes the mechanics of a shot into a compact representation, new attempts can be visualized in relation to others using the t-SNE projection. This allows athletes to quickly identify when their form begins to drift from a known baseline. For example, a player might observe that their shot embeddings begin to cluster differently after multiple repetitions or when executing specific shot types, indicating a form breakdown potentially caused by fatigue or instability—issues that can then be targeted in training.

The learned embedding space also enables comparative insights across players. Using t-SNE visualizations, an athlete can explore whose form their shots most closely resemble, gaining inspiration from particular mechanics or consistency. Alternatively, a player can select a target shooter they want to emulate and leverage the previously described tools to analyze how far their current form is from that

ideal—and more importantly, how to adjust their training to close the gap.

To demonstrate the practical value of our analysis tools, we examine LW's "Bad" shots in comparison to his "Good" shots (defined here as his standard Catch and Shoot form). As context, **Figure 9** illustrates a synchronized freeze-frame from both forms at the same point in the shooting motion. The "Bad" form is visibly inefficient, with a behind-the-head release that introduces unnecessary variability and reduces efficiency.

The first step in identifying such breakdowns is simply recognizing that a problem exists. In **Figure 10**, the t-SNE embedding clearly separates LW's "Bad" shots from his "Good" ones, suggesting that the model has learned to represent these distortions meaningfully. In a real training context, this kind of visualization could help a coach or player immediately flag a deviation and investigate further using targeted tools.

The Vertical Deviation Heatmap especially provides a very striking evidence of form breakdowns, highlighting substantial misalignments in the vertical dimension. Focusing on the same moment captured in the freeze-frames (Frame $\sim$50, after sequences are normalized to 64 frames), we observe that LW's arms—and consequently his release point—are significantly shifted upward, as shown by the deep red coloration in **Figure 2**. This reflects the primary flaw in his "Bad" form: a behind-the-head release that disrupts both efficiency and consistency.

To further analyze and correct this flaw, LW could then turn to the interactive Scrubber Tool (**Figure 3**), which provides a frame-by-frame overlay of his normal and distorted forms. By observing joint-level deviations throughout the motion cycle, he can better understand when and how his form diverges. Equipped with this insight, LW could then design targeted drills to reinforce proper mechanics, helping him make concrete improvements and enhance his shooting performance overall.

6.2. Limitations

While the model shows strong performance across most tasks, there are a few limitations to note. One of the most prominent issues arises in the shot-type classification task for shooter LD as the model struggles to distinguish between LD's Catch and Shoot shots and Step Backs. This suggests that LD likely exhibits less mechanical variation between the two shot types than, for example, LW. As a result, embeddings for these shots significantly overlap in the learned space (**Figure 11**), leading to a reduced classification accuracy of 59.40%—a noticeable drop compared to the 90%+ accuracy achieved for LW. Even this, however, offers some insight: it indicates that LD's Step Back motion is quite similar to his Catch and Shoot, which may be an encouraging sign of consistency for the shooter.

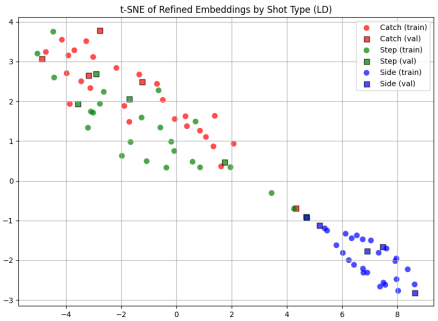

Figure 11. Learned embedding space for shooter LD after Phase 2 supervision. While Side shots remain well-separated, there is significant overlap between Catch and Step shots, reflecting the difficulty the model has in distinguishing those shot types for this shooter.

Another important limitation is the curated nature of our dataset. All videos were recorded under controlled conditions, with clean framing and no occlusions. Real-world applications—such as in-game footage or spontaneous shooting practice—would introduce additional challenges, including camera motion, partial occlusion, lighting variation, and more. Robustness to these factors would require further development, such as domain adaptation strategies or training on a more diverse dataset.

6.3. Future Extensions

While our current system performs well under controlled conditions, there are several promising directions for future work. First, expanding the dataset to include in-game or unstructured shooting footage would test the robustness of the pipeline and broaden its applicability. Additionally, integrating real-time feedback tools—such as alerting shooters when their form deviates significantly from baseline—could transform the system into an actionable, on-court training aid.

7. Conclusion

This work introduces a novel pipeline for analyzing basketball shooting form using pose-sequence-based embeddings learned through contrastive learning. By combining self-supervised pretraining with supervised fine-tuning, our two-phase approach achieves strong performance on both shooter identification and fine-grained shot-type classification tasks—even when visual differences, at least to the human eye, are quite subtle. Beyond classification, we present a suite of interpretability tools—including t-SNE visualizations, deviation heatmaps, and a pose scrubber—that enable athletes and coaches to gain actionable insights into shooting mechanics. Our results highlight the potential of embedding-based analysis as a powerful framework for personalized training, with promising applications both in basketball and in other domains involving motion analysis.

8. Contributions

Both authors contributed jointly to the core idea and collaborated closely throughout the project. Landon led data collection and was primarily responsible for the development and implementation of the Phase 2 supervised fine-tuning pipeline. Michael focused on Phase 1 self-supervised pretraining and the design of the interpretability tools. All major components were developed in active collaboration.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[2] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 1

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[4] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[6] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1